Trabalho de Conclusão de Curso

Bayesian Inference for Uncertainty Quantification in Binary Classification Tasks with Limited Data

Wagner da Silva Fontes
wsf@ic.ufal.br

Orientador:
Prof. Dr. Bruno Almeida Pimentel

Maceió, Maio de 2023

Wagner da Silva Fontes

Bayesian Inference for Uncertainty Quantification in Binary Classification Tasks with Limited Data

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Prof. Dr. Bruno Almeida Pimentel

Maceió, Maio de 2023

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecária Responsável: Lívia Silva dos Santos CRB - 1670

F682b Fontes, Wagner da Silva.
Bayesian inference for uncertainty quantification in binary classification tasks with limited data / Wagner da Silva Fontes. – 2023.
49 f.:il

Orientador: Bruno Almeida Pimentel.
Monografia (Trabalho de Conclusão de Curso em Ciência da Computação) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2023.

Bibliografia: f. 45-49

1. Inferência bayesiana. 2. Classificação binária. 3. Quantificação de incerteza. I. Título.

CDU: 004.22

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my family. To my parents, who have been a constant source of support and encouragement throughout my educational journey. To my grandparents, whose love and wisdom have been a guiding light. A special thanks to my brother, whose incredible sacrifices and unwavering support have been invaluable. The love and support of my family have been my strength and motivation, and this work is a testament to their belief in me.

I am immensely grateful to my professor and mentor, Bruno Pimentel, for accepting to advise me this year after a three-year hiatus from university. I am also thankful to professors Rafael Amorim and Fábio Coutinho, who have nurtured my entrepreneurial and scientific spirit during my time at the university. My appreciation extends to all my professors, their high standards and relentless challenges eventually led me to appreciate the true beauty of Math and Computer Science.

To my friends, França Mac Dowell, Nelson Gomes Neto, Lucas Raggi, Gabriel Barbosa, Daniel Vassalo, and Lucas Amorim, who have made the university experience memorable. Their camaraderie made the long lines at the Restaurante Universitário bearable and turned disappointments into laughter. I am also grateful to all the friends I made during my time at the university, including those from other institutes who encouraged me to persevere through the challenging times.

Finally, I would like to extend my gratitude to my coworkers, who are also my friends. Their knowledge and passion have been instrumental in fostering my interest in various subjects, including Bayesian Statistics. This work is as much a product of their influence as it is of my efforts.

This journey has been a collective effort, and I am fortunate to have been surrounded by such inspiring individuals. Thank you all!

"Doubt is not a pleasant condition, but certainty is absurd."

– Voltaire

Resumo

Nos últimos anos, a Quantificação de Incerteza emergiu como um componente essencial dos modelos de Aprendizado de Máquina, particularmente no contexto de tomada de decisões e avaliação de riscos. Neste trabalho, realizamos uma investigação abrangente sobre o desempenho da Inferência Bayesiana para a Quantificação de Incerteza em problemas de classificação binária com dados limitados. Aplicamos modelos de regressão logística bayesiana, bem como os outros dois métodos, a diversos conjuntos de dados pequenos e amplamente conhecidos. A performance dessas abordagens é avaliada utilizando a métrica F1 Score e comparada em diferentes níveis de incerteza preditos por cada método. Nossos resultados demonstram que a abordagem de Inferência Bayesiana não apenas exibe um desempenho competitivo, mas também fornece uma saída mais flexível na forma de uma distribuição, que viabiliza uma análise mais completa dos resultados e pode ser utilizada de diversas maneiras para capturar diferentes níveis de incerteza. Apesar dos desafios de convergência associados a conjuntos de dados pequenos e espaços paramétricos complexos, os modelos bayesianos conseguem capturar tanto as incertezas epistêmicas quanto as aleatórias, oferecendo uma estrutura robusta para a compreensão dos fenômenos subjacentes. Em contrapartida, os métodos de região de incerteza e Inferência Conformal apresentam certas limitações, como o tamanho reduzido do conjunto de treinamento para Inferência Conformal e inflexibilidade na saída, que podem potencialmente dificultar suas aplicações práticas. Nossas descobertas destacam o potencial da Inferência Bayesiana como uma abordagem eficaz para a Quantificação de Incerteza em problemas de classificação binária, mesmo quando confrontada com a limitação de dados, com performance equiparável aos métodos frequentistas especialmente com a escolha de distribuições priori não informativas.

Palavras-chave: Quantificação de Incerteza, Inferência Bayesiana, dados pequenos, classificação binária.

Abstract

In recent years, Uncertainty Quantification has gained significant attention as an essential aspect of Machine Learning models, particularly in the context of decision-making and risk assessment. In this study, we provide a comprehensive investigation of the performance of Bayesian Inference for Uncertainty Quantification in binary classification problems with limited data, and we compare its effectiveness with alternative approaches, such as the uncertainty region and Conformal Inference methods. We apply Bayesian logistic regression models, as well as the other two methods, to various well-known small datasets. The performance of these approaches is evaluated using F1 scores at different levels of the uncertainty predicted by each method. Our results demonstrate that the Bayesian Inference approach not only exhibits competitive performance, particularly when using non-informative priors, but also provides a more flexible output in the form of a distribution, this adaptability enables a more comprehensive analysis of the results and can be leveraged in various ways to capture different levels of uncertainty. Despite the convergence challenges associated with small datasets and high-dimensional parameter spaces, the Bayesian models capture both epistemic and aleatoric uncertainties, offering a robust framework for understanding the underlying phenomena. In contrast, the uncertainty region and Conformal Inference methods display certain limitations, such as reduced training set size for Conformal Inference and inflexibility in the output, which could potentially hinder their practical applications. Our findings highlight the potential of Bayesian Inference as an effective approach for Uncertainty Quantification in binary classification problems, even when faced with limited data.

Key-words: Uncertainty Quantification, Bayesian Inference, small data, binary classification.

List of Figures

List of Tables

Contents

1

Introduction

1.1 Motivation

The power of Machine Learning lies in its ability to make accurate and reliable predictions, a capability that has transformative implications across a range of fields [1], from finance and healthcare to transportation and environmental studies. This transformative power is largely due to its ability to sift through vast amounts of information and extract valuable insights. However, the real world often presents us with situations where we have to make predictions based on a limited amount of data [2]. This is where the true strength of a Machine Learning model is tested — its ability to discern meaningful patterns and make accurate predictions even when the data is scarce. When we apply Machine Learning to situations with limited data, we encounter unique challenges and opportunities. For instance, when the data is limited, there's a risk that the model might latch onto random patterns in the data that don't represent the underlying reality, failing to generate reliable outcomes or providing useful guidance for decision-making.

As we delve deeper into the realm of limited data, we realize the inherent fragility of the predictive power so often attributed to Machine Learning models. When there is scarcity in the wealth of information that these models typically thrive on, the probability of error in predictions escalates, while the accuracy of these predictions often dwindles. At the heart of this issue lies the concept of uncertainty. Uncertainty, in this context, can be understood as the degree of confidence we have in the predictions made by our model [3]. It is a reflection of the potential variability in these predictions, which may arise from various sources, including the inherent noise in the data, the limited size of the dataset, and the model's simplifying assumptions. Quantifying this uncertainty, therefore, becomes an indispensable part of the predictive process, especially when dealing with limited data.

Uncertainty Quantification provides a mechanism to measure, understand, and communicate the potential variability in our predictions [4]. It adds an additional layer of transparency

to the predictive process, by not only providing a prediction but also an estimate of the confidence we have in this prediction. This enables decision-makers to understand the possible range of outcomes and the likelihood of each outcome, facilitating a more informed and risk-aware decision-making process. When dealing with limited data, where the margin for error is slim, the importance of Uncertainty Quantification becomes even more pronounced. Therefore, it's crucial to equip our models with the ability to quantify uncertainty, thereby augmenting their predictive power with a deeper understanding of the underlying complexities and uncertainties inherent in the available data.

In this monograph, we traverse the landscape of Uncertainty Quantification with a particular focus on Bayesian Inference [5]. Bayesian Inference, named after the Reverend Thomas Bayes, is a method of statistical inference that allows one to update probability for a hypothesis as more evidence or information becomes available. It stands as an alternative to frequentist statistical inference that is often taught and practiced widely. At its core, Bayesian Inference is a way of learning from evidence as it accumulates. In Machine Learning, it provides a practical method for adapting to new data and dealing with uncertainty, and is therefore of significant interest in our study. The potential of Bayesian Inference in handling uncertainty is substantial, yet, it is not without its complexities and nuances. Thus we delves into the intricate labyrinth of Bayesian Inference and the comparative study of alternative methods to illuminate the unique strengths and challenges associated with each approach.

1.2 Objectives

The primary objective of this work is to investigate the efficacy of Bayesian Inference as a probabilistic approach for handling uncertainty in binary classification problems, particularly when faced with limited data. By comparing the performance of Bayesian Inference to alternative Uncertainty Quantification techniques, this study aims to provide a comprehensive evaluation of the advantages and limitations of the method. The comparison will be based on empirical results from experiments conducted on various datasets, highlighting the differences in performance, versatility, and computational complexity of each approach.

Lastly, this work aims to provide insights and recommendations for practitioners and researchers working with binary classification problems and limited data. By examining the benefits and drawbacks of each Uncertainty Quantification approach, the study intends to offer valuable guidance for selecting the most appropriate method, considering the problem context, available resources, and the desired level of interpretability. Through a comprehensive analysis of the experimental results and a discussion of potential future directions, this work can contribute to the advancement of knowledge in the field, promoting more informed decision-making and better risk assessment across domains and applications.

1.3 Structure

This monograph is structured to provide a comprehensive understanding of the application of Bayesian Inference for Uncertainty Quantification in binary classification problems with limited data. Each chapter serves a specific purpose and contributes to the overall objective of the research. After this introductory chapter, we provide a literature review in the Chapter 2, addressing the a comprehensive overview of the current state of research in the field of Uncertainty Quantification. The objective of this Chapter is to familiarize the reader with the existing body of knowledge and to highlight the gaps in the literature that this study aims to address.

The Chapter 3 delves into the theoretical underpinnings of predictive uncertainty, Bayesian Inference, and their differences from classical statistics. The objective here is to equip the reader with the necessary theoretical knowledge to understand the methodology and results of the research.

In the Chapter 4, we detail the procedures employed in our study. It outlines the datasets used, the implementation of the Bayesian Inference and the frequentist methods, and the evaluation strategy. The objective of this chapter is to provide a clear and replicable account of the research process.

The Chapter 5 presents the findings of the experiments and provides a thorough discussion of their implications. The objective is to interpret the results in the context of the research objectives and to provide insights into the performance and practical implications of Bayesian Inference for Uncertainty Quantification in binary classification problems with limited data.

Finally, Chapter 6 provides a summary of the study, highlighting the key findings and their implications. It also suggests potential avenues for future research. The objective of this chapter is to encapsulate the research journey and to provide a clear direction for future work in this field.

2

Literature Review

2.1 Introduction

Machine Learning, a subfield of Artificial Intelligence, is dedicated to the development of algorithmic models that aim to capture the underlying patterns and structures of a complex phenomenon or system. These models are designed to learn from data, make predictions, and guide decisions based on that learning. But given the intrinsic complexity of the real world, such models is but an approximation of the underlying reality, and in the process of constructing such an approximation, it may fail to capture some of the intricate details and nuances of the real world, which can introduce various forms of uncertainty or error in the model's predictions. The earliest discussions on this topic and the need for a better evaluation while working with the output of statistical models in a real scenario was addressed by McCullagh and Nedler, in their notable, pioneering book Generalized Linear Models [6]:

> Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeller. The first is that *all models are wrong*; some, though, are better than others and we can search for the better ones. At the same time we must recognize that eternal truth is not within our grasp.

This idea sparked interest among researchers in several domains and practical applications, leading to further investigations into this topic that guided to deeper formulations about the sources and types of uncertainty [3], that will be explored in the theoretical foundation section, as well as more complete methods to assess its effects in modeling [7].

The need to measure what we do not know, as well as the extent of such unknowns, has led the field of Uncertainty Quantification (UQ) to be studied in the most diverse areas of mathematical applications. Mervin [8] emphasizes the significance of this topic in enabling computational simulations for drug discovery, citing applications such as adverse drug reaction prediction and

molecular toxicity prediction through frequentist and Bayesian methods for estimating confidence in model outputs. The author also points out the UQ as a fundamental factor to expedite decision making saving considerable time and effort, allowing the development of applications that need more interpretable results in addition to high accuracy. In a complete distinct domain, Jung [9] accent the criticality of uncertainty estimation for risk assessment in flood-prone areas through flood inundation models predictions in order to avoid both human and economic losses, assist rescue and relief operations during floods, and helping to design prevention strategies.

In this chapter, we aim to present a comprehensive yet concise overview of the most advanced techniques for Uncertainty Quantification currently available. It is crucial to recognize that these state-of-the-art methods are predominantly tailored for Deep Learning algorithms. As a result, we will not only discuss these cutting-edge approaches but also delve further into alternative strategies for estimating uncertainty. Our focus will be on methodologies that are particularly suitable for handling small data and employing conventional Machine Learning algorithms, ensuring a thorough understanding of the various options for uncertainty estimation in diverse scenarios.

2.2 Cutting-Edge Approaches

Due to its growing importance for a wide spectrum of tasks, the ability to accelerate the decision process, and reduce the risk associated with using the Machine Learning model's predictions, numerous methods for Uncertainty Quantification have been developed. To this date, the state-of-the-art methods are related to sophisticated techniques applied to Deep Learning models. Monte Carlo Dropout (MC-dropout) is an approach initially proposed by Gal and Ghahramani [10], which uses the Dropout mechanism in inference time to estimate predictive uncertainty. Transforming the technique, originally thought as a means of regularization to reduce overfitting in artificial neural networks, into a mechanism to produce variations in the algorithm predictions without necessarily adding a new layer of complexity. In this way, a single model becomes capable of generating different predictions for the same sample, composing a distribution used to measure the confidence in the obtained output.

Bayesian Neural Networks (BNNs) [11] and Probabilistic Backpropagation (PBP) [12] present approximations of Bayesian methods that naturally account for uncertainty in parameter estimates and can translate this uncertainty into predictions. Both methods result in a structure in which the weights associated with each neuron follow a probability distribution and are not expressed only by a scalar as in a vanilla neural network. The former is, as its authors describe, "at its core a probabilistic model augmented with neural network as a universal function approximator", but it becomes computationally expensive because it depends on sampling or variational inference steps. The latter describes a modification of the backpropagation algorithm employed during the training phase, and by modifying only the information propagation

mechanism through the network, it poses an alternative to the scalability problem.

Lakshminarayanan et al. (2016), while proposing the Deep Ensembles method [13], argue that the quality of predictive uncertainty obtained while using the approximation of Bayesian methods for neural networks, such as BNNs and PBP aforementioned, depends on the degree of approximation of such methods, which are limited by being harder to implement and computationally slower to train compared to non-Bayesian approaches. The suggested ensemble of neural networks uses an adversarial training strategy that helps to capture ambiguity in training targets and, as learning is represented by multiple models, the method naturally becomes more robust to model misspecification. The solution outperforms or gets a similar outcome for most of the datasets when Lakshminarayanan compares it to results from BNNs, PBP, and MC-Dropout. This is used by the authors to justify Deep Ensembles as the best alternative among the evaluated techniques, mainly due to the simplicity of implementation since it requires only minor modifications to the standard training pipeline, and for being suited for distributed computation enabling large-scale Deep Learning applications.

2.3 Alternative Approaches for Conventional Algorithms

In the expansive realm of Artificial Intelligence, a multitude of sophisticated techniques have been developed for Deep Learning models. These techniques, which range from image pattern detection to emotion recognition, have been extensively studied and applied in various contexts [4]. However, as Kang et al. have pointed out [14], the lion's share of research efforts in Uncertainty Quantification (UQ) have been channeled towards Deep Learning and reinforcement learning frameworks.

On the other hand, traditional Machine Learning algorithms, which are often employed in scenarios with limited computational power and data availability, have received comparatively less attention. These algorithms, despite their simplicity, can be highly effective, particularly in situations where data is scarce. In this subsection, we delve deeper into strategies that are not only effective on small datasets but also capable of providing uncertainty measurements for the predictions. This exploration is crucial in the context of traditional Machine Learning algorithms, as it offers a more nuanced understanding of their performance and the associated uncertainties.

2.3.1 Calibrated Point Estimates

Most traditional algorithms for classification tasks (Logistic Regression, Decision Trees, Naive Bayes, etc.) are capable of predicting a "probability score" for each associated class. One primary and simplistic way to compute UQ involves interpreting the point estimate of the model's conditional probability as a gauge of its level of confidence. Then, when the predicted values are either extremely low or high, it may be read as a high level of confidence in the model's

predictions. On the other hand, if the outputs are close to 0.5, it indicates a lower level of confidence. The major problem with this approach is that not all models are well-calibrated [15], i.e. most models do not have the requirement to predict outcomes consistent with the actual frequencies of events occurring. In other words, a well-calibrated algorithm is able to produce a score of 0.8 for an event when the event occurs roughly 80% of the time, making such a score immediately interpretable as a probability.

Feng [16] applies the Platt Scaling method [17] to calibrate a Machine Learning model in pursuit of improving the uncertainty analysis in lithofacies classification. To transform the original output into calibrated probabilities, first, the prediction $f(X)$ is passed through a sigmoid function:

$$P(y = 1|f) = \frac{1}{1 + exp(\beta_0 X + \beta_1)} \tag{2.1}$$

A and B are parameters fitted using maximum likelihood estimation from a holdout set. Gradient descent is then used to find A and B:

$$\arg\min_{A,B} \sum_i (y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)) \tag{2.2}$$

where:

$$p_i = \frac{1}{1 + exp(\beta_0 X_i + \beta_1)} \tag{2.3}$$

Platt Scaling and other similar calibration methods are model-agnostic and can be applied in a variety range of ML algorithms [18; 19; 20], and in most cases can successfully generate more calibrated outputs. However, even making the prediction more interpretable, as it is a post-processing operation, it needs a subset of data unseen during the training step to avoid unwanted biases [16], which may be harmful to small datasets. Besides, as discussed more deeply in [8], as the calibration process is generally independent of the training step, the produced UQ cannot account for the uncertainty in the model parameters and thus does not carry the expected attributes of uncertainty estimation.

2.3.2 Conformal Inference

Conformal prediction [21] is a statistical learning framework that can be employed to estimate predictive uncertainty for Machine Learning models without the need to modify their fitting process. The construction of prediction regions in conformal inference is based on the concept of "nonconformity score", which measures the distance of a new data point from the training data. Such regions translate into a set of labels for classification problems and intervals for regression tasks that may be read as error bounds around the predictions. The conversion of point estimation to set estimation intrinsically involves the notion of uncertainty, thus predictive uncertainty would be the size of the conformal set (or interval) [22].

This method enables the transformation of potentially uncalibrated point estimates into con-

formal sets with marginal coverage guarantees. With $C(x)$ as the conformal set of predictions for a sample x, and y the true label, the conformal method ensures that:

$$P(y \in C(x)) = 1 - \alpha \tag{2.4}$$

Where α is a parameter, and $1 - \alpha$ is averaged over a holdout dataset. As discussed in [23], although marginal coverage is assured, conditional coverage is unfeasible using the conformal method since the resultant set is independent of the samples specific features, being defined instead by the average of the distances to other instances in the dataset.

This frequentist approach, although facilitates assembling prediction intervals for problems where the target is continuous [24], does not provide a clear uncertainty notion easily applicable to classification tasks. Compared to methods for UQ through point estimates, it has the advantage of being able to guarantee, with a parameterizable significance level, the coverage of the predicted set. However, it presents some of the negative points discussed in the previous method, such as usage of a dataset with samples not seen in the training, and it is indifferent to the uncertainty inherent to the model parameters.

2.3.3 Uncertainty Region

Lavazza and Morasca [25], seeking to introduce Uncertainty Quantification in Machine Learning models applied to detect faulty modules in software systems, proposed a novel method that does not require any post-processing step as in conformal inference and model calibration. The process utilizes the points estimate from a classifier without the demand for calibration, relying on the idea that there is always a region for the predicted score so close to a random estimation that a prediction is hardly accurate and shouldn't be regard as reliable. For random estimate, the authors mean the frequency of a class in the dataset, following the frequentist definition of probability, which defines the probability of an event as the limit of the proportion of occurrences when the number of observations tends to infinity.

To explicitly specify the prediction interval for a given score it's necessary, first and foremost, to compute the standard deviation of such prediction. Considering a binary classification problem where x is a sample and $f(x)$ its respective score, $f(x)$ can be associated with a Bernoulli distribution with two possible events of variance defined by:

$$var(x) = p(x)(1 - p(x)) \tag{2.5}$$

and then, the prediction interval lower, $lb(x;k)$ and upper bounds, $ub(x;k)$, are expressed:

$$[lb(x;k), ub(x;k)] = [p(x) - k \cdot \delta(x), p(x) + k \cdot \delta(x)] \tag{2.6}$$

where $\delta(x)$ is the standard deviation,

$$\delta(x) = \sqrt{p(x)(1-p(x))} \tag{2.7}$$

and k the parameter to scale the interval. Despite not providing a methodology to define the parameter k, the authors apprise that its value should not be too large making all intervals unreliable, nor too short making the model overconfident.

Besides quantifying how spread the values of $p(x)$ are around the predicted value, the concept of uncertainty region is presented to classify every prediction as reliable or unreliable. If the random estimate is within the predicted interval, the score is too close to the random and should be regarded as unreliable. Thus, a reliable score is defined by:

$$up(x;k) < PR < lb(x;k) \tag{2.8}$$

This method provides a straightforward uncertainty interpretation by delimiting the frontiers of an uncertainty region in which the prediction should be as far as possible to have a high confidence level. While benefiting from not needing additional data for UQ, this process is not easily adaptable for regression tasks and introduces some issues when applied to classification problems with more than two classes.

2.3.4 Bayesian Inference

In broad terms, Bayesian Inference is a learning technique that employs probabilities to describe and reason about our convictions. More specifically, this approach offers a way to accurately revise our beliefs when new data becomes available. In this method, we can easily describe a prior knowledge to constrain the model when calculating the probability of a related event, and then get uncertainty estimation in the form of posterior distributions [5]. As its name suggests, Bayes' theorem is the key concept behind this technique, and can be defined as:

$$P_{post}(\theta|X,y) = \frac{P(\theta)P(Y|X,\theta)}{P(y|X)} \tag{2.9}$$

where (X,y) are the samples, with X being the input variables or predictors, and y being the output variables or targets, θ is a set of unobserved parameters that define a probabilistic model, and $P(\theta)$ represents our prior beliefs stated in the initialization of the problem. The Bayesian Inference gives us a way to infer the posterior distribution $P_{post}(\theta|X,y)$ by multiplying the prior $P(\theta)$ by the likelihood function $P(Y|X,\theta)$ each time a new observation is available. $P(X|y)$ can be read as a normalization factor that ensures the posterior distribution sums to 1.

Essentially, any parametric model such as linear regression, logistic regression, or even a neural network can be expressed as a Bayesian model by defining prior probabilities for its parameters θ. Nevertheless, as mentioned in [26], it may be impractical to derive the posterior analytically for some distributions when dealing with complex models, or even when the

data is high-dimensional. Some approximation methods have been developed in an effort to overcome this intractability, sampling techniques like Monte Carlo Markov Chain (MCMC) [27] and Gibbs [28], or even optimization approaches where the posterior is approximated with some simpler parametric distribution like in Variational Inference [29].

The resulting posterior from Bayesian modeling, which essentially encapsulates the updated belief about the model parameters, can be significantly influenced by the chosen prior distributions, especially when the dataset at hand is small or limited. This noteworthy sensitivity to priors illustrates one of the distinct nuances of the Bayesian approach, where the balance between prior beliefs and data evidence is key. Other characteristics, intricacies and facets of the Bayesian framework, along with the requirement and detailed definitions for these approximation methods, will be delved into with more depth and detail in the ensuing section. This in-depth exploration aims to shed light on the complexities, opportunities, and challenges inherent in the Bayesian approach to uncertainty quantification.

Besides the sophisticated methods employed in Deep Learning models discussed earlier, Bayesian Inference has also been applied in simpler algorithms to generate uncertainty estimations [26; 30; 31]. Utilizing the potential of Bayesian method approximation techniques in linear models, this monograph investigates the application of logistic regression to obtain uncertainty estimation in situations where data are scarce.

3

Theoretical Foundation

Predictive uncertainty plays a crucial role in various fields and applications, as it enables practitioners to quantify the level of confidence in their predictions, providing a more nuanced understanding of the outcomes. This concept is particularly relevant in machine learning, where models are trained to make predictions based on available data. In order to account for the inherent variability in the data and the limitations of the model, it is essential to have a robust framework for assessing and quantifying uncertainty. This chapter aims to provide the necessary theoretical foundation to understand the concepts of predictive uncertainty, Bayesian inference, and their differences from classical statistics.

This chapter will also explore the logistic regression model, a widely used technique for binary classification problems, and how it can e designed from a Bayesian perspective. The Bayesian logistic regression model allows us to not only make predictions but also estimate the uncertainty associated with each prediction, which can be invaluable for decision-making and risk assessment. To evaluate the performance of a classification model, it is essential to have a set of metrics that can measure the quality of its output. This chapter will also introduce some machine learning metrics that are commonly used to assess the performance of classification models. These metrics will be fundamental for understanding the experimental results throughout this work, where a Bayesian logistic regression model will be applied to a real-world classification problem. By leveraging the Bayesian framework for uncertainty quantification, we will demonstrate how it can be used to make more informed predictions and better understand the limitations of the model. This theoretical foundation will provide the necessary background for interpreting and appreciating the experimental results and their implications.

3.1 Predictive Uncertainty

Machine learning, at its core, is about creating models from data to make predictions or inform decision-making processes, and since it involves generalizing from specific observations

to broader models, uncertainty is an inherent aspect of the field. The very process of learning from a limited sample requires some inductive reasoning, which, by nature, is prone to some level of conjecture. Thus, predictive uncertainty refers to the quantification of the variability or lack of confidence in predictions made by a model, which for the most part stemming from the variance and noise present in the data commonly deriving from measurement errors.

Moreover, the model itself can be a significant source of uncertainty. Limitations in its structure or assumptions can lead to inaccurate approximations of a system. For instance, a model may not capture the complexity of the underlying data-generating process or may be too simplistic, resulting in less reliable predictions. Even with a suitable model structure, uncertainty can arise from the its parameters estimation process. This is especially true when the data is limited, or hold specific characteristics that may impose some difficulty in the learning process, such as highly correlated variables or underrepresented groups.

Predictive uncertainty is widely studied across various fields, with researchers seeking to understand its sources, quantification methods, and how to reduce its impact on models. Accurately representing uncertainty is crucial and considered a fundamental aspect of almost all machine learning techniques, especially in safety-critical areas like medicine or technical systems. In fact, estimating and quantifying uncertainty on a case-by-case basis is arguably more important and practical than focusing on average accuracy or confidence, which is commonly reported in machine learning studies.

The first form of uncertainty is commonly described as epistemic, for which the data quality and model specification is usually pointed out as the main reasons for its existence [3; 8]. Most real-world applications face the challenge of data quality, which may have data rich in quantity but poor in representativeness — as illustrated in the upper right area of Figure 3.1 — since a dataset contains just a sample of the real data and never all observations, thus it virtually never represents the entire population. The chosen algorithm or model architecture also poses an obstacle since it may be improper to represent the desired learning. Nevertheless, both issues can be mitigated through the acquisition of more data points and adjustments in the model design, respectively, hence the epistemic uncertainty is assumed as reducible.

On the other hand, aleatoric uncertainty is the irreducible aspect of uncertainty, stemming from the inherent stochasticity in observations and the uncertain relationship between independent and dependent variables [32]. Captured data is not a perfect representation of reality and is influenced by noise and randomness, as exemplified at the bottom left of Figure 3.1. Each observation carries inherent noise that can't be controlled, and this noise accumulates across observations and propagates through the system, forcing the model's predictions to deviate from the true outcome. Since this uncertainty is a property of the data distribution rather than the model itself, it can't be reduced by adding information or choosing a different representation.

However, despite the seemingly clear definitions of the types of uncertainty, distinguishing between these concepts can be challenging in some scenarios, and they may even become interchangeable. In practice, modifying the instance space — by adding a new feature, for ex-

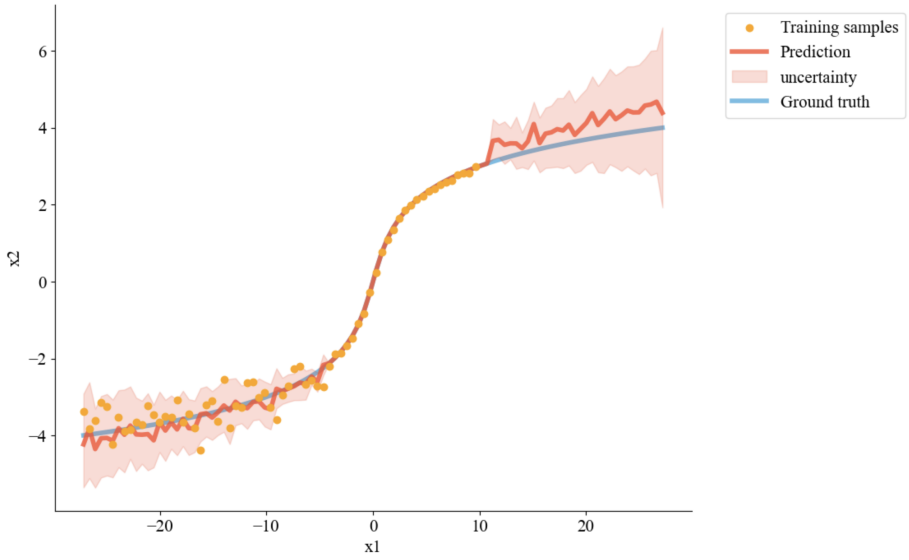

Figure 3.1: A synthetic example of epistemic and aleatoric uncertainties and predictive uncertainty originated for both cases. At the bottom left, one can see the noise in the samples posing difficulties to the learning process (aleatoric), while at the top right, the lack of data it's the root cause of a poor generalization (epistemic).

ample — can reduce noise or fix a class overlap, turning aleatoric uncertainty into epistemic and vice versa. This blurs the distinction between the two types of uncertainty and makes it even more difficult to quantify. Hüllermeier and Waegeman [33] thoroughly explore and exemplify this idea, reporting that "generally, embedding data in a higher-dimensional space will reduce aleatoric and increase epistemic uncertainty because fitting a model will become more difficult and require more data".

3.2 Bayesian Statistics

Bayesian statistics is a branch of statistical analysis that provides a systematic and logical framework for incorporating prior knowledge, beliefs, and uncertainties into the statistical inference process. Rooted in the work of Thomas Bayes and Pierre-Simon Laplace in the 18th century, over the centuries this approach has evolved into a powerful and flexible framework for reasoning under uncertainty. Unlike its frequentist counterpart, Bayesian statistics employs probabilities to represent not only long-term frequencies but the degree of belief or certainty in a particular event or hypothesis. In this section we will explore the primary distinctions between Bayesian and classical statistics and delve into the underlying theory of its key components.

3.2.1 Bayesian versus Frequentist Inference

In the field of statistical inference for uncertainty estimation, two primary schools of thought prevail: the frequentist and Bayesian approaches. Generally speaking, the Bayesian method

was developed in the 18th century through the work of Bayes and Laplace, and since the beginning required a lot of effort when dealing with complex distributions, which later translated into a high computational cost. In the early 20th century, Fisher and Pearson introduced frequentist concepts such as confidence intervals, significance level, and hypothesis testing, which required less computational power and consequently gained widespread acceptance. While frequentist measures like confidence intervals and P values still dominate research, the Bayesian approach has experienced a resurgence due to advancements in computing power and increased data availability. Although some differences persist, most experts agree that solving complex problems often requires insights from both approaches working together [34].

The primary differences between these two statistical methods largely stem from their respective interpretations of probability, its intrinsic nature, and the means by which this property can be accessed. Frequentist statistics consider a data sample as a result of one instance within an infinite number of precisely replicated experiments. Any conclusions drawn from this perspective are predicated on the assumption that events transpire with probabilities, which represent the long-term frequencies of those occurrences in an unbounded series of experimental repetitions. For example, when flipping a coin continuously, the proportion of observed heads within an infinite number of tosses defines the probability of obtaining heads. This concept is founded on the notion that such probabilities genuinely exist and remain constant for each series of coin tosses conducted. In frequentist statistics, data are presumed to be random, originating from sampling a fixed and well-defined population distribution.

Conversely, Bayesian statisticians do not envision infinite experiment repetitions as a prerequisite for defining and specifying probabilities. Instead, probability serves as a measure of certainty regarding a specific belief. Consequently, the probability of obtaining heads in a coin toss quantifies the underlying belief that the coin will land in that particular orientation before it is flipped. In this context, Bayesians do not perceive probabilities as fundamental laws of cause and effect but rather as abstractions to express our uncertainty. In this sense, it is unecessary for events to be repeatable to define a probability. Rather than being a fixed attribute, probabilities in Bayesian statistics are viewed as expressions of subjective beliefs, which can be updated based on new data. Bayes' rule (Equation 3.1) is the key concept behind Bayesian inference, where probabilities are utilized to convey the uncertainty in parameter values following new observations, offering a logical method to execute the process of updating our prior beliefs as intuitively demonstrated in Figure 3.2. A detailed exploration of such learning process and the constituent parts of Bayes' Theorem will be conducted in the next section.

$$P_{post}(\theta|X,y) = \frac{P(Y|X,\theta)P(\theta)}{P(y|X)} \tag{3.1}$$

Bayesians believe that since we observe the data, it is fixed and does not change. There is no need to consider an infinite number of potential samples or assume that the data results from a random sampling process. We never have perfect knowledge of an unknown parameter (e.g., the

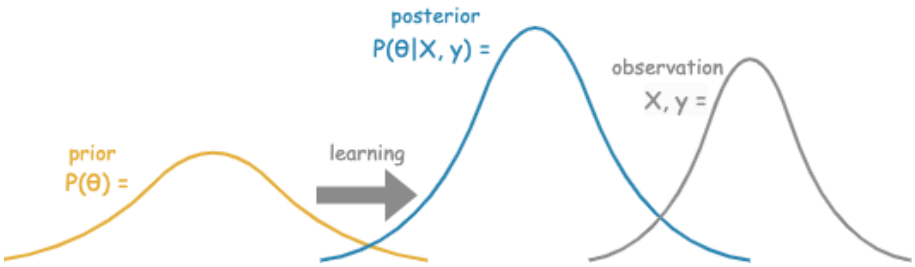

Figure 3.2: Learning process through Bayes' rule. A new observation is used to update a prior probability distribution yielding a new posterior distribution.

probability of a coin landing heads up), and this epistemic uncertainty, which relates to our lack of knowledge, leads to treating the parameter as probabilistic. There are two ways to interpret this assumption [35]. Firstly, we can view the unknown parameter as genuinely fixed in an absolute sense, but our beliefs are uncertain, so we use probability to express this uncertainty. That way the sample is just a noisy representation of the signal, resulting in different outcomes for each set of coin tosses. Alternatively, we can assume that there is no definitive, unchanging probability of obtaining heads, meaning that each sample we take unknowingly yields a slightly different parameter. In this case, we get varying results from each coin-flipping round because we expose our system to a slightly different probability of landing heads up each time. Although these two perspectives differ philosophically, they are mathematically equivalent, allowing us to apply the same analysis to both.

Frequentist statisticians adopt a different perspective when it comes to the unseen and known parts of a system. They consider the parameters of the probability model, which represent the unseen aspects of the system, to be fixed and unchanging. In contrast, the known parts of the system, such as the data, are viewed as variable and subject to change. This viewpoint is less flexible in comparison to the Bayesian, as it assumes that these parameters remain constant or represent the average outcome of a long series of identical experiments. While this approach has its merits, it can also be limiting, as it may not adequately account for the inherent uncertainty in the parameter estimates, especially when dealing with complex systems or limited data.

In statistical inference, the goal is to form conclusions based exclusively on probability rules. To summarize the evidence for a particular hypothesis, the probability of the hypothesis given the data is utilized. The challenge is that when performing the statistical inference through a probability model to describe a scenario, it permits the calculation of the probability of obtaining the observed data given the hypothesis is true, which is exactly the opposite of the intended objective since this probability accounts for all potential samples that might have been observed from the population if the hypothesis were valid. Frequentists employ inverse probability as proof supporting a specific hypothesis, assuming a hypothesis to be true and estimating the probability of securing the observed data sample based on that assumption. If the probability is low the hypothesis is deemed improbable, resulting in its rejection. However, if

the probability does not fall below a predetermined threshold, the hypothesis is not discarded.

3.2.2 Elements of Bayesian Inference

Bayes' theorem streamlines the accurate quantification of how observed data changes our beliefs. In this context, the primary goal is to establish $P(\theta|X,y)$, which can be thought as the degree of conviction in our beliefs given the data we have observed. This component of the formula is referred to as the posterior probability, and it is the aspect that Bayes' theorem seeks to resolve through the correctly estimation of the set of parameters, denoted as θ in equation 3.1. These parameters can be regarded as knobs that can be adjusted to modify the behavior of a statistical model and may represent tangible quantities (such as the average rainfall in a specific region) or abstract concepts (like the variance in a normal distribution).

As a means to compute the posterior probability, several other components are required. The first is $P(Y|X,\theta)$, representing the probability of the data given our beliefs about the data, and is commonly referred to as the likelihood. The likelihood conveys the plausibility of the data based on our belief. Subsequently, the probability of our initial belief, $P(\theta)$, must be determined. This element is known as the prior probability, or simply "the prior", as it encapsulates the strength of our belief before observing the data. The likelihood and the prior combine to yield a posterior probability. Typically, the probability of the data, $P(y|X)$, is employed to normalize the posterior, ensuring it accurately represents a probability ranging from 0 to 1. In this section, we will examine each component of the theorem individually, illustrating their importance and roles in the learning process.

Likelihood

In probability models, various system behaviors are generated by adjusting a set of defining parameters. If the chosen model is appropriate, it should be possible to modify these parameters in a way that allows the model to replicate the behavior of the real-world system being studied. When setting the parameter values and using the model to generate data, the resulting data distribution acts as a valid probability distribution. In Bayesian inference, however, the goal is to determine a posterior belief for each set of parameter values. This implies that the data is held constant while the parameter values are changed. In this context, the probability model no longer acts as a valid probability distribution, i.e., the distribution no longer sums (for discrete distributions) or integrates (for continuous distributions) to 1. To emphasize the distinction, likelihood is the preferred term in Bayesian inference as opposed to probability distribution.

Consider a statistical model designed to describe a typical dice roll. In this scenario, we do not account for any environmental factors that could impact the outcome, such as the force exerted during the throw, the surface it lands on, or even the weight of the die. Furthermore, we assume that previous results have no influence on future outcomes. Due to this lack of information, the statistical model cannot precisely predict the behavior of the dice, resulting in

a probabilistic rather than a deterministic model. For simplification purposes, we denote the faces of the die as D1, D2, D3, D4, D5, and D6, representing all six possible outcomes. We also assume that the die is fair, so the probability of a specific face appearing is given by $\theta = \frac{1}{6}$. Additionally, based on the independence assumption, if the die is rolled twice, we assume that the first roll's outcome does not influence the second roll's outcome. Thus, we can calculate the probability of obtaining the same face in two consecutive rolls:

$$\begin{aligned} P(D1, D1|\theta) &= P(D1|\theta) \times P(D1|\theta) \\ &= \theta^2 \\ &= (\frac{1}{6})^2 = \frac{1}{36} \end{aligned} \tag{3.2}$$

Adopting a Frequentist perspective, where the probability of an event is fixed and assuming that the die in question is indeed fair, we can calculate the probability of obtaining a specific sum when rolling the die twice, each possible permutation is shown in Figure 3.3. Figure 3.4 displays the probabilities for each outcome. As all individual events in the depicted discrete distribution are non-negative and their sum equals 1, it constitutes a valid probability distribution.

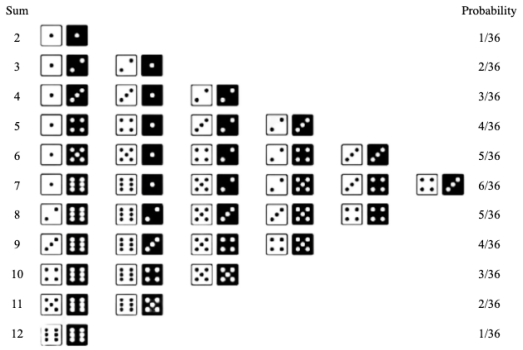

Figure 3.3: Visual representation of the sum of two dice.

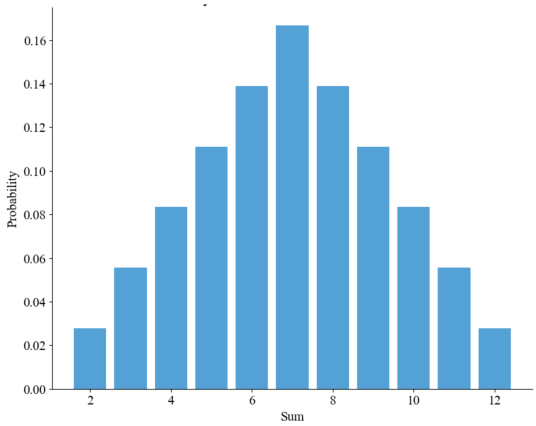

Figure 3.4: Probability distribution of the sum of two dice.

In this probability distribution, we consider a specific case where the value of θ is held constant, and the data encompass all possible outcomes when rolling the dice. However, as previously mentioned, Bayesian inference does not maintain fixed parameters in the model. Instead, the data remain constant while the model parameters vary. Bayes' theorem provides a method to calculate the posterior probability density for any given value of the parameter. Suppose we have a die with an unknown inherent θ. To estimate the posterior belief for any value of θ, we must compute $P(Y|\theta)$ (where Y represents the outcome of rolling the dice) for each potential value of θ. Imagine we roll the dice twice, seeking for any combination of values without considering the order, in which D1 appears exactly once. We can utilize our model to calculate the probability of this outcome for any value of θ:

$$\begin{aligned} P(D1, \neg D1|\theta) + P(\neg D1, D1|\theta) &= \theta(1-\theta) + (1-\theta)\theta \\ &= 2\theta(1-\theta) \end{aligned} \tag{3.3}$$

Although it may appear to resemble a continuous probability distribution (as seen in Figure 3.5), the integral over the domain of θ does not equal 1 (as demonstrated in Equation 3.4). Therefore, when we vary θ, $P(Y|\theta)$ does not constitute a valid probability distribution. To highlight this difference, in Bayesian inference the term "likelihood" is used to describe $P(Y|\theta)$ when varying the parameter, θ.

$$\begin{aligned} \int_0^1 2\theta(1-\theta)\,d\theta &= \int_0^1 2\theta - 2\theta^2\,d\theta \\ &= 2\int_0^1 \theta\,d\theta - 2\int_0^1 \theta^2\,d\theta \\ &= 2(\frac{1^2}{2} - \frac{0^2}{2}) - 2(\frac{1^3}{3} - \frac{0^3}{3}) \\ &= \frac{1}{3} \end{aligned} \tag{3.4}$$

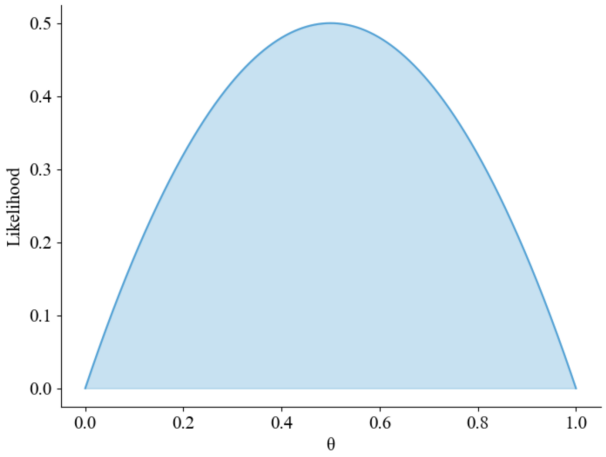

Figure 3.5: The likelihood function representing the event of receiving D1 exactly once when rolling dice twice.

Prior

The prior probability component in Bayesian inference is an essential element that captures our preliminary belief or understanding concerning a parameter or hypothesis before incorporating any observed data. As the name suggests, the term "prior" indicates that it is formulated "prior to utilizing the data." This component is represented by an unconditional probability and, unlike likelihood, it must always constitute a valid probability distribution. The selection of prior probability can be subjective, potentially influenced by expert knowledge, historical data, or

reasonable assumptions regarding the parameters [5]. The choice of prior probability often falls into one of two categories: non-informative priors and informative priors, each with its unique characteristics and applications.

A non-informative prior, alternatively referred to as a flat, vague, or non-informative prior, represents a category of prior probability that conveys minimal or no initial knowledge about the parameters or hypotheses. In such instances, the prior is typically chosen to exert minimal impact on the posterior distribution. Uninformative priors are frequently employed when there is little or no information available about the parameters, or when the aim is to circumvent the inclusion of subjective beliefs in the analysis. These priors can be uniform, ascribing equal probabilities to all possible parameter values, or they may follow specific distributions that maximize a particular measure.

Conversely, an informative prior constitutes a type of prior probability that integrates our existing knowledge or beliefs about the parameters or hypotheses. This prior is determined based on expert opinions, historical data, or other pertinent information. Informative priors prove valuable when we have reason to believe that certain parameter values are more probable than others, and we wish to incorporate this information into our analysis. The selection of an informative prior can be subjective, thus care should be taken to ensure the prior accurately and impartially reflects the available information.

Priors undeniably represent one of the most debated aspects of Bayesian statistics, mainly due to the innate subjectivity they introduce to the analysis. In situations where prior information is either ambiguous or challenging to articulate as a probability distribution, this subjectivity can become a sticking point. Critics often argue that the presence of subjectivity undermines the rigor and objectivity of the statistical analysis. These disagreements surrounding the specification, interpretation, and implications of informative priors have given rise to two distinct philosophical perspectives within Bayesian statistics: objectivism and subjectivism [36].

In the objectivist approach, the goal is to choose prior probabilities that reflect minimal prior knowledge, aiming for non-informative or weakly informative priors. This approach attempts to eliminate or reduce the influence of subjectivity in the selection of priors, seeking to provide a more impartial starting point for the analysis, allowing the data to have a greater influence on the posterior probability. In the subjectivist approach, the selection of prior probabilities is based on the subjective beliefs or expert knowledge of the analyst. This approach acknowledges that prior beliefs are subjective and allows for the incorporation of relevant contextual information in the analysis, embracing the fact that different analysts may have different prior beliefs, which could lead to varying posterior probabilities based on the same data [36].

The primary concern raised by subjectivist critics is that selecting non-informative priors could potentially compromise the performance of a statistical model and bring it close to a conventional frequentist model [37; 38]. This debate accentuates the significance of correctly specifying the prior, especially when the available data is limited or subject to noise. The impact of changing the prior before observing the data is demonstrated in Figure 3.6, where the

posterior distribution is notably influenced by the choice of prior. This highlights the crucial role that priors play in shaping the outcomes of Bayesian analyses and underscores the importance of careful consideration when determining appropriate priors for a given problem.

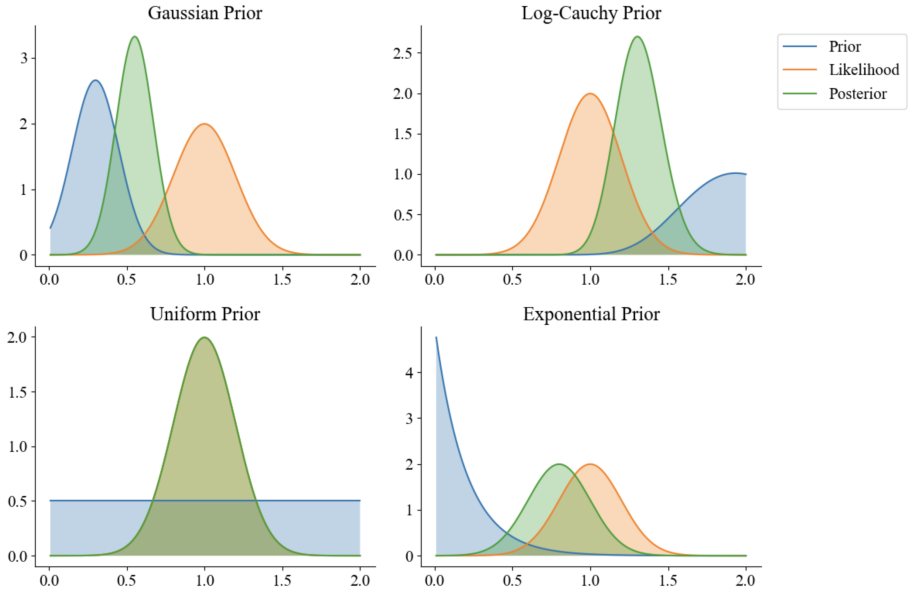

Figure 3.6: Variations in prior distributions yield distinct posterior distributions.

Denominator

As previously discussed, the likelihood functions might not represent a valid probability distribution. Even though the prior is a valid probability distribution when both are multiplied to form the numerator of Bayes' rule, we could still end up with an invalid probability distribution, as it may not sum to 1. The denominator, $P(Y)$, can be understood as a probability distribution representing our beliefs over possible data samples. Upon obtaining a specific data sample, this distribution collapses into a single value that normalizes the numerator of Bayes' rule, ensuring the posterior distribution is valid. The calculation of this term requires marginalizing out all parameter dependencies in the numerator (the product of the likelihood and prior).

In discrete models, computing the denominator is generally not problematic, as marginalization involves a sum that is computationally tractable, except in cases with a large number of parameters. However, for continuous models, marginalization requires calculating a multidimensional integral. Accurate calculation or appropriate approximation of this integral becomes challenging when there are more than a few parameters, which is common in applied Bayesian analyses involving multiple continuous parameters.

To address this challenge, two methods have been proposed. The first method employs conjugate priors to the likelihood, ensuring that the posterior shares the same distributional form as the prior, but with different parameter values [39]. Although the formulae for these posterior parameters have been tabulated, the application of conjugacy is typically limited to relatively

simple problems since it requires that the prior and posterior belong to the same parametric family. The alternative method involves sampling from the posterior distribution and using the properties of the obtained sample to approximate the corresponding properties of the posterior distribution. This approach serves as the basis for most modern computational techniques used in applied Bayesian analysis [27; 28].

Posterior

The posterior distribution, the final outcome of Bayesian Inference, combines our prior beliefs with information from observed data to create an updated understanding of our knowledge. This synthesis of past experience and data often (though not always) results in reduced uncertainty compared to the prior, as the data provides deeper insight into the world. Following the learning process (iteration over the data or evidence), this probability distribution can be employed to estimate the parameters of a statistical model, which can then be used for forecasting, generating predictions as distributions or point estimates.

The posterior distribution is particularly valuable in expressing predictive uncertainty, as it accounts for both the variability in the data and the uncertainty linked to our prior beliefs. Rather than relying on point estimates, which may be misleading or insufficient in capturing real-world complexity, the posterior distribution offers a spectrum of potential parameter values and their corresponding probabilities. This allows for improved decision-making and more robust risk assessment by quantifying our confidence in various outcomes or predictions. By utilizing the posterior distribution, Bayesian inference provides a powerful and adaptable framework for managing diverse sources of uncertainty and making well-informed inferences across a wide range of applications.

While it is possible to estimate the full posterior distribution for a parameter, point estimates are often necessary for comparisons with frequentist methods or for simplifying the model application. Several methods exist to extract point estimates from the posterior [40], with the mean, median, and maximum a posteriori (MAP) being the most commons. The posterior mean is simply the expected value of the posterior distribution, and it provides a sensible, representative estimate of the distribution's central location. The median, often close to the mean, is the point where 50% of the probability mass lies on either side, and it can be indicative of the posterior distribution's center and is preferable in cases where the mean is heavily skewed. The MAP estimator, corresponding to the highest point in the posterior (also known as the posterior mode), is straightforward to calculate as it depends solely on the numerator of Bayes' rule. However, as illustrated in Figure 3.7, it can be far from the majority of the probability mass, making it a less desirable option when considering uncertainty.

$$E[\theta|Y] = \int \theta P(\theta|Y) d\theta \qquad \text{(Mean)} \tag{3.5}$$

$$m \mapsto \int_0^m P(\theta|Y)d\theta = \frac{1}{2} \qquad \text{(Median)} \tag{3.6}$$

$$\arg\max_{\theta} f(\theta \mid Y) \qquad \text{(MAP)} \tag{3.7}$$

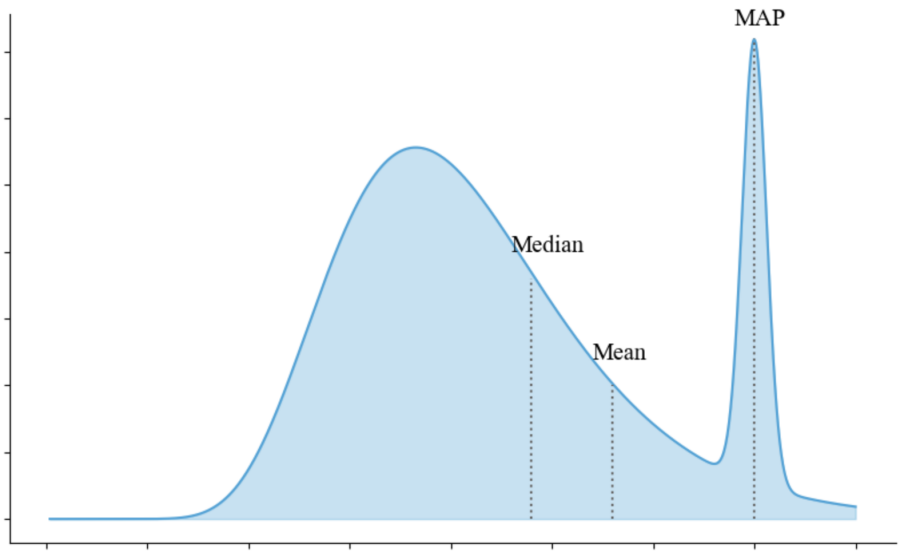

Figure 3.7: Multimodal posterior distribution.

3.2.3 Markov Chain Monte Carlo

A notable practical challenge in Bayesian inference is that obtaining analytical or exact solutions is often unattainable, except for the simplest models. In previous sections, we discussed the fundamental components of Bayes' theorem: the user-specified prior beliefs $P(\theta)$, the likelihood function which quantifies the probability of the observed data given the model parameters $P(Y|\theta)$, and the denominator $P(Y)$ that serves as a normalizing constant to transform the numerator into a valid probability distribution, ultimately resulting in the desired posterior distribution. While the calculation of the last term follows a specific rule, it frequently represents a highly intricate and computationally demanding operation, as it relies on computing the marginal likelihood, which entails integration over multiple variables. This term is alternatively referred to as Evidence, which can be mathematically expressed as:

$$P(Y) = \int_i P(Y|\theta_i)P(\theta)d\theta_i \tag{3.8}$$

Here, i denotes the index for the set of parameters in the likelihood function.

Considering the significant computational cost and frequent intractability of this problem, various computational methods have been developed to approximate the posterior distribution without the need to compute $P(Y)$. These approximations rely solely on the numerator of Bayes' rule, thereby simplifying the problem by circumventing the integral entirely. One approach is

Variational Inference [29], which transforms the Bayesian inference problem into an optimization problem by identifying an approximate distribution that minimizes the divergence from the true posterior. However, this method has notable disadvantages, as it relies on combining simple distributions to simulate the posterior, and is unable to approximate highly complex distributions effectively.

As an alternative to Variational Inference, one can sample directly from the distribution. If unbiased samples from the posterior distribution can be readily obtained, the mean and variance can be extracted and utilized for estimation purposes. Regrettably, sampling from a complex and unknown distribution is not always straightforward. Given that we possess an unnormalized probability distribution, conventional sampling algorithms for common distributions, such as exponential or Gaussian, are inadequate. Consequently, we resort to advanced sampling methods specifically designed to efficiently explore the parameter space and generate correlated samples from the posterior distribution. Markov Chain Monte Carlo (MCMC) encompasses a distinct set of sampling methods for approximating a distribution without requiring the exact probability density function [27].

These approaches combine two fundamental concepts: Monte Carlo methods and Markov chains. The Monte Carlo aspect refers to the process of random sampling, which involves repeatedly drawing samples from a probability distribution to approximate the desired characteristics. For instance, if one aims to estimate the probability of obtaining heads when flipping a coin, the Monte Carlo approach would involve flipping the coin numerous times and using the resulting observations to make predictions. On the other hand, Markov chains are mathematical models that describe the behavior of sequences of events or states, where each state is connected by links representing the transition probabilities between them. The defining feature of a Markov chain is that the next state in the sequence solely depends on the current state, rather than the entire history of states. MCMC algorithms combine these two concepts by constructing a Markov chain whose stationary distribution is the target probability distribution of interest. By iteratively drawing samples from the chain, MCMC methods enable efficient exploration of the parameter space and approximation of complex probability distributions.

Metropolis-Hastings Algorithm

Building upon the MCMC methods, the Metropolis-Hastings (MH) algorithm [41] is a popular technique widely used in Bayesian inference. It allows for an efficient approximation of the posterior distribution while circumventing the need to compute the normalizing factor. The algorithm generates a Markov chain whose stationary distribution converges to the posterior distribution, only requiring the computation of the unnormalized posterior itself and avoiding the computationally expensive evidence calculation. Thus, as we only know the numerator of Bayes' rule, we can represent it as a distribution F, which is proportional to the desired posterior probability distribution P:

$$p(\theta) \propto f(\theta) \tag{3.9}$$

The Metropolis-Hastings algorithm operates by selecting an arbitrary starting point θ_0 in the parameter space, as well as a proposal distribution Q that is easy to sample from, such as a Gaussian or uniform distribution centered around the current state. Iteratively, the algorithm draws candidate states θ^* from the proposal distribution and calculates an acceptance probability α, based on the ratio of unnormalized posterior probabilities and proposal probabilities:

$$\begin{aligned} \alpha &= \frac{f(\theta^*)/Q(\theta^*|\theta_{i-1})}{f(\theta_{i-1})/Q(\theta_{i-1}|\theta^*)} \\ &= \frac{f(\theta^*)Q(\theta_{i-1}|\theta^*)}{f(\theta_{i-1})Q(\theta^*|\theta_{i-1})} \end{aligned} \tag{3.10}$$

The candidate state is accepted or rejected by the following rule:

$$\theta_i = \begin{cases} \theta^*, & \text{with probability } \min(\alpha,1) \\ \theta_{i-1}, & \text{otherwise} \end{cases} \tag{3.11}$$

This rule act as a correction factor, since the proposal distribution Q is not the target distribution.

After running the algorithm for a sufficient number of iterations, the Markov chain might converges to the stationary distribution, which corresponds to the desired posterior distribution. To ensure convergence, it is advisable to run multiple chains with distinct initial states, allowing for a more effective assessment of convergence. If all chains converge to the same distribution, despite starting from different points, it provides evidence that the chains have indeed reached the target posterior distribution. For each chain, a burn-in period (iterations at the beginning of the MCMC run) is usually discarded to account for potential dependence on the initial state θ_0, and the remaining samples can be used to approximate various properties of the target posterior distribution, such as the mean, variance, or other summary statistics.

3.3 Bayesian Logistic Regression

Logistic regression is a widely-used statistical method for modeling the relationship between a binary outcome variable (target) and one or more predictor variables (features). It is a type of generalized linear model (GLM) that extends linear regression to classification problems [6], where the response variable represents two distinct categories (e.g., success/failure, yes/no, 0/1). The logistic regression model estimates the probability of the outcome variable belonging to a particular class, given the values of the predictor variables. It does so by employing the logistic function (also known as the sigmoid function), which maps the output of a linear combination of predictor variables to a probability value between 0 and 1.

Suppose we aim to use the Bayesian framework to model the occurrence of a specific event,

which can be represented as a binary variable y (Event = 1 or Non-event = 0), with corresponding probabilities p and $(1-p)$. The logistic regression for such scenario can be defined as [42]:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \tag{3.12}$$

Here, $X = (x_1, x_2, \cdots, x_i)$ denotes a vector of independent explanatory variables (features), $\beta = (\beta_0, \beta_1, \cdots, \beta_i)$ represents a vector of unknown regression parameters of the model, and $i = 1, 2, \cdots, k$ refers to the number of features. Consequently, the predicted value of y can be derived from the above equation as follows:

$$P(y=1) = \frac{\exp(\beta_1 + \beta_2 x_1 + \cdots + \beta_i x_i)}{1 + \exp(\beta_1 + \beta_2 x_1 + \cdots + \beta_i x_i)} \tag{3.13}$$

$$\text{Where, } y = \begin{cases} 1; & \text{Event} \\ 0; & \text{Non-event} \end{cases}$$

In the prior sections, θ was employed as a placeholder for a set of generic model parameters. Here, we're focusing on logistic regression, so we've switched to the use of β to denote its particular parameters.

Equation 3.13 can alternatively be expressed as:

$$P = \frac{\exp(\beta X)}{1 + \exp(\beta X)} \tag{3.14}$$

Given that y can be represented as the outcome of a Bernoulli experiment with a probability of success provided by Equation 3.13 [26], the likelihood function for n training samples is defined as:

$$L(y|\beta, X) = \prod_{j=1}^{n} P_j^{y_j} (1 - P_j)^{1-y_j} \tag{3.15}$$

In classical statistical inference, the logistic regression parameter vector β for the aforementioned model can be estimated using the Maximum Likelihood Estimation (MLE) method [43]:

$$\sum_{j=1}^{n} \left[y_j \log(P_j) + (1 - y_j) \log(1 - P_j) \right] \tag{3.16}$$

However, to express the uncertainty for each parameter β, Bayesian methods can be employed to estimate the model parameters of this logistic regression model. To perform Bayesian inference for the unknown parameters $\beta_1, \beta_2, \cdots, \beta_i$, a prior distribution must be defined for each model parameter. For simplicity, let's assume a non-informative Gaussian prior distribution for each parameter with mean μ and variance σ^2.

$$\beta_j \sim Normal(\mu_j, \sigma_j^2) \tag{3.17}$$

$$P(\beta_j|\mu_j,\sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\beta_j-\mu_j)^2\right] \tag{3.18}$$

As discussed in previous sections, the posterior probability distributions can be obtained by multiplying the prior distribution with the complete likelihood function. Consequently, the posterior distribution of the unknown parameters β_j for Bayesian logistic regression with a Gaussian prior is given by [44]:

$$\begin{aligned} P(\beta_j|y,X) &\propto L(y|\beta,X) \times P(\beta_j) \\ &\propto \left[\prod_{j=1}^{n} P_j^{y_j}(1-P_j)^{1-y_j}\right] \\ &\quad \times \left[\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\beta_j-\mu_j)^2\right)\right] \\ &\propto \prod_{j=1}^{n} \left[\left(\frac{\exp(\beta X)}{1+\exp(\beta X)}\right)^{y_j} \left(1-\frac{\exp(\beta X)}{1+\exp(\beta X)}\right)^{1-y_j}\right] \\ &\quad \times \left[\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\beta_j-\mu_j)^2\right)\right] \end{aligned} \tag{3.19}$$

In summary, Equation 3.19 represents only the numerator of Bayes' theorem in the context of Bayesian logistic regression. Nevertheless, the denominator of the theorem, which corresponds to the marginal likelihood, would be expressed by a more complex equation that may be computationally intractable. To overcome this challenge and approximate the posterior probability distribution of each model parameter, sampling techniques such as the Metropolis-Hastings algorithm can be employed. This powerful and versatile approach allows for efficient exploration of the parameter space and accurate estimation of the posterior distributions.This configuration, which combines the strengths of Bayesian inference and MCMC sampling methods, will be utilized throughout the experiment developed in this work, demonstrating its applicability and comparing its effectiveness with other uncertainty quantification techniques.

4

Methodology

In this chapter, we detail the procedures employed to evaluate the performance of Bayesian Inference for uncertainty quantification in the context of machine learning models applied to scenarios with limited data. As the primary focus of this monograph, Bayesian Inference is thoroughly examined, and its performance is compared with two other uncertainty quantification approaches — Conformal Inference [21] and Frequentist Uncertainty Region [25]. The objective of the experiments conducted is to assess the efficacy of Bayesian Inference in extracting uncertainty estimates in conjunction with machine learning predictions from a model, as well as to determine its suitability in comparison to the other techniques in a propabilistic way.

For this, five public datasets from various sources are utilized, each containing different features and targets suitable for binary classification tasks in the context of small data. To facilitate a comparison of the performance of each approach, the machine learning metric F1-score is used, and a cross-validation strategy is employed to ensure the fairness and robustness of the results. The comparison of each uncertainty quantification approach is conducted by examining the top quantiles of samples, ranked based on their uncertainty output. Specifically, we focus on the 25th, 50th, and 75th percentiles, which represent the most confident predictions made by each technique. By analyzing these distinct quantiles, we can gain insights into how each method performs at different levels of uncertainty and determine their effectiveness in separating reliable predictions from uncertain ones. In this study, we refer to these groups as the first (Q1), second (Q2), and third (Q3) quartiles of the ranked samples, where each quartile corresponds to the 25%, 50%, and 75% most confident predictions, respectively.

Subsequent subsections will cover the data and preprocessing steps, provide an overview of the implementation and parameter details for each uncertainty quantification approach, expound on the uncertainty output transformation, and elucidate the evaluation and cross-validation strategy employed. The primary goal is to assess the effectiveness of Bayesian Inference in this

scenario and determine its advantages and drawbacks over the other approaches for uncertainty quantification.

4.1 Datasets

The datasets selected for this study have been carefully chosen to represent a diverse range of domains and feature compositions, which include a mix of categorical and numerical attributes. This diversity allows us to thoroughly evaluate the performance of the uncertainty quantification approaches under consideration, specifically in binary classification tasks. By using datasets with varying complexities and characteristics, we can gain a comprehensive understanding of the strengths and weaknesses of each uncertainty quantification approach, as well as their applicability to a wide array of real-world problems.

Heart Failure [45]: This dataset is the most complete heart disease dataset available for research purposes, created by combining five individual datasets (Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog) with a total of 918 observations. The dataset contains 11 independent variables, including five categorical and six numerical features. The primary goal of this dataset is to predict heart disease events based on the given clinical features.

QSAR Biodegradation [46]: The QSAR dataset is comprised of 1055 chemicals, classified into two classes: ready and not ready biodegradable. It contains 41 numerical attributes (molecular descriptors) that were used to develop QSAR (Quantitative Structure-Activity Relationships) models. This dataset allows for the study of relationships between chemical structure and biodegradation of molecules.

Australian Credit Approval (Statlog) [47]: This dataset concerns credit card applications, consisting of 690 samples with 6 numerical and 8 categorical attributes. The dataset provides a good mix of attribute types, including continuous and nominal variables. The objective is to predict credit approval based on the provided attributes.

Pima Indians Diabetes [48]: This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases, and it is focused on predicting diabetes in female Pima Indian patients at least 21 years old. It contains 780 samples with 8 numerical features, including several medical predictor variables and one target variable, Outcome.

Titanic [49]: The Titanic dataset includes data for 887 real Titanic passengers, with each row representing one person. The dataset contains 10 features that describe different attributes about the person, including whether they survived, their age, their passenger-class, their sex, and the fare they paid. The objective is to predict passenger survival using machine learning.

A summary of the datasets is presented in Table 4.1, which provides information on the number of samples, features, and feature types for each dataset. These datasets consist of a mix of numerical and categorical input variables. In order to prepare the data for analysis, we applied various preprocessing techniques, including ordinal encoding and one-hot encoding for

categorical features, as well as treatment of null values.

Dataset	Samples	Features	Feature Types	Positive Class %
Heart Failure	918	11	5 categorical, 6 numerical	55%
QSAR Biodegradation	1055	41	41 numerical	33%
Australian Credit	690	14	6 numerical, 8 categorical	44%
Pima Indians Diabetes	780	8	8 numerical	34%
Titanic	887	10	4 categorical, 6 numerical	38%

Table 4.1: Summary of datasets used in the experiments.

For ordinal categorical features, ordinal encoding was applied to preserve the ordinal information present in these features. This encoding method assigns an integer value to each category in a way that the order of the categories is maintained. This is particularly useful for categorical variables with intrinsic ordering, such as the levels of chest pain type and the slope of the peak exercise ST segment in the Heart Failure dataset. For nominal categorical features, one-hot encoding was applied to ensure that the relationships between categories are not misrepresented as ordinal. One-hot encoding creates binary features for each unique category of a given feature, with each binary feature indicating the presence or absence of that category. This method is appropriate for categorical variables without intrinsic ordering, such as the 'sex' feature in the Titanic dataset, where the categories 'male' and 'female' should be treated as distinct and unrelated.

Regarding the treatment of null values, we filled the missing values with the median value for numerical features and with the mode for categorical ones. This approach helps to minimize the impact of missing data on the performance of the algorithms, as it replaces the missing values with a representative central tendency measure. This method is commonly used in machine learning applications where missing data is inevitable, but it is crucial to maintain the integrity of the dataset for accurate analysis and model performance. For the numerical features, no rescaling or standardization was performed, as it would not provide any significant benefits for the analysis in this context.

It is important to note that all three approaches being compared in this study utilize the same algorithm definition and the same datasets. Consequently, the data preprocessing steps applied are consistent across all methods, ensuring that any potential benefits or drawbacks resulting from these steps would affect all approaches equally. This consistency in data preprocessing is essential for a fair comparison of the performance of the different approaches to uncertainty estimation.

4.2 Bayesian Inference Experiment Setup

In the Bayesian Inference approach, the choice of prior distribution plays a crucial role in the analysis. For this study, non-informative priors are chosen to avoid introducing any potential

biases in the estimation process. Non-informative priors are designed to have minimal influence on the posterior distribution, allowing the data to drive the conclusions. This choice is particularly appropriate when there is no strong prior knowledge or belief about the parameters being estimated. Based on the findings discussed by Rasines and Young [50], uniform distributions are a suitable representation for non-informative priors, as they have small influence on the posterior distribution. We use uniform distributions for all parameters in our Bayesian Logistic Regression model.

The Metropolis-Hastings algorithm, a popular MCMC sampling method, is employed to obtain samples from the posterior distribution of the model parameters. This algorithm is well-suited for Bayesian Inference problems with complex and high-dimensional parameter spaces. To ensure a thorough exploration of the parameter space and reduce the risk of convergence issues, four independent Markov chains are run in parallel, each with 25,000 iterations. These settings were chosen based on empirical evidence and recommendations in the literature [51], which suggest that using multiple chains and a sufficiently large number of iterations can improve the reliability of the results. Running multiple chains enables a more accurate assessment of the convergence of the MCMC algorithm and helps obtain reliable results. Convergence diagnostics, such as the Gelman-Rubin statistic [52], can be applied to evaluate the performance of the chains and ensure that they have adequately explored the posterior distribution.

To quantify uncertainty in the predicted posterior probability, the width of the highest density interval (HDI) is calculated, and the mean value of the distribution is taken as the predicted score. The HDI is a credible interval containing a specified proportion of the posterior probability mass and is the shortest interval containing that proportion. In this study, the HDI is computed to represent 95% of the posterior distribution. By measuring the width of the HDI, an estimate of the uncertainty level can be obtained, which is useful for comparing the performance of different approaches to uncertainty quantification. The entire Bayesian approach is implemented using some modules from the PyMC Python library [53], a powerful and flexible framework for Bayesian modeling and probabilistic programming. PyMC provides an intuitive interface for defining models, specifying prior distributions, and running MCMC algorithms, such as the Metropolis-Hastings. This implementation allowed for a robust and efficient analysis of the Bayesian Inference approach in this study.

4.3 Frequentist Experiments Setup

In the frequentist experiment, two methods were employed to quantify the uncertainty of individual predictions: Conformal Inference and Uncertainty Region, both of which have been previously discussed in the literature review. These frequentist methods were selected because they provide a straightforward way to compute uncertainty in the form of a binary label, indicating whether a prediction is reliable based on some predefined parameters. Alternative methods,

such as model calibration also explored in the literature review, do not provide a simple way to interpret the uncertainty in individual predictions, so they were not considered in this study.

Conformal inference provides a method to construct prediction sets for new observations, which have a guaranteed coverage probability under some mild assumptions. In classification settings, conformal inference uses nonconformity scores to construct prediction sets that have the desired level of confidence $(1-\alpha)$, where α is a user-specified parameter representing the error tolerance. For this study, the Inverse Probability Nonconformity Measure was used, this measure returns the value of $1-p$, where p represents the probability assigned to the actual class by the underlying classification model. This choice is suitable for classification problems, as it takes into account the confidence of the model in assigning an instance to a particular class. As this is a binary classification problem, there are four possible conformal sets for a single test sample [54]: {positive}, {negative}, {positive, negative}, and { }. We will call the predictions with sets of length 1 as "reliable". Note that the parameter α directly affects the sets' length: the smaller α is, the less tolerance we allow, and the prediction set has a higher chance of containing the true label.

In the uncertainty region approach, we followed the specifications outlined in the original paper where the technique was proposed [25], as it has already been applied to a logistic regression model for a binary classification problem. As discussed earlier, the uncertainty region approach constructs prediction intervals for individual scores by calculating the standard deviation of the predictions. Upon computing the standard deviation for a score prediction, a scaling parameter k is applied to determine the final interval. The original paper does not provide a specific method for choosing the parameter k; however, the authors recommend that it should not be too large, which would make all intervals unreliable, nor too short, which would make the model overconfident. Besides quantifying the spread of the values of $p(x)$ around the predicted value, the uncertainty region approach introduces the concept of classifying each prediction as either reliable or unreliable. If the random estimate is within the predicted interval, the score is considered too close to the random value and should be treated as unreliable.

For both the conformal inference and the uncertainty region approach, the respective parameters α and k were adjusted to obtain different sizes of uncertainty groups. This adjustment enables a direct comparison of the methods' performance across various levels of uncertainty. The simplicity of these methods allows for straightforward implementation without the need for external libraries. As a result, the comparative analysis of these methods provides valuable insights into their suitability and effectiveness when compared to Bayesian Inference. The implementation details, parameter settings, and source code utilized for the experiments in this study are publicly accessible for further scrutiny and reuse. The relevant resources can be found in the dedicated GitHub repository at the following web address: https://github.com/WagnerFLL/bayesian-uncertainty-quantification/.

4.4 Evaluation

Comparing the performance of the three uncertainty quantification approaches in binary classification tasks is inherently challenging. The key difficulty lies in the distinct output types for each method: posterior distributions for Bayesian Inference, prediction sets for Conformal Inference, and uncertainty intervals for the Uncertainty Region. To facilitate a straight comparison, we transform the uncertainty output of each method into a binary variable, indicating whether a prediction is deemed reliable. By adjusting the parameters for each method (interval width, α, and k), we can obtain different proportions of "reliable" predictions. The primary comparison is conducted by selecting equally sized reliable groups from each method and examining their performance using the F1 score. This approach enables us to assess the efficacy of each method in distinguishing reliable predictions from uncertain ones.

Three comparisons are conducted to evaluate the performance of the uncertainty quantification approaches, focusing on reliable groups that represent the 1st, 2nd, and 3rd quantiles (top 25%, 50%, and 75%) of the most reliable predictions. By examining these subsets of predictions, we can gain a deeper understanding of how well each method performs when handling varying levels of uncertainty in tasks with significantly small datasets. This comparative analysis is critical for determining the best approach to use in practice, given specific requirements and desired levels of confidence.

Beyond the primary comparison of reliable groups, we also assess the overall performance of the logistic regression models fitted using each method. While this is not the main objective of our study, it is essential to understand the impact of the uncertainty quantification methods on the models' overall performance and predictive power. By comparing the performance of the models using the F1 score, we can identify potential trade-offs between the methods and make more informed decisions about which approach is best suited for a given task. Additionally, it allows us to examine how the methods impact the learning process of the algorithm, shedding light on their influence on the underlying logistic regression models.

To ensure a fair comparison, a stratified k-fold cross-validation strategy with k = 5 was employed. This method maintains the proportion of class labels in each fold, ensuring a balanced representation of the data throughout the evaluation process. In each iteration, the logistic regression model was trained on the training set, and predictions were subsequently generated for the test set. It is important to note that the conformal inference technique requires an additional portion of data for calibration purposes. In this study, the auxiliary calibration set is extracted from the training data, effectively reducing the number of available samples for model training. This approach is taken to ensure that the calibration data remain unseen by the model and are not included in the test samples, thus preserving the integrity of the evaluation process.

To assess the performance of the models and compare their efficacy, we used the F1 score. F1 score is a harmonic mean of precision and recall, offering a balanced measure of a model's performance in terms of both false positives and false negatives. Precision is the proportion of

true positive predictions among all positive predictions, while recall is the proportion of true positive predictions among all actual positive instances. The F1 score is particularly useful when there is an imbalance in the class distribution or when it is necessary to give the same importance to both precision and recall. The F1 score ranges from 0 to 1, with 1 indicating perfect precision and recall, and 0 signifying the worst possible performance.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{4.1}$$

By evaluating the performance of the models using the F1 score, we can obtain a deeper understanding of how well each approach is able to separate reliable predictions from uncertain ones. This, in turn, allows us to compare the effectiveness of the Bayesian Inference, conformal inference, and uncertainty region approaches in providing meaningful uncertainty quantification when dealing with limited data across multiple domains.

5

Results and Discussion

In this chapter, we delve into the analysis and interpretation of the experimental results obtained from our study, focusing primarily on the Bayesian Inference approach. Our aim is to provide a comprehensive understanding of the convergence properties of the Bayesian models, compare the performance of the three Uncertainty Quantification methods, and explore the relationship between the predicted Bayesian posterior distributions and model uncertainty. This discussion will shed light on the strengths and limitations of the Bayesian Inference method and its practical implications in binary classification tasks with limited data.

The chapter is organized as follows: First, we examine the convergence and mixing properties of the Markov chains for the Bayesian models, including trace plots, autocorrelation plots, and convergence diagnostics. Next, we compare the performance of the Bayesian Inference, Uncertainty Region, and Conformal Inference approaches in terms of F1-scores for various subsets of predictions.

5.1 Convergence of Bayesian Models

The convergence diagnostics help us understand the reliability of the posterior estimates obtained from the MCMC sampling process, in this study, employed using the Metropolis Algorithm. To perform this evaluation, we consider the final models obtained after the cross-validation procedure. To assess convergence, we examine the effective sample size (ESS) and the Gelman-Rubin R-hat statistic for each parameter in the models. ESS measures the number of effectively independent draws from the posterior distribution, while the R-hat statistic compares the within-chain and between-chain variance to detect non-convergence.

For some models, we observe that the ESS is smaller than 200 for certain parameters, which may indicate potential issues with the mixing of the MCMC sampler, suggesting that the sampler could be struggling to explore the parameter space efficiently [55]. However, for all models,

we have at least 100 effective samples, which is a reasonable number for our purposes. It is important to note that having a few parameters with relatively low ESS might not be detrimental if those parameters are not our primary interest or if they represent a small proportion of the total parameters in the model.

R-hat helps assess whether the chains have converged to the same target distribution or if they are still exploring different regions of the parameter space. An R-hat value close to 1 indicates that the between-chain and within-chain variances are similar, suggesting that the chains have converged to the same distribution. In practice, an R-hat value below 1.1 is generally considered acceptable, indicating good convergence [51]. However, a more stringent threshold of 1.05 is often recommended for better confidence in the convergence assessment. In our analysis, we find some parameters with R-hat values between 1.05 and 1.1, as well as a few with values greater than 1.1. These findings are summarized in the table 5.1, which reports the number of parameters with ESS smaller than 200 and the number of parameters with R-hat values in various ranges for each dataset.

Table 5.1: Summary of convergence diagnostics for each dataset.

Dataset	ESS < 200	$\hat{r} < 1.05$	$1.05 \leq \hat{r} \leq 1.1$	$\hat{r} > 1.1$	Total Parameters
Heart	1	9	2	1	12
Australian	1	11	2	2	15
Titanic	3	6	4	1	11
Diabetes	4	5	1	3	9
Qsar	6	33	4	5	42

The columns in Table 5.1 are as follows:

- ESS < 200: Number of parameters with effective sample size less than 200.
- $\hat{r} < 1.05$: Number of parameters with an R-hat statistic less than 1.05.
- $1.05 \leq \hat{r} \leq 1.1$: Number of parameters with an R-hat statistic between 1.05 and 1.1.
- $\hat{r} > 1.1$: Number of parameters with an R-hat statistic greater than 1.1.
- Total Parameters: Total number of parameters in the model.

It is worth noting that convergence issues are not uncommon when dealing with small datasets or more complex models. There are several possible solutions to address or mitigate these convergence problems, such as increasing the number of sampling iterations, changing the sampling algorithm, or incorporating more informative prior distributions. When informative priors are used, they can provide additional information to guide the sampling process and lead to better convergence properties [56]. However, it is essential to be cautious with the choice of informative priors, as they may introduce bias if the prior information is not accurate or representative of the true underlying process [57].

Adding informative priors can be particularly helpful in situations where the data alone may not provide enough information for the model to converge. One such example is when the dataset is small, and the model struggles to capture the complexities in the data. In these cases, the incorporation of informative priors can act as a form of regularization, guiding the model towards more plausible parameter values based on prior knowledge or expert opinion [56].

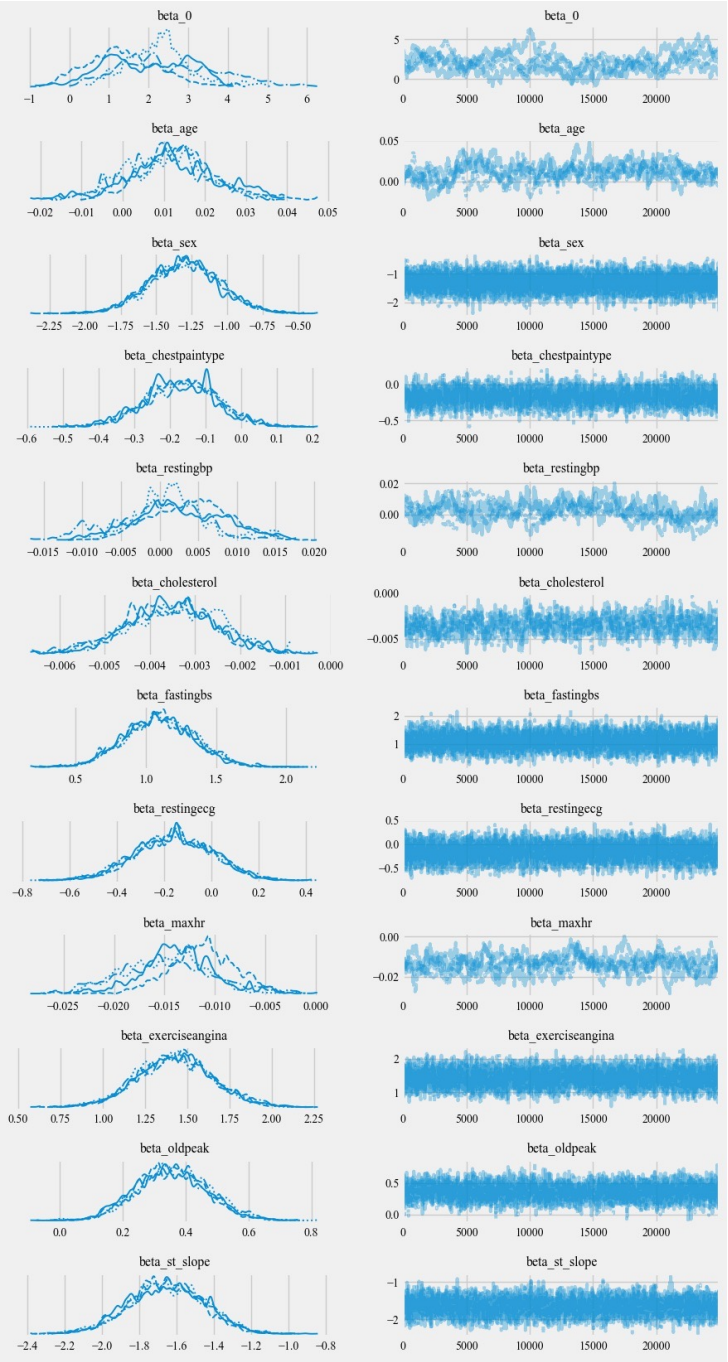

Figure 5.1: Trace plot for the heart failure dataset.

Despite these potential convergence issues, a closer inspection of the trace plots for the heart failure dataset (Figure 5.1) reveals that most of the features show good convergence among

the four Markov chains. The intercept value (beta_0) of the logistic regression appears to be the most problematic distribution, possibly having the highest R-hat value. By examining the autocorrelation graph of this variable, we observe a high autocorrelation, indicating that the sampler is struggling to sample its distribution efficiently.

It is essential to recognize that even with some convergence warnings, the Bayesian models still provide valuable information and insights. The convergence issues are not necessarily an indication of the models being entirely useless or unreliable. Instead, the concerns about the convergence will be converted into uncertainty, reflecting into the predicted posterior distribution. This feature allows the Bayesian models to capture both epistemic and aleatoric uncertainty and make more useful posterior distributions. As a result, the Bayesian models offer a robust framework for Uncertainty Quantification, which can be compared and contrasted with alternative approaches in the subsequent sections. The ability to account for uncertainty in the models enables more informed decision-making and improves the overall understanding of the underlying phenomena.

5.2 Comparison of Approaches

In this section, we compare the performance of the Bayesian Inference approach with the Uncertainty Region and Conformal Inference approaches. Table 5.2 provides a summary of the results for each dataset, showcasing the F1 score and the standard deviation across five distinct cross-validation steps. This table offers a comprehensive view of the performance consistency and variability of each approach across multiple iterations of the validation process.

Dataset	Bayesian Inference	Uncertainty Region	Conformal Inference
Heart Failure			
Overall	**0.857 ± 0.04**	0.856 ± 0.04	0.854 ± 0.04
Q1	**0.960 ± 0.01**	0.936 ± 0.02	0.952 ± 0.02
Q2	**0.935 ± 0.03**	0.917 ± 0.03	0.929 ± 0.02
Q3	0.900 ± 0.04	0.903 ± 0.04	**0.909 ± 0.03**
Australian Credit Approval			
Overall	**0.849 ± 0.02**	0.838 ± 0.02	0.835 ± 0.03
Q1	0.946 ± 0.08	**0.969 ± 0.03**	0.922 ± 0.05
Q2	0.929 ± 0.03	**0.934 ± 0.01**	0.930 ± 0.04
Q3	**0.912 ± 0.01**	0.911 ± 0.02	0.909 ± 0.01
Titanic			
Overall	**0.727 ± 0.04**	0.726 ± 0.04	0.721 ± 0.05
Q1	0.897 ± 0.03	**0.932 ± 0.04**	0.900 ± 0.07
Q2	0.845 ± 0.04	**0.864 ± 0.03**	0.863 ± 0.04

Q3	0.790 ± 0.06	**0.820 ± 0.04**	0.810 ± 0.04
Pima Diabetes			
Overall	0.633 ± 0.01	**0.635 ± 0.01**	0.630 ± 0.03
Q1	0.832 ± 0.04	**0.863 ± 0.07**	0.856 ± 0.06
Q2	0.801 ± 0.04	**0.822 ± 0.03**	0.812 ± 0.04
Q3	**0.743 ± 0.03**	0.740 ± 0.02	0.723 ± 0.02
Qsar			
Overall	0.790 ± 0.02	**0.793 ± 0.02**	0.780 ± 0.04
Q1	0.944 ± 0.05	0.941 ± 0.05	**0.957 ± 0.06**
Q2	**0.939 ± 0.05**	0.924 ± 0.05	0.926 ± 0.03
Q3	0.877 ± 0.04	**0.882 ± 0.03**	0.862 ± 0.02

Table 5.2: F1 scores. Comparison of Bayesian Inference, Uncertainty Region, and Conformal Inference Approaches

The overall metric of the Bayesian Inference approach is generally close to or slightly better than the Uncertainty Region and Conformal Inference approaches. For example, the overall F1 score for the Heart dataset is 0.857 ± 0.04 for the Bayesian Inference, while the Uncertainty Region and Conformal Inference approaches have scores of 0.856 ± 0.04 and 0.854 ± 0.04, respectively. This trend is consistent across all datasets.

When comparing the performance at different quantiles, the Bayesian Inference approach exhibits comparable or better performance than the other two approaches. For instance, at the 25% quantile, the Bayesian Inference approach yields better performance for the Heart and Diabetes datasets, while the Uncertainty Region approach performs better for the Australian Credit and Titanic datasets, and the Conformal Inference approach is better for the Qsar dataset. At the 50% and 75% quantiles, the performance differences between the approaches are generally small, with no approach consistently outperforming the others.

An intriguing observation from our analysis is that the overall performance of the Conformal Inference method consistently falls short of surpassing that of the Uncertainty Region method. A plausible explanation for this outcome is the requirement of a calibration set in the Conformal Inference method, which is extracted from the training data, thereby reducing its size. Given that both methods employ the same logistic regression implementation and identical hyperparameters, the reduced training set size in the Conformal Inference approach could be a contributing factor to its comparatively lower performance.

Another aspect worth considering is the flexibility of the output generated by each method. In order to produce different certainty groups for Conformal Inference, it is necessary to vary the alpha parameter while running predictions, which renders the uncertainty estimation less flexible since it is presented as a prediction set. The Uncertainty Region method encounters

a similar issue, as adjusting the K parameter necessitates examining the probability interval. Conversely, Bayesian Inference provides a more versatile output in the form of a distribution that can be applied in numerous ways. This adaptability enables a more comprehensive analysis of the results and facilitates well-informed decision-making based on the information at hand.

In summary, the Bayesian Inference approach demonstrates competitive performance when compared to the Uncertainty Region and Conformal Inference. While there are some variations in performance at different quantiles, the overall F1 score of the Bayesian Inference is generally comparable or slightly better than the other two options. The results indicate that the Bayesian Inference method is a viable alternative for predicting probabilities and generating prediction intervals in binary classification problems.

5.3 Strengths and Challenges of the Bayesian Method

Addressing the challenges of small datasets requires careful consideration, as conventional machine learning models are frequently susceptible to overfitting, which exacerbates predictive uncertainty. Oftentimes, traditional machine learning models do not explicitly account for this uncertainty, resulting in limited user comprehension of the inherent uncertainty embedded within the predictions. Consequently, users predominantly rely on conventional metrics to evaluate model performance. The Bayesian framework, however, intrinsically encapsulates uncertainty in the form of the predicted posterior distribution. This enables users to interpret uncertainty via diverse methods, such as extracting the Highest Density Interval (HDI), computing the standard deviation, or assessing the width of a specified credible interval.

Moreover, the uncertainty present in the posterior distribution can be systematically examined using tests applied to individual parameter distributions, such as the Gelman-Rubin test, which was previously discussed in this chapter. By conducting such tests, it becomes possible to identify sampling issues, pinpoint potentially problematic variables, and detect any adverse relationships between them. These concerns can be alleviated, to a certain extent, by incorporating additional data. Notably, Bayesian models are well-suited for sequential learning—an advantage when dealing with small datasets. As more data is acquired, the model can be iteratively updated, offering a flexible and adaptive approach to learning from limited data.

The incorporation of prior knowledge represents a key advantage of the Bayesian approach. By integrating this prior information, the model is not solely dependent on the small dataset, which aids in alleviating issues arising from data scarcity. This integration serves as a natural regularization mechanism, as it assists in preventing overfitting, a prevalent problem when working with small datasets. The prior distribution encourages the model to discern more general patterns, rather than merely memorizing noise within the training data. Furthermore, this method enables the utilization of expert knowledge, capitalizing on existing understanding of the task at hand. Nonetheless, the selection of an appropriate prior distribution presents a con-

siderable challenge, particularly in the context of small datasets. An overly informative prior may introduce bias into the model's predictions, while a weakly informative prior might not offer substantial advantages. Additionally, the availability of domain experts who possess specialized knowledge pertaining to the specific task is not always a guarantee in the machine learning domain. Consequently, striking the right balance when choosing a prior distribution is critical to ensuring optimal performance and mitigating potential pitfalls.

While the Bayesian method offers several notable advantages, it also comes with its own set of limitations. One of the primary drawbacks is its computational complexity. Table 5.3 illustrates the total execution time for each complete experiment conducted on the various datasets, highlighting the increased computational demands associated with the Bayesian approach. The Qsar dataset, which contains the largest number of independent variables, exhibits an even longer execution time. This underlines the challenges faced by the Bayesian algorithm, particularly the Metropolis-Hastings sampling method, as the number of variables increases. In contrast, the frequentist approach to constructing a logistic model is predicated on maximum likelihood estimation, which can be expedited by employing intelligent optimization techniques such as Stochastic Gradient Descent (SGD), as utilized in our study. As a result, the frequentist method proves to be substantially more efficient in comparison to the Bayesian approach, which necessitates the use of robust sampling methods to achieve accurate results. This difference in computational demands should be taken into consideration when selecting an approach for handling uncertainty in logistic regression tasks, particularly when dealing with large-scale or high-dimensional datasets.

Dataset	**Method**		
	Uncertainty Region	**Conformal Inference**	**Bayesian Inference**
Heart	1.09 s	1.18 s	6min 50s
Titanic	1.1 s	1.27 s	6min 21s
Australian	1.32 s	1.4 s	8min 19s
Diabetes	738 ms	846 ms	4min 47s
Qsar	6.17 s	6.29 s	37min 30s

Table 5.3: Execution time of Bayesian Inference, Conformal Inference, and Uncertainty Region experiments across all datasets.

An additional limitation of the Bayesian method is its sensitivity to model misspecification. This issue can be particularly pronounced when dealing with small datasets, as the limited amount of data available may not be sufficient to detect and rectify instances of model misspecification, potentially leading to biased or unreliable predictions. Furthermore, the Bayesian framework is predominantly limited to linear models and does not readily extend to more complex or non-linear models, such as boosting trees. This constrains the applicability of the Bayesian approach to specific types of problems and datasets, potentially necessitating the exploration of alternative techniques to address the limitations inherent in linear models. As such, it is essential for practitioners to carefully weigh benefits and drawbacks of the Bayesian method

in relation to their specific problem context, taking into account the limitations with respect to computational complexity, model misspecification, and the scope of applicable model types.

6

Conclusion

As we unravel the threads of our exploration, it becomes evident that Bayesian Inference provides a robust and effective approach to handling uncertainty in binary classification problems, particularly useful when confronted with the complexities of limited data. This work has navigated the intricacies of Bayesian Inference, compared it with alternative approaches such as the Uncertainty Region and Conformal Inference, and illustrated the nuanced differences that emerge across these methods. The results from our experiments have revealed that the Bayesian method, in terms of F1 score, consistently matches or even slightly outperforms the alternative approaches across a variety of datasets. Beyond its competitive performance, the true strength of Bayesian Inference lies in its unique ability to generate posterior distributions offering a wealth of information that transcends a mere point estimate. These probabilistic outputs encapsulate the inherent uncertainty present in predictions that often arise from the noise and incompleteness of the data, and model misspecification, offering an enriched perspective enables analysis that extends beyond the commons Machine Learning metrics. This affords a level of granularity in the interpretation of predictions that is typically absent in alternative approaches, as users are not limited to a singular prediction, but are equipped with a range of plausible outcomes, each accompanied by an associated degree of belief. The versatility embedded within these distributions permits tailored interpretations, catering to specific decision-making scenarios.

Two salient attributes of Bayesian Inference emerge as particularly impactful in the domain of binary classification problems with constrained data. Primarily, the concept of Bayesian learning unfolds progressively, unlike conventional models that need retraining with the influx of new data. This trait, known as sequential learning, allows the Bayesian framework to absorb new information iteratively, making it especially advantageous under limited data conditions. As the Bayesian model receives additional data over time, it refines its understanding, leading to an enhancement in accuracy, even in the face of initially sparse data. This capacity for ongoing adaptation and improvement embodies the resilience of Bayesian Inference, allowing it to deliver dependable results despite the complexities and uncertainties tied to small data.

Secondly, Bayesian Inference breaks away from the typical data-dependent learning paradigm by weaving in prior information into the very fabric of the model. This amalgamation acts as an inherent regularization mechanism, offering a buffer against the risk of overfitting, a prevalent issue in scenarios with limited data. By leveraging this prior information, the Bayesian model is nudged towards recognizing broader patterns, instead of falling into the trap of memorizing noise within the small training dataset. Importantly, this feature opens up the possibility of harnessing expert knowledge, thus enriching the model with an established understanding of the task in focus. This ability to integrate empirical evidence with expert insights enhances the robustness of Bayesian Inference, making it a potent tool in the face of data scarcity, and enhancing its potential for generating reliable and robust predictions.

Navigating the landscape of Bayesian Inference, however, is not without its challenges. Particularly when juxtaposed with the more computationally efficient frequentist methods such as the Uncertainty Region and Conformal Inference. The intricacies and computational demands of the Bayesian method represent a significant hurdle, as it relies on computationally expansive sampling methods for approximating posterior distributions, an approach that becomes increasingly intricate and resource-intensive as data dimensionality increases. Further, while the Bayesian framework excels in its ability to accommodate prior knowledge, it also exposes a potential vulnerability. The selection of an appropriate prior distribution, particularly in the context of limited data, can be fraught with difficulty. The balancing act between overly informative and weakly informative priors is critical, as the former may inadvertently introduce bias into the model's while the latter might not offer substantial regularization benefits.

As we distill the insights gleaned from this work, the pivotal role of methodological suitability for a given task comes into sharp focus. It's not a contest of supremacy between frequentist and Bayesian methodologies, but a strategic alignment of the approach with the specific data conditions and requirements of the task at hand. Frequentist methods may shine when data is plentiful and the underlying assumptions about the data generation process are met, but Bayesian Inference, with its nuanced handling of uncertainty and its ability to incorporate prior knowledge, presents a compelling alternative when dealing with limited or noisy data.

In an era increasingly steered by data-driven decision-making, this investigation goes beyond merely advocating for Bayesian Inference; it also emphasizes that the objective should not be the eradication of uncertainty, but the understanding, quantification, and apt communication of it. The culmination of this research resonates with the promising potential of Bayesian Inference as a robust and versatile tool for Uncertainty Quantification in binary classification problems with limited data. The findings of this study provide a launchpad for further exploration and application in myriad domains, where the quest for accurate, reliable predictions is inexorably linked with a robust quantification of uncertainty. As we step into the future, we are better equipped to navigate the challenges of uncertainty in Machine Learning, armed with a deeper understanding of the strengths and pitfalls of Bayesian Inference.

Bibliography

[1] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[3] G. Apostolakis. A commentary on model uncertainty. 01 1994.

[4] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[5] Stephen M Stigler. Thomas bayes's bayesian inference. *Journal of the Royal Statistical Society: Series A (General)*, 145(2):250–258, 1982.

[6] P. McCullagh; J.A. Nelder. *Generalized Linear Models (Monographs on Statistics and Applied Probability Book 37)*. Chapman Hall/CRC, London, 1983.

[7] Chris Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419–466, 1995.

[8] Lewis H Mervin, Simon Johansson, Elizaveta Semenova, Kathryn A Giblin, and Ola Engkvist. Uncertainty quantification in drug design. *Drug discovery today*, 26(2):474–489, 2021.

[9] Younghun Jung and Venkatesh Merwade. Uncertainty quantification in flood inundation mapping using generalized likelihood uncertainty estimate and sensitivity analysis. *Journal of Hydrologic Engineering*, 17(4):507–520, 2012.

[10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[11] Vikram Mullachery, Aniruddh Khera, and Amir Husain. Bayesian neural networks. *CoRR*, abs/1801.07710, 2018.

[12] José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pages 1861–1869. PMLR, 2015.

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[14] Dae Y Kang, Pamela N DeYoung, Justin Tantiongloc, Todd P Coleman, and Robert L Owens. Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine. *NPJ digital medicine*, 4(1):142, 2021.

[15] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.

[16] Runhai Feng. Improving uncertainty analysis in well log classification by machine learning with a scaling algorithm. *Journal of Petroleum Science and Engineering*, 196:107995, 2021.

[17] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.

[18] Omar Khan, Jetan H Badhiwala, Muhammad A Akbar, and Michael G Fehlings. Prediction of worse functional status after surgery for degenerative cervical myelopathy: a machine learning approach. *Neurosurgery*, 88(3):584–591, 2021.

[19] Kenneth CY Wong, Yong Xiang, and Hon-Cheong So. Uncovering clinical risk factors and prediction of severe covid-19: A machine learning approach based on uk biobank data. *MedRxiv*, pages 2020–09, 2020.

[20] Florian Baumann, Karsten Vogt, Arne Ehlers, and Bodo Rosenhahn. Probabilistic nodes for modelling classification uncertainty for random forest. In *2015 14th IAPR International Conference on Machine Vision Applications (MVA)*, pages 510–513, 2015.

[21] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[22] Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. Predicting with confidence: using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, 2021.

[23] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

[24] Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11559–11569. PMLR, 18–24 Jul 2021.

[25] Luigi Lavazza and Sandro Morasca. Dealing with uncertainty in binary logistic regression fault-proneness models. In *Proceedings of the Evaluation and Assessment on Software Engineering*, pages 46–55. 2019.

[26] Ruihan Wang, Hui Chen, and Cong Guan. A bayesian inference-based approach for performance prognostics towards uncertainty quantification and its applications on the marine diesel engine. *ISA transactions*, 118:159–173, 2021.

[27] Henry De-Graft Acquah. Bayesian logistic regression modelling via markov chain monte carlo algorithm. *Journal of Social and Development Sciences*, 4(4):pp. 193–197, Apr. 2013.

[28] Alan E. Gelfand. Gibbs sampling. *Journal of the American Statistical Association*, 95(452):1300–1304, 2000.

[29] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[30] Yarin Gal, Petros Koumoutsakos, Francois Lanusse, Gilles Louppe, and Costas Papadimitriou. Bayesian uncertainty quantification for machine-learned models in physics. *Nature Reviews Physics*, 4(9):573–577, 2022.

[31] Pejman Honarmandi and Raymundo Arróyave. Uncertainty quantification and propagation in computational materials science and simulation-assisted materials design. *Integrating Materials and Manufacturing Innovation*, 9:103–143, 2020.

[32] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. Risk Acceptance and Risk Communication.

[33] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, mar 2021.

[34] Robin Willink and Rod White. 1-disentangling classical and bayesian approaches to uncertainty analysis. 2012.

[35] Scott M. Lynch. Bayesian theory, history, applications, and contemporary directions. In James D. Wright, editor, *International Encyclopedia of the Social Behavioral Sciences (Second Edition)*, pages 378–382. Elsevier, Oxford, second edition edition, 2015.

[36] Emmanuel Torsen. Objective versus subjective bayesian inference: A comparative study. 3:56–65, 01 2015.

[37] Wioletta Grzenda. Informative versus non-informative prior distributions and their impact on the accuracy of bayesian inference. *Statistics in Transition. New Series*, 17:763–780, 12 2016.

[38] John W. Seaman III, John W. Seaman Jr., and James D. Stamey. Hidden dangers of specifying noninformative priors. *The American Statistician*, 66(2):77–84, 2012.

[39] Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, 20(2):147–156, 1993.

[40] Harold W. Sorenson. *Parameter Estimation : Principles and Problems*. Marcel Dekker, New York, 1980.

[41] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[42] Mary Gladence L., M. Karthi, and Maria Anu. A statistical comparison of logistic regression and different bayes classification methods for machine learning. *ARPN Journal of Engineering and Applied Sciences*, 10:5947–5953, 01 2015.

[43] Lucien Le Cam. Maximum likelihood: an introduction. *International Statistical Review/Revue Internationale de Statistique*, pages 153–171, 1990.

[44] PA Lukman, S Abdullah, and A Rachman. Bayesian logistic regression and its application for hypothyroid prediction in post-radiation nasopharyngeal cancer patients. In *Journal of Physics: Conference Series*, volume 1725, page 012010. IOP Publishing, 2021.

[45] Federico Palacios. Heart failure prediction dataset. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction, 2021.

[46] Ringsted Tine Ballabio Davide Todeschini Roberto Consonni Viviana Mansouri, Kamel. QSAR biodegradation. UCI Machine Learning Repository, 2013. DOI: 10.24432/C5H60M.

[47] Ross Quinlan. Statlog (Australian Credit Approval). UCI Machine Learning Repository. DOI: 10.24432/C59012.

[48] Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association, 1988.

[49] Titanic: Machine learning form disaster. https://www.kaggle.com/competitions/titanic/data, 2017.

[50] Daniel G. Rasines and G. Alastair Young. Bayesian selective inference: Non-informative priors, 2021.

[51] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, 2011.

[52] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[53] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[54] Damjan Krstajic. Critical assessment of conformal prediction methods applied in binary classification settings. *Journal of Chemical Information and Modeling*, 61(10):4823–4826, 2021.

[55] Robert E Kass, Bradley P Carlin, Andrew Gelman, and Radford M Neal. Markov chain monte carlo in practice: a roundtable discussion. *The American Statistician*, 52(2):93–100, 1998.

[56] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. 2008.

[57] Wolf Vanpaemel. Prior sensitivity in theory testing: An apologia for the bayes factor. *Journal of Mathematical Psychology*, 54(6):491–498, 2010.