UNIVERSITY OF CALIFORNIA, SAN DIEGO

A Fully Bayesian Approach to Logistic Regression

A Thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Electrical Engineering
(Intelligent Systems, Robotics, and Control)

by

Joanne L. Shin

Committee in charge:

Professor Todd P. Coleman, Chair
Professor Sanjoy Dasgupta
Professor Gert Lanckriet

2015

The Thesis of Joanne L. Shin is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

Chair

University of California, San Diego

2015

DEDICATION

To my mother Wai Ling, father Frank and sister Jennie.

EPIGRAPH

The effective theory is only moderately successful in predicting behavior because, as we all know, decisions are often not rational or are based on a defective analysis of the consequences of the choice. That is why the world is such a mess.

—Stephen Hawking

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my advisor Prof. Todd Coleman for the continuous support of my studies and research, for his motivational idioms, enthusiasm, patience, and extensive knowledge in a number of cross-disciplinary fields. I could have not imagined having a better advisor for my graduate studies.

Besides my advisor, I would like to thank my undergraduate mentor and good friend Saura Naderi for the support throughout these past six years, for her encouragement, patience, advice and continuous life guidance in this journey into adulthood.

I thank my fellow lab and officemates for the colorful discussions we have shared, fancy espressos, and for all the fun we have spent together causing a ruckus on the 4th floor of Calit2.

Last, but not least I thank my parents and sister for the non-stop support and encouragement in completing my graduate studies; along with Alice Lin, Taylor LoNigro, Christopher Chandler, Kevin Yeap and many other close friends who remind me there exists a life beyond the confines of school.

ABSTRACT OF THE THESIS

A Fully Bayesian Approach to Logistic Regression

by

Joanne L. Shin

Master of Science in Electrical Engineering
(Intelligent Systems, Robotics, and Control)

University of California, San Diego, 2015

Professor Todd P. Coleman, Chair

Binary logistic regression is often used in clinical applications to predict the occurrence of medical conditions that arise within a patient population. Point estimations are often made to approximate the unknown regression weights. In doing so, information about the underlying posterior distribution of the weights is lost. We propose a method that views the logistic regression model from a Bayesian perspective and takes into consideration the full posterior of the unknown regression coefficients when computing the probability of belonging to the positive class. This method will be referred to as the *Fully Bayesian* method. The Fully Bayesian method allows us to quantify the uncertainty in our

probability calculations. The work in this paper builds on Kim and Ma's previous work in which they demonstrated efficiently solvable fully Bayesian estimation techniques. By solving a (convex) Kullback-Leibler divergence problem they were able to obtain a mapping from any log-concave prior to its corresponding posterior distribution thus enabling one to draw independent samples from the posterior with ease. Having the full posterior is useful in revealing how credible a prediction is and can be utilized to define an abstain strategy. The data set was created from a subsample of de-identified patient data from Kaiser Permanente and consists of a highly imbalanced number of patients that have and have not been diagnosed with asthma. The results show that the overall performance of a Fully Bayesian scheme produces a higher measure of accuracy than the point estimate method.

Chapter 1

Introduction

1.1 Background

Regression models are predictive models widely used across many disciplines and applications. These models aim to fit some functional relationship between a number of categorical dependent variables (regressors) to an independent variable. The relationship between the dependent and independent variable determines the type of regression model. In situations where the independent variable only takes on two states (binary classes), logistic regression is often employed.

In clinical applications, binary logistic regression is a popular method for predicting medical outcomes. The regressors are often seen to be a mixture of binary (yes or no) and real valued measurements (heart rate, body mass index, age, etc.) [1]. The possible outcomes are usually 1 or 0 indicating that a patient *does* or *does not* have a particular disease or condition. To fit the model according to a particular data set, we need to estimate the parameters involved - namely the weight of each regressor. The fitting phase will be referred to as the *learning* phase and the data involved during the learning phase will be

referred to as the *training* data. To test our model we hold out a portion of data that is unseen by our model and this data is referred to as the *testing* data. As we will see in the following section, a point estimation that minimizes the loss associated with logistic regression is made to approximate the weight on each regressor. By performing a point estimation to learn the regression weights information about the uncertainty of the estimation is lost. However, in clinical domains one may want to ask - how credible is the outcome of the model?

1.2 Notation

The following notation will be consistently used throughout this document.

- Capitalization will be used to indicate random variables
- D represents the number of regressors/features (regressors and features will be used interchangeably)
- $\vec{x} \in \{0,1\}^{D+1}$ represents the regressors in vector form, $\vec{x} = [1\ x_1\ x_2\ ...\ x_D]$
- $y \in \{0,1\}$ represents the class label
- Class labels 0 and 1 are referred to as the "negative" and "positive" class
- $\vec{w} \in \mathcal{R}^{D+1}$ represents the weights or coefficients in vector form, $\vec{w} = [w_0\ w_1\ w_2\ ...\ w_D]$
- Superscripts are used to indicate different samples (i.e. $\vec{x}^{(i)} = [1\ x_1^{(i)}\ x_2^{(i)}\ ...\ x_D^{(i)}]$ indicates the ith sample)

Other notation will be defined as needed.

Chapter 2

Fully Bayesian Decision-Making

2.1 Traditional Logistic Regression

Logistic regression assumes the following relationship between the conditional probability of belonging to the positive class (p) and the regressors $\vec{x}$ (2.1).

$$\log \frac{p}{1-p} = w_0 + w_1 x_1 + ... + w_D x_D \tag{2.1}$$

where

$$p = P(Y = 1 | X = \vec{x}\,; \vec{w}) \tag{2.2}$$

Note that the semi-colon is used in the notation above to indicate that $\vec{w}$ is a deterministic unknown variable. In the spirit of using logistic regression to model medical outcomes let each sample represent different patients and let each outcome of belonging to the positive class be analogous to a patient being diagnosed with a medical condition. All patients

are assumed to be independent of one another. Every patient k has their own features and probability of being diagnosed: $(\vec{x}^{(k)}, y^{(k)})$, but all patients are assumed to share the same weights ($\vec{w}$) on each regressor. Using a labeled set of training data the coefficients can be estimated.

Before getting to how $\vec{w}$ is estimated, let us first set up the overall problem in more detail. The relationship from (2.1) can be rewritten as (2.3) and (2.4).

$$P(Y = 1|X = \vec{x}^{(i)}\,;\vec{w}) = \frac{1}{1+e^{-\vec{w}^T\vec{x}^{(i)}}} \tag{2.3}$$

$$P(Y = 0|X = \vec{x}^{(i)}\,;\vec{w}) = \frac{1}{1+e^{\vec{w}^T\vec{x}^{(i)}}} \tag{2.4}$$

Since $X \in \{0,1\}^{D+1}$, we can represent (2.3) and (2.4) compactly as (2.5).

$$P(Y = y^{(i)}|X = \vec{x}^{(i)}, W = \vec{w}) = \left[\frac{1}{1+e^{-\vec{w}^T\vec{x}^{(i)}}}\right]^{y^{(i)}} \left[\frac{1}{1+e^{\vec{w}^T\vec{x}^{(i)}}}\right]^{(1-y^{(i)})} \tag{2.5}$$

The earlier assumption of each patient being independent allows us to represent the probability of all outcomes of every patient as the product of each individual outcome (2.7).

$$L(\vec{w}) = P(Y = y^{(1)},...,y^{(N)}|X = \vec{x}^{(1)},...,\vec{x}^{(N)}\,;\vec{w}) \tag{2.6}$$

$$= \prod_{i=1}^{N} P(Y = y^{(i)}|X = \vec{x}^{(i)}\,;\vec{w}) \tag{2.7}$$

This is also known as the likelihood function ($L(\vec{w})$) that is parameterized by $\vec{w}$.

2.1.1 Point Estimation

Given a set of labeled training data, the goal is to find the values of $\vec{w}$ that maximizes (2.7). Note that (2.7) is not concave with respect to $\vec{w}$ and is therefore difficult to solve. However the log-likelihood whose structure is the log-sum of exponents is well-known to be concave [2]. Moreover, since $\log(u)$ is a monotonically increasing function the maxima/minima is preserved (2.8) - (2.9). Maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood as shown in (2.10) - (2.11).

$$\vec{w}^* = \underset{\vec{w}}{\operatorname{argmax}}\, L(\vec{w}) \tag{2.8}$$

$$= \underset{\vec{w}}{\operatorname{argmax}}\, \log L(\vec{w}) \tag{2.9}$$

$$= \underset{\vec{w}}{\operatorname{argmax}}\, -\sum_{i=1}^{N} y^{(i)} \log(1+e^{-\vec{w}^T\vec{x}^{(i)}}) - (1-y^{(i)}) \log(1+e^{\vec{w}^T\vec{x}^{(i)}}) \tag{2.10}$$

$$= \underset{\vec{w}}{\operatorname{argmin}} \sum_{i=1}^{N} y^{(i)} \log(1+e^{-\vec{w}^T\vec{x}^{(i)}}) + (1-y^{(i)}) \log(1+e^{\vec{w}^T\vec{x}^{(i)}}) \tag{2.11}$$

The negative log-likelihood is often referred to as the log-loss (2.12) - (2.13).

$$l(\vec{w}) = -L(\vec{w}) \tag{2.12}$$

$$= \sum_{i=1}^{N} y^{(i)} \log(1+e^{-w^T x^{(i)}}) + (1-y^{(i)}) \log(1+e^{\vec{w}^T\vec{x}^{(i)}}) \tag{2.13}$$

Finding the optimal value $\vec{w}^*$ in this manner is referred to as the Maximum Likelihood Estimate (MLE). A penalty or regularization term is often added on to prevent over-fitting by discouraging the values of $\vec{w}$ to grow impractically large (2.14). Imposing this structure onto $\vec{w}$ is equivalent in a Bayesian sense of imposing some prior belief on the regression

weights and interpreting the weights themselves as being a random variable. This equivalent problem in a Bayesian sense is referred to as a Maximum a Posteriori (MAP) point estimation problem and will be further explored in the next section.

$$\vec{w}^* = \underset{\vec{w}}{\operatorname{argmin}}\, l(\vec{w}) + \lambda R(\vec{w}) \tag{2.14}$$

Let us shift to a Bayesian perspective and interpret the regression weights as being a random variable. The choice of $R(\vec{w})$ determines the structure of the prior on W. For instance, choosing an L2 regularizer corresponds to imposing a Gaussian prior on W whereas an L1 regularizer corresponds to imposing a Laplacian prior on W. Once $\vec{w}^*$ is estimated, new samples ($\vec{x}^{(new)}$) can be classified as belonging to the positive class according to (2.15).

$$\begin{aligned} & P(Y=1|X=\vec{x}^{(new)}, W=\vec{w}^*) \geq P(Y=0|X=\vec{x}^{(new)}, W=\vec{w}^*) \\ & P(Y=1|X=\vec{x}^{(new)}, W=\vec{w}^*) \geq 1 - P(Y=1|X=\vec{x}^{(new)}, W=\vec{w}^*) \\ \text{Declare 1 if: } & P(Y=1|X=\vec{x}^{(new)}, W=\vec{w}^*) \geq \frac{1}{2} \end{aligned} \tag{2.15}$$

Note that when the costs associated with the two types of error are not equal, 1/2 can be replaced by some threshold value, τ, where τ is a ratio of the costs.

2.2 Fully Bayesian Logistic Regression

As seen previously, the common Maximum Likelihood point estimation with a regularization on $\vec{w}$ has an equivalence to a Bayesian MAP point estimate. This section is primarily focused on the Bayesian interpretation of the point estimation problem. Although point estimation is commonly used to estimate unknown parameters, there are fundamental

limitations on how well a parameter can be estimated. For example, suppose we want to estimate the regression weights in a logistic regression framework with D = 1 regressor. Let $P(W = \vec{w}|X = \vec{x}^{(1)},...,\vec{x}^{(N)}, Y = y^{(1)},...,y^{(N)}) \sim N(0,\sigma^2)$ be the posterior on W. For short, let the posterior be $\pi(\vec{w})$. Figure 2.1 shows two possible Gaussian posteriors for W. Although the green curve has a 95% credibility interval that is much wider than the blue curve, the MAP point estimates are equal among the two curves ($\vec{w}^* = 10$ shown in red). As the name suggests - credibility intervals are useful in revealing how "good" or credible the estimation is. If our point estimate came from the green posterior, relatively large changes in the value that W takes on result in small changes in the probability of W. When evaluating the posterior, it may seem as though any value of W within the 95% credibility interval of $\vec{w}^*$ is suitable. However, large changes in the values that W takes on tend to result in large enough changes in the probability of belonging to the positive class to make the decision unstable (flip between 0 and 1).

We propose a Fully Bayesian logistic regression method that robustly learns the posterior of W and takes the distribution along with its credibility intervals into consideration when performing a classification task. Let us first assume that we are able to obtain the full posterior distribution and in detail show how it is useful. Recall that the rule for classifying a new sample $\vec{x}^{(new)}$ in a MAP point estimate framework is given by (2.15). Note that the training data, $I = (\vec{x}^{(1)}, y^{(1)}),...,(\vec{x}^{(N)}, y^{(N)})$, no longer provides any additional information once W is learned. Thus, conditional independence leads us to (2.16).

$$P(Y = 1|I, X = \vec{x}^{(new)}, W = \vec{w}) = P(Y = 1|X = \vec{x}^{(new)}, W = \vec{w}) \tag{2.16}$$

Let the posterior distribution on the weights be defined by (2.17).

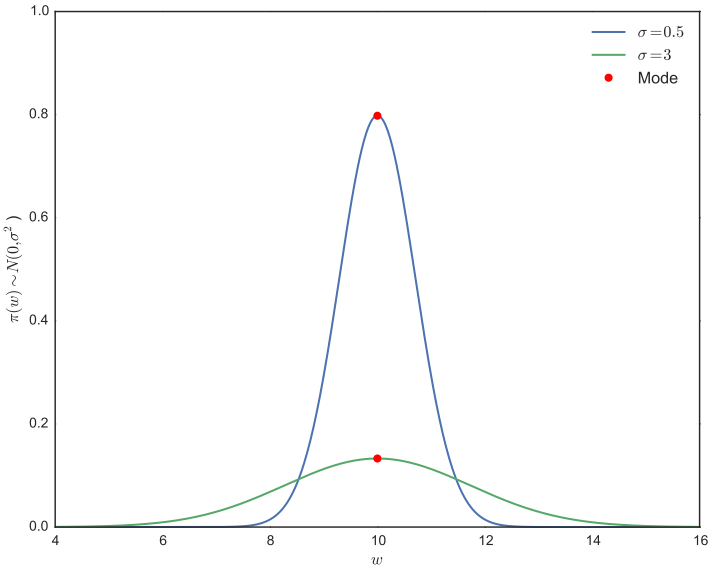

Figure 2.1: *Two possible posterior distributions on W. Note although the two distributions are different they have the same mode (shown in red).*

$$\pi(\vec{w}) = P(W = \vec{w}|I) \tag{2.17}$$

The Law of Total Probability can be used to determine which weights are the most fitting (2.19) - (2.20) by marginalizing out W. The marginal probability (2.18) is equivalent to averaging the sigmoid function over the distribution of all values of W. According to the Law of Large Numbers, the integral in (2.20) can be estimated by drawing many independently

identically distributed (i.i.d) samples from $\pi(\vec{w})$ (2.21).

$$P(Y=1|X=\vec{x}^{(new)},I) \tag{2.18}$$

$$= \int_{\vec{w}} P(Y=1|X=\vec{x}^{(new)},W=\vec{w},I)P(W=\vec{w}|I)d\vec{w} \tag{2.19}$$

$$= \int_{\vec{w}} \frac{1}{1+e^{-\vec{w}^T\vec{x}^{(new)}}}\pi(\vec{w})d\vec{w} \tag{2.20}$$

$$\underset{(Z_k)_{k=1}^M\sim\pi}{\approx} \frac{1}{M}\sum_{k=1}^{M}\frac{1}{1+e^{-Z_k^T\vec{x}^{(new)}}} \tag{2.21}$$

This method provides a different way of computing the probability of belonging to the positive class by taking into consideration the distribution of the posterior on W. Note that equation (2.21) is an average of $P(Y=1|X=\vec{x}^{(new)},W=Z_k)$, or the probability of belonging to the positive class given that W takes on some value Z_k. Thus, the Fully Bayesian method can be interpreted as an average of a collection of point estimations. This allows us empirically calculate the uncertainty of the probability of belonging to the positive class.

Let us revisit the two cases depicted by the blue and green curves in figure 2.1 using the Fully Bayesian method for decision-making. When $\pi(\vec{w})$ is narrow (i.e. the blue curve in figure 2.1) one would (most likely) see the distribution of $P(Y=1|X=\vec{x}^{(new)},W=Z_k)$, for $k=1,...,M$ looks similar to figure 2.2. Figure 2.2 shows that the empirical 95% credibility interval is below the $\tau=0.5$ threshold. This suggests that our classifier's decision for $X=\vec{x}^{(new)}$ is fairly confident.

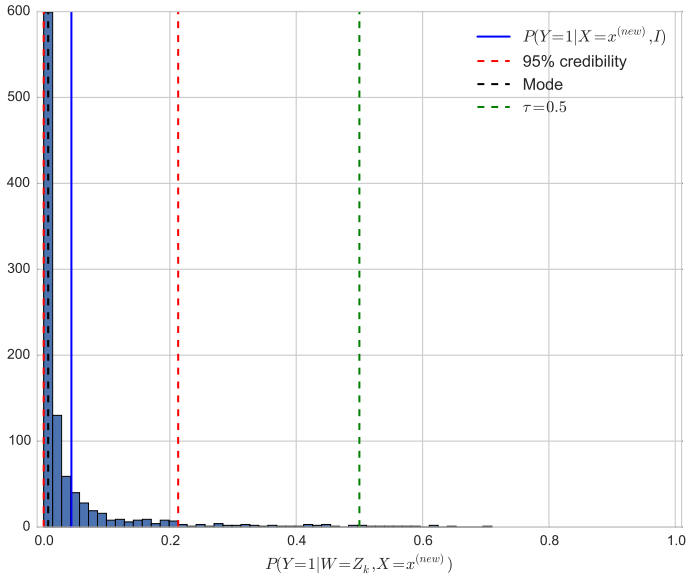

Figure 2.2: *For each sample of W at a fixed new sample $X=\vec{x}^{(new)}$, $P(Y=1|W=Z_k,X=\vec{x}^{(new)})$ changes. Thus, drawing M samples of W results in an estimated distribution of $P(Y=1|W=Z_k,X=\vec{x}^{(new)})$ for $k=1,...,M$*

Now, imagine $\pi(\vec{w})$ is wide (i.e. the green curve in figure 2.1), then our histogram would (most likely) resemble figure 2.3. Although the mode and average are well below the 0.5 threshold in figure 2.3, the empirical 95% credibility interval crosses 0.5. These cases are indicative of our classifier being unsure or less confident about its decision. In these cases, one might seek an alternative classification rule such as abstaining from making a decision.

In many clinical settings where there is a high cost associated with making an error, it is not uncommon to abstain from making a decision when the confidence of the prediction is low [3]. In fact, Ferri and Hernandez-Orallo [4] measure the confidence of a decision by equating it to the probability of belonging ot the positive class. They define a range of low confidence by selecting a finite region near τ in which their classifier abstains. A problem

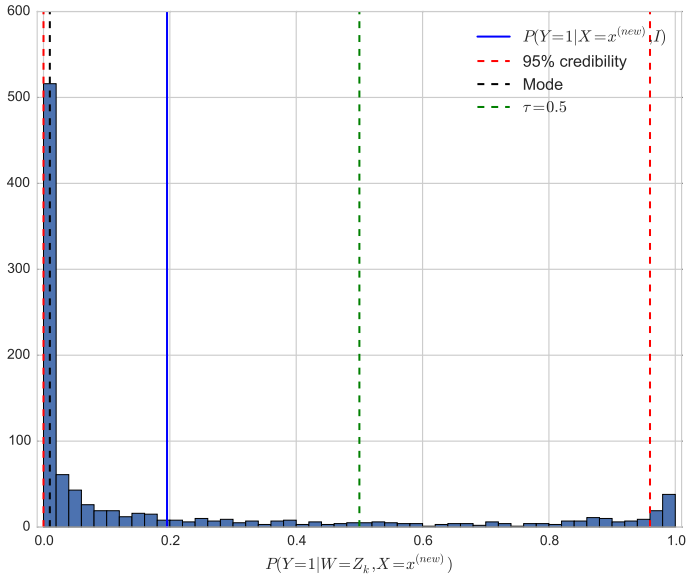

Figure 2.3: *The histogram above shows the empirical distribution of* $P(Y=1|W=Z_k, X=\vec{x}^{(new)})$ *for a fixed sample* $X=\vec{x}^{(new)}$ *and* M *samples from the posterior. Note that the 95% credibility interval crosses the decision threshold,* τ.

with this method is that it ignores the credibility of the probability measure itself. As a result of the Fully Bayesian method, credibility intervals are readily available and can be used in designing a more careful abstain strategy. Such a strategy would involve abstaining when the 95% credibility interval is on the opposing side of a threshold value relative to its probability of belonging to the positive class. An example of this will be shown in the Experiments chapter.

2.2.1 Learning the Posterior

One of the main motivations behind developing point estimation methods is avoiding the hassle of having to compute non-trivial integrals in learning the full posterior distribution. However, one can approximate the full posterior distribution as well as any statistic belonging

to it so long as a large number of samples from the posterior are attainable. A well-known sampling method for attaining such samples is Gibbs Sampling [5]. A Gibbs Sampler is a Markov Chain Monte Carlo (MCMC) algorithm. One of the biggest caveats of MCMC algorithms is they tend to have unknown rates of convergence; meaning, there is no guarantee on how long it will take to train such a classifier. Another significant problem arises from changing the underlying probability distributions of interest. Since MCMC algorithms are structured uniquely for specific probability distributions, altering these distributions often entirely changes the structure of the algorithm.

Through recent work by Kim and Ma, [6], an efficient way to attain samples from a posterior distribution has been developed that does not suffer from the same issues seen in MCMC algorithms. As we will see, Kim, Ma and Mesa's [7] results provide a robust way sample from the posterior given any log-concave prior. To summarize their results let us define a few terms. Let $\mathbb{P}$ and $\mathbb{Q}$ correspond to the prior and posterior distribution, respectively, of the some random variable and let S be a map that pushes or transforms $\mathbb{P}$ to $\mathbb{Q}$. Kim and Ma state that if our likelihood and prior distributions are log-concave, then there exists some diffeomorphism map S that will push $\mathbb{P}$ to $\mathbb{Q}$. Moreover, the densities corresponding to $\mathbb{P}$ and $\mathbb{Q}$ - p, q, respectively are related through (2.22).

$$p(u) = q(S(u))|\det(J_S(u))| \tag{2.22}$$

This result allows us to draw samples from a (known) prior distribution and transform them into samples drawn from its corresponding posterior distribution.

In further summary, the necessary details will be included to give an idea of how the the map is attained. Let S^* be the optimal or desired map that pushes $\mathbb{P}$ to $\mathbb{Q}$ and let all other maps, S, push some $\tilde{\mathbb{P}}$ that is not necessarily equal to $\mathbb{P}$ to $\mathbb{Q}$ as shown in figure 2.4 [6].

Notice if $\tilde{\mathbb{P}} \to \tilde{\mathbb{P}}$ then $S \to S^*$, thus minimizing the "distance" between $\tilde{\mathbb{P}}$ and $\mathcal{P}$ will lead to finding the desired map. The Jacobian equation (2.22) restricts the search to only consider positive-definite maps and guarantees the optimal map S^* is unique. The optimization problem that Kim and Ma formulate state exactly this (2.23), where the distance metric in a probability sense is the well-known Kullback-Leibler (KL) divergence.

$$S^* = \underset{S \in \mathcal{D}^+}{\operatorname{argmin}} D(P||\tilde{P}) \tag{2.23}$$

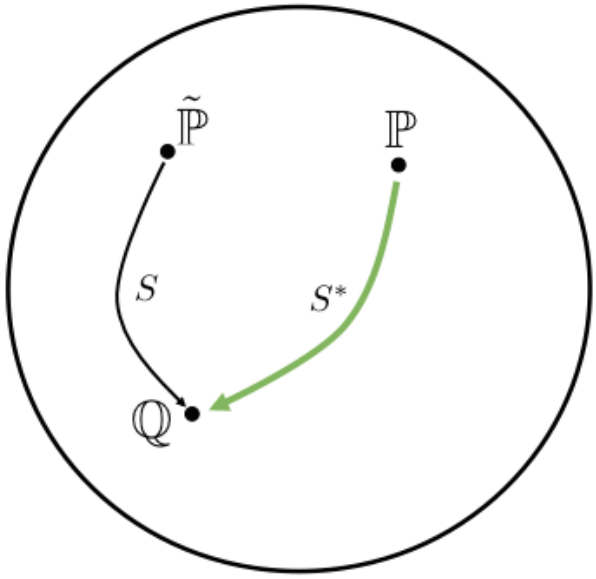

Figure 2.4: *Shows the mapping from P to Q by S^*, $\tilde{P}$ S and Q to $\tilde{P}$ by S^{-1} (where S is not necessarily S^*)*

In terms of the logistic regression framework, the posterior on W is the one of interest. Let $f_W(\vec{w})$ denote the probability density of W. Equation (2.23) becomes (2.24) - (2.25).

$$S^* = \underset{S_y \in \mathcal{D}^+}{\operatorname{argmin}} \int_{\vec{w} \in W} f_W(\vec{w}) \log \frac{f_W(\vec{w})}{\tilde{f}_W(\vec{w})} d\vec{w} \tag{2.24}$$

$$= \underset{S_y \in \mathcal{D}^+}{\operatorname{argmin}} \log \beta_y - \int_{\vec{w} \in W} f_W(\vec{w}) T(S_y, \vec{w}) \tag{2.25}$$

$$\approx \underset{S_y \in \mathcal{D}^+}{\operatorname{argmax}} \frac{1}{M} \sum_{i=1}^{M} \tilde{T}(S_y, W_i)$$

where $W_1, W_2, ..., W_M$ are drawn i.i.d from P_W and

$$\beta_y = \int_{v \in W} f_{Y|W}(y|\vec{w}) f_W(v) dv \tag{2.26}$$

$$T(S_y, \vec{w}) = \log f_{Y|W}(y|S(\vec{w})) + \log f_W(S(\vec{w})) \tag{2.27}$$

$$+ \log \det(J_S(\vec{w})) - \log f_W(\vec{w})$$

Kim and Ma's paper proceeds to approximate S as a linear basis expansion (2.28).

$$S(\vec{w}) = \sum_{\forall i} g_j \phi^{(j)}(\vec{w}) \tag{2.28}$$

Where $\phi^{(i)} \in \mathcal{R}$, $g_j \in \mathcal{R}^D$ (where D is the dimension of X) and $A = [\phi^{(1)}(\vec{w}), ..., \phi^{(K)}(\vec{w})]$ are chosen to be orthogonal with respect to the prior distribution. By approximating S with a linear basis expansion, we no longer need to search over all possible diffeomorphic functions. Instead, the search is limited to the space of possible basis coefficients corresponding to diffeomorphisms. The key terms with their corresponding dimensions are defined by (2.29)

- (2.33).

$$F = [g_1, ..., g_K] \qquad (D \times K) \tag{2.29}$$

$$A(\vec{w}) = [\phi^{(1)}(\vec{w}), ..., \phi^{(K)}(\vec{w})]^T \qquad (K \times 1) \tag{2.30}$$

$$S(\vec{w}) = FA(\vec{w}) \qquad (D \times 1) \tag{2.31}$$

$$J_A(\vec{w}) = \left[\frac{\partial \phi^{(i)}}{\partial \vec{w}_j}(\vec{w})\right]_{i,j} \qquad (K \times d) \tag{2.32}$$

$$J_S(\vec{w}) = FJ_A(\vec{w}) \qquad (D \times D) \tag{2.33}$$

Kim and Ma's final optimization problem (denoted as (P4) in their paper) is represented by (2.34).

$$F^* = \underset{FJ_A(W_1) \succ 0, ..., FJ_A(W_M) \succ 0}{\operatorname{argmax}} \frac{1}{M} \sum_{i=1}^{M} \tilde{T}(F, W_i) \tag{2.34}$$

where again, $W_1, W_2, ..., W_M$ are drawn i.i.d. from P_W and $\tilde{T}(F, \vec{w})$ is defined by

$$\tilde{T}(F, \vec{w}) \triangleq \log f_{Y|W}(y|FA(\vec{w})) + \log f_W(FA(\vec{w})) \tag{2.35}$$
$$+ \log\det(FJ_A(\vec{w})) - \log f_W(\vec{w})$$

Since (2.34) or (P4) is convex it theoretically can be efficiently solved. However, the number of constraints scales linearly with the number of samples we draw from the prior (M) and in order to approximate the map well, a large number of samples drawn from the prior are needed. As a result, the number of constraints become a computational bottleneck. More recent developments by Mesa and Kim [7] resolve this issue by taking (P4) and turning it into a distributed problem via the Alternating Direction Method of Multipliers (ADMM)

(2.36).

$$\begin{aligned}
\min_{F,Z,p,B} \quad & \frac{1}{M}\sum_{i=1}^{M} g(p_i) - \log\det(Z_i) + \frac{1}{2}\rho||F_i - B||_2^2 \\
& + \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\rho||BA_i - p_i||_2^2 + \frac{1}{2}\rho||BJ_i - Z_i||_2^2 \\
\text{s.t.} \quad & BA_i = p_i : \quad \gamma_i \quad (d \times 1) \\
& BJ_i = Z_i : \quad \lambda_i \quad (d \times d) \\
& F_i - B = 0 : \quad \alpha_i \quad (d \times K)
\end{aligned} \tag{2.36}$$

where γ_i, λ_i and α_i are Lagrange multipliers and $g(p_i) = -\log q(p_i)$. This allows us to form the penalized Lagrangian (2.37).

$$\begin{aligned}
\mathcal{L}_\rho(F,Z,p,B;\gamma,\lambda,\alpha) = & \frac{1}{M}\sum_{i=1}^{M} g(p_i) - \log\det Z_i \\
& + \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\rho||F_i - B||_2^1 + \frac{1}{2}\rho||BA_i - p_i||_2^2 \\
& + \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\rho||BJ_i - Z_i||_2^2 + \gamma_i^T(p_i - BA_i) \\
& + \frac{1}{M}\mathbf{tr}(\gamma_i^T(Z_i - BJ_i)) + \mathbf{tr}(\alpha_i^T(F_i - B))
\end{aligned} \tag{2.37}$$

ADMM allows us to separate (2.37) into sequential optimization problems for $B^{k+1}, F_i^{k+1}, Z_i^{k+1}, p_i^{k+1}, \gamma_i^{k+1}, \lambda_i^{k+1},$ and α_i^{k+1} for $i = 1, ..., M$ as shown in (2.38)-(2.44).

$$B^{k+1} = \underset{B}{\operatorname{argmin}}\ \mathcal{L}_\rho(F_i^k, Z_i^k, p_i^k B; \gamma^k, \lambda^k, \alpha^k) \tag{2.38}$$

$$F_i^{k+1} = \underset{F_i}{\operatorname{argmin}}\ \mathcal{L}_\rho(F_i, Z_i^k, p_i^k B^{k+1}; \gamma^k, \lambda^k, \alpha^k) \tag{2.39}$$

$$Z_i^{k+1} = \underset{Z_i}{\operatorname{argmin}}\ \mathcal{L}_\rho(F_i^{k+1}, Z_i, p_i^k B^{k+1}; \gamma^k, \lambda^k, \alpha^k) \tag{2.40}$$

$$p_i^{k+1} = \underset{p_i}{\operatorname{argmin}}\ \mathcal{L}_\rho(F_i^{k+1}, Z_i^{k+1}, p_i B^{k+1}; \gamma^k, \lambda^k, \alpha^k) \tag{2.41}$$

$$\gamma_i^{k+1} = \gamma_i^k + \rho(p_i^{k+1} - B^{k+1} A_i) \tag{2.42}$$

$$\lambda_i^{k+1} = \lambda_i^k + \rho(Z_i^{k+1} - B^{k+1} J_i) \tag{2.43}$$

$$\alpha_i^{k+1} = \alpha_i^k + \rho(F_i^{k+1} - B^{k+1}) \tag{2.44}$$

All the update terms above except p_i^{k+1} are guaranteed to have a closed form solution and can be found in the Appendix section of [7]. The update term, p_i^{k+1}, is the only term that is directly dependent on the likelihood and prior. The update corresponding to logistic regression with a L1 penalty term is shown in (2.45)-(2.46)

$$p_i^{k+1} = \underset{p_i}{\operatorname{argmin}}\ g(p_i) + \frac{1}{2}\rho ||B^{k+1} A_i - p_i||_2^2 + \gamma_i^{kT}(p_i - B^{k+1} A_i) \tag{2.45}$$

$$\begin{aligned} g(p_i) \triangleq & -\log P(Y = y^{(1)}, ..., y^{(N)} | X = \vec{x}^{(1)}, .., \vec{x}^N, W = p_i) \\ & - \log P(W = p_i) \\ = & \sum_{i=1}^{N} y^{(i)} \log(1 + e^{-p_i^T x^{(i)}}) + (1 - y^{(i)}) \log(1 + e^{p_i^T x^{(i)}}) + \beta ||p_i||_1 \end{aligned} \tag{2.46}$$

where β corresponds to a regularization coefficient. An L1 penalty was chosen because the data set used in the experiments was binary and sparse and it is well known that

L1 regularizations are preferred in such situations. The ADMM dual form of (P4) is computationally easier to handle. The number of constraints no longer scales with the number of samples from the prior. Instead we have M optimization problems for each update term. This is advantageous because the M optimization problems can now be solved in parallel by commonly used Map-Reduce routines.

Chapter 3

Experiments

The objective was to provide comparisons (in performance and capabilities) between a Fully Bayesian and traditional logistic regression classifier with and without an abstain option.

3.1 Data

The data set was created using de-identified real patient data from Kaiser Permanente. The medical condition we aimed to classify was asthma (ICD9: 493.00). The cases are defined as patients that have been diagnosed with asthma and the controls are defined as patients that have not been diagnosed with asthma. The data set contained 10,100 patients in which 100 were cases and 10,000 were controls. In the real patient population it was observed that there were approximately 10x more control patients than case patents. There were 10 features that characterized each patient shown in Table 3.1. Specifically, these features were all various medications that are often administered to treat asthma or asthma-like symptoms.

Table 3.1: Feature Descriptions

1	Antiasthmatics
2	Asthma/COPD Therapy - Beta 2-Adrenergic Agents, Inhaled, Short Acting
3	Albuterol Sulfate
4	Beta-Adrenergic Agents
5	Albuterol Sulfate HFA 90 mcg/actuation aerosol inhaler
6	Glucocorticoids, Orally Inhaled
7	Asthma Therapy, Glucocorticoids
8	Beclomethasone Dipropionate
9	Beclomethasone Dipropionate 80 mcg/actuation aerosol inhaler
10	Albuterol Sulfate 2.5 mg/3 mL (0.083 %) solution for nebulization

In a binary classification task, an ideal classifier has the ability to correctly identify the class label for a new observation. In practice this task is feasible when the distribution of the samples belonging to each class do have not any overlap or can be separated by some boundary. The simplest boundary is linear, but boundaries can take on hyperbolic, elliptic, and many other geometries. Our data set is not separable; all samples that belong to the positive class have twin samples that belong to the negative class. In other words, all the distinct combinations of medications prescribed to the case group were also prescribed to some patients in the control group. For our collection of case and control asthma patients, it is not uncommon for the medications listed in Table 3.1 to be prescribed as treatment for ailments that have similar symptoms to asthma. For instance, feature 2 is cross-referenced as COPD or chronic obstructive pulmonary disease and features 3 and 5 (albuterol) are used as a quick relief medication prescribed to treat breathing problems not limited to asthma.

3.2 Error Metrics

The error metrics used were balanced accuracy [8] precision, recall and confusion matrix:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{P}} + \frac{\text{TN}}{\text{N}}\right)$$

$$\text{N} = \text{TN} + \text{FP}$$

$$\text{P} = \text{FN} + \text{TP}$$

$$\text{Precision (PPV)} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

	Predicted 0	Predicted 1	
Truth 0	TN	FP	N
Truth 1	FN	TP	P

Note that balanced accuracy is equal to the traditional definition of accuracy when the data set is balanced. It is particularly useful in characterizing the accuracy of an imbalanced data set. For instance, if 99/100 of the patients belong to class 0, then the accuracy for a classifier that classifies all patients as belonging to class 0 is 99%. Clearly, this is a bit misleading. On the other hand, balanced accuracy for the same classifier equates to 0.5. Since balanced accuracy considers the ratio of the correctly classified patients for each class separately it is more useful over its traditional counterpart as a metric for error.

3.2.1 Experiment

All scripts created for the experiment were created in Python. The sklearn Logistic Regression package was used to create the point estimate model. Given a new patient, a classifier declares them as a case if the probability of belonging to the case group (p) is greater than the probability of belonging to the control group ($1-p$) or simply $p > 0.5$. However, we would also like to introduce an option for the classifier to abstain from making a decision when the uncertainty of a decision is high. Figure 3.1 depicts an ideal probability distribution of p for patients belonging to each class. These distributions are ideal because all the control patients have a low ($p < 0.5$) probability of belonging to the case group and vice-versa.

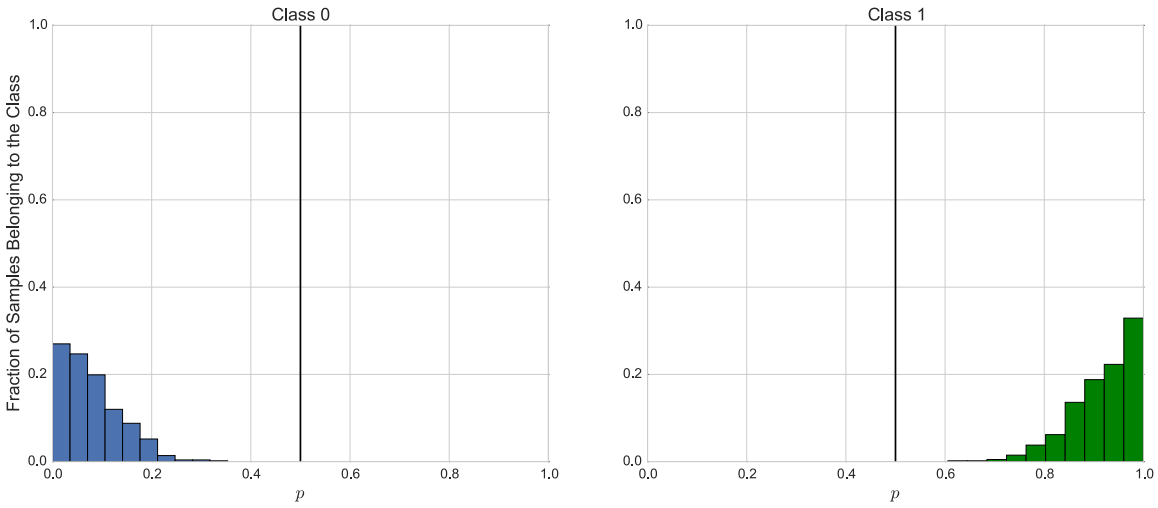

Figure 3.1: *An ideal histogram of the probability of belonging to the positive class (p) for all the patients within the training set.*

We trained logistic regression models using the Fully Bayesian and MAP point estimation approach and observed the distribution of p_{FB} and p_{MAP} given a particular patient as shown in figures 3.2 and 3.3, respectively. The patients from the training set were fed back into each model to observe the ability of each classifier to identify patients it had

already seen. Figure 3.2 shows the Fully Bayesian classifier was able to recover roughly half the case patients and approximately 95% of the control patients. In contrast, figure 3.3 shows the point estimate classifier was unable to distinguish between case and control patients; it believed all the patients belonged to the control group.

Let us define the credibility interval more formally. Let $p_{FB} + \delta^{+}$ to be the upper edge of the 95% credibility interval on p_{FB} and let $\Delta\tau$ be the difference between $p_{FB} + \delta^{+}$ and the threshold (τ) where $\tau = 0.5$ as shown in figure 3.4. When $p_{FB} + \delta^{+}$ is above τ, then $\Delta\tau > 0$ and when $p_{FB} + \delta$ is below τ, then $\Delta\tau < 0$. Although figure 3.2 shows approximately half the case patients being below $\tau = 0.5$, figure 3.5 shows that $\Delta\tau > 0$ for the majority of the case patients. This implies the credibility interval of p_{FB} straddles the threshold and suggests that the credibility interval can be used to identify possible errors that the classifier might make in identifying case patients.

Although the same argument can be made for the lower edge of the 95% credibility interval when identifying control patients figures 3.2 and 3.3 show both classifiers are fairly good at identifying control patients, thus we will focus our efforts in recovering case patients.

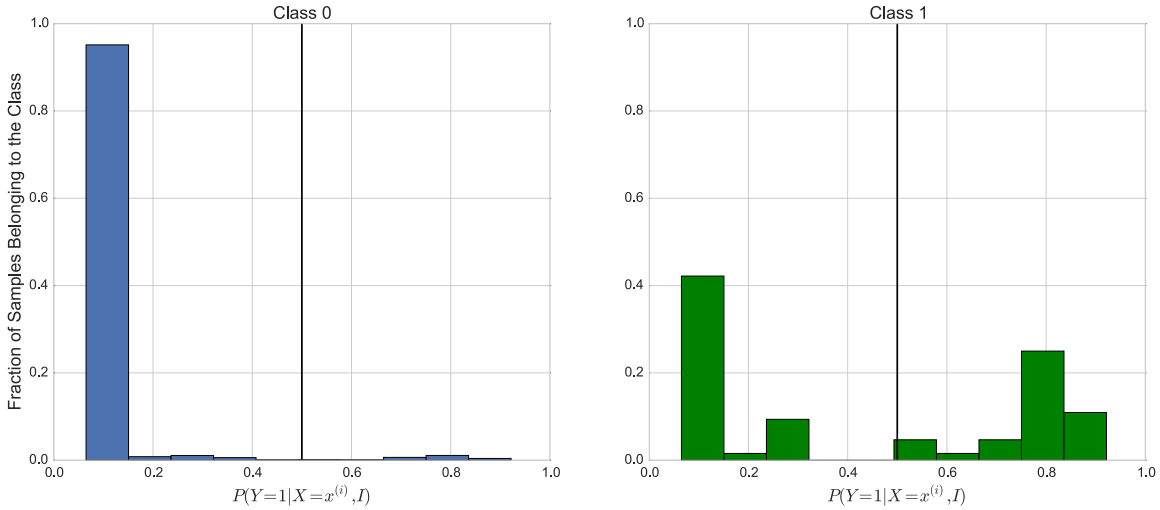

Figure 3.2: *The histogram of p_{FB} for all the patients within the training set.*

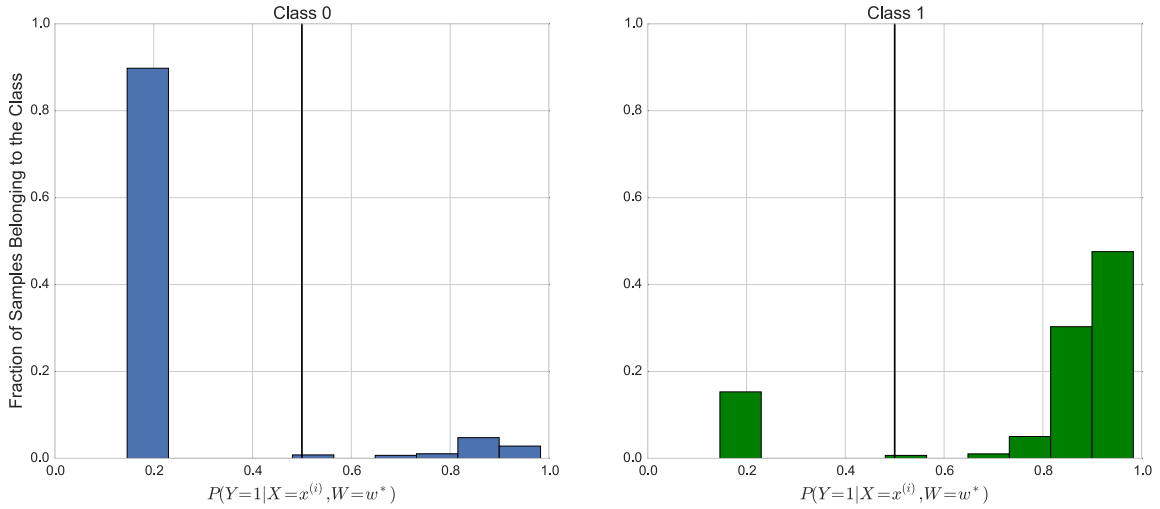

Figure 3.3: *The histogram of p_{MAP} for all the patients within the training set.*

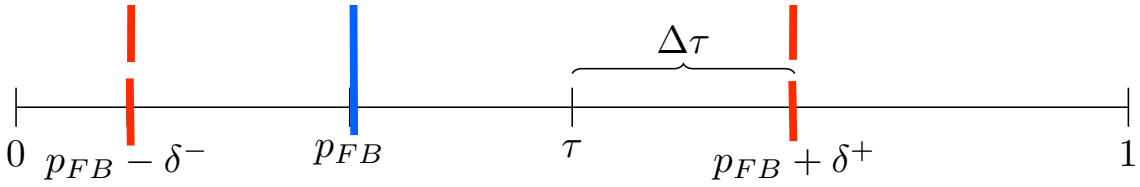

Figure 3.4: *$\Delta\tau$ is defined as the distance from τ to the upper edge of the p_{FB}'s 95% credibility interval ($p_{FB} + \delta^+$). When the cost of both errors are equal, $\tau = 0.5$.*

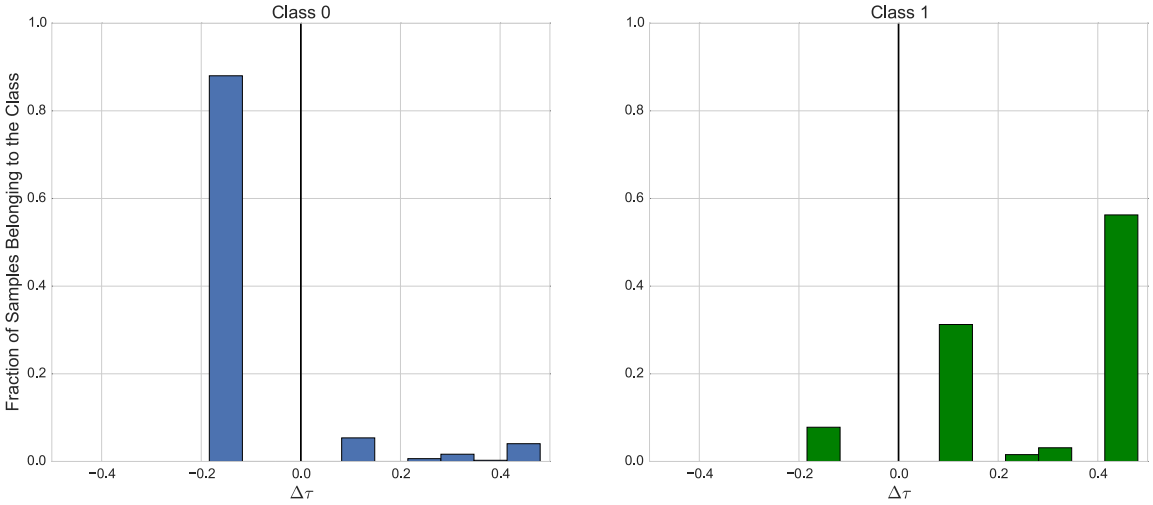

Figure 3.5: *The fraction of patients belonging to class 0 or 1 (from training) whose 95% credibility interval on p_{FB} exceeds the threshold (the black line).*

3.2.2 Results: Declaring a Class for Each Patient

The first experiment conducted was without an abstain option. The full data set of 10,100 patients was split up into a training (70%) and testing (30%) three times and three separate trials were observed. In all three trials, the point estimate classifier was not able to recall any of the case patients while the Fully Bayesian approach was able to recall approximately 50% of the case patients as shown in Tables 3.2 - 3.3.

Table 3.2: Summary of Error: Balanced Accuracy, Precision, Recall

Trial 1	Fully Bayesian	Point Estimate
Balanced Accuracy	0.7535	0.5
Precision	0.2345	N/A
Recall	0.5278	0.0

Trial 2	Fully Bayesian	Point Estimate
Balanced Accuracy	0.6850	0.5
Precision	0.1685	N/A
Recall	0.3947	0.0

Trial 3	Fully Bayesian	Point Estimate
Balanced Accuracy	0.7293	0.5
Precision	0.1585	N/A
Recall	0.4815	0.0

Table 3.3: Summary of Error: Confusion Matrix

Trial 1

Fully Bayesian	Predicted 0	Predicted 1
Truth 0	2932	62
Truth 1	17	19

Point Estimate	Predicted 0	Predicted 1
Truth 0	2994	0
Truth 1	36	0

Trial 2

Fully Bayesian	Predicted 0	Predicted 1
Truth 0	2918	74
Truth 1	23	15

Point Estimate	Predicted 0	Predicted 1
Truth 0	2992	0
Truth 1	38	0

Trial 3

Fully Bayesian	Predicted 0	Predicted 1
Truth 0	2934	69
Truth 1	14	13

Point Estimate	Predicted 0	Predicted 1
Truth 0	3003	0
Truth 1	27	0

The overall performance of the Fully Bayesian classifier according to precision-recall outperformed the point estimate classifier during the first two trials. In the third trial the precision-recall curve of the point estimate classifier performs better than the Fully Bayesian classifier at various values of τ. Note that throughout our analysis, τ was set to 0.5 to reflect weighing the cost of each type of error equally and the precision-recall curves are provided to get a sense of the performance at different values of τ.

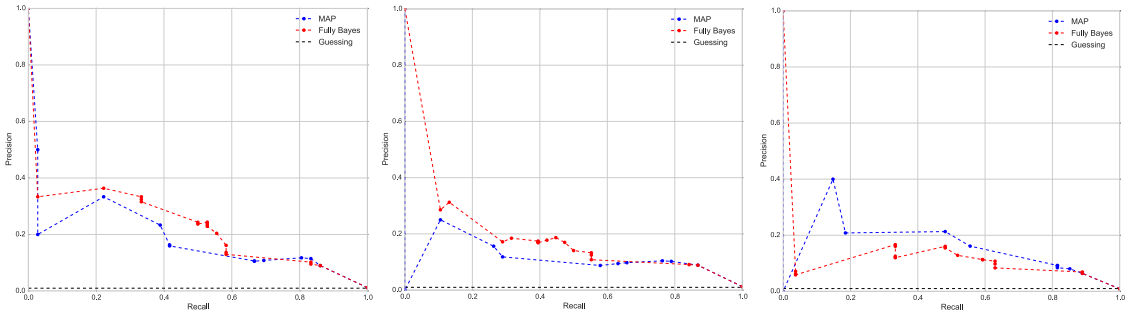

Figure 3.6: *Precision-Recall curve for MAP point-estimation (blue) and Fully Bayesian (blue) Logistic Regression for three separate trials.*

3.2.3 Results: Allowing an Abstain Option

In this scenario, the classifier was allowed to abstain from classifying up to K patients at no cost. Each abstention was then reviewed by a human expert that could correctly identify the case and control patients. With this scheme in mind, the goal was to come up with a strategy that abstains from making a decision when the classifier thinks it will make a

mistake. The credibility interval was used to quantify when the classifier felt uncertain about its decision. The Fully Bayesian classifier with an abstain option is given by Algorithm 1. Note that the algorithm was written to assume p_{FB} is a n length vector containing the probability of a patient belonging to the case group for n patients.

Algorithm 1 Fully Bayesian abstain rule

Require: p_{FB} is a n length vector of probabilities for n patients where $n \geq K$ and edgesUpper is an array with the upper edge of the 95% for each p_{FB}

```
function FULLYBAYESIANABSTAIN(p_FB, K,edgesUpper)
    x ← [patient 1, patient 2,..., patient n]                    ▷ Sample indexes
    Δτ, abstain ← [ ],[ ]
    for i = 1,...,n do
        if p_FB[i] < 0.5 then
            Δτ ← (edgesUpper[i] - 0.5)                           ▷ add to list
            xabstain ← x[i]                                      ▷ add to list
    ind ← Δτ.argsort()          ▷ sort Δτ by descending order and return indexes
    xabstain ← xabstain[ind[:K]]
    return xabstain
```

A similar scheme was defined for the MAP point estimate classifier where the K samples closest to τ were abstained shown in Algorithm 2. The performance of the two classifiers with an abstain option are summarized in figure 3.7.

Algorithm 2 MAP point estimate abstain rule

Require: p_{MAP} is a n length vector of probabilities for n patients where $n \geq K$

```
function POINTESTIMATE(p_FB, K)
    x ← [patient 1, patient 2,..., patient n]                    ▷ Sample indexes
    Δτ ← [ ]
    Δτ ← |0.5 − p_MAP|
    ind ← Δτ.argsort()                          ▷ Descending; ind: sort index
    xabstain ← x[:K]
    return xabstain
```

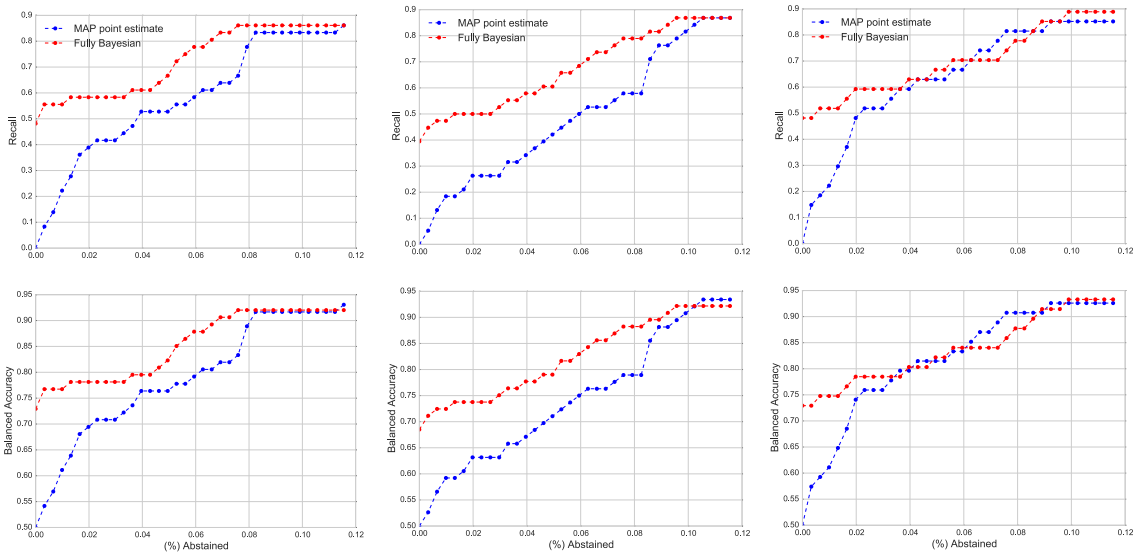

Figure 3.7: *Recall and balanced accuracy plotted versus the percentage of samples abstained for three separate trials.*

Figure 3.7 shows a comparison between the performance of the Fully Bayesian classifier and the MAP point estimate classifier with abstain options for three separate trials. Overall, leveraging the credibility interval (in the Fully Bayesian method) to identify abstentions performs better than its point estimate counterpart. When we are only allowed to abstain a few samples (less than 100), the steep slopes of the recall curves for the point estimate classifiers (top row, blue lines) suggests that they are better at identifying abstentions that result in case patients than the Fully Bayesian classifier. A smaller slope indicates a smaller improvement in performance as more patients are abstained. Between 150 to 250 abstentions, both the abstain strategies are able to recover approximately an equal number of case patients (approximately equal slope). The overall performance of the Fully Bayesian classifier with an abstain option outperforms the point estimate classifier with an abstain option because it is able to recognize some case patients on its own without a human expert. These trends are also reflected in the overall measurement of balanced accuracy.

Chapter 4

Conclusion

The Fully Bayesian logistic regression framework offers credibility intervals that can be used in a variety of different applications. This paper specifically illustrated a use case where it is difficult for a classifier to identify the underrepresented class with real patient data from a population of asthma patients. The credibility intervals were used to provide an additional measure of confidence and gave us a more robust way in identifying potential misclassifications. These misclassifications were abstained from being classified in a scheme in which abstentions were reviewed by human experts that correctly identified the class labels for all abstentions. Three trials were conducted and the results showed that leveraging the credibility interval through the Fully Bayesian method outperformed the traditional point estimation method overall under the assumption that the cost of each type of error was equal.

Bibliography

[1] M. P. LaValley. Statistical primer for cardiovascular research. *Circulation*.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[3] Yair Wiener and Ran El-yaniv. Agnostic selective classification. In J. Shawe-taylor, R.s. Zemel, P. Bartlett, F.c.n. Pereira, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1665–1673. 2011.

[4] J. Hernandez-Orallo C. Ferri. Cautious classifiers. *ROCA*, 4.

[5] Ilker Yildirim. Bayesian inference: Gibbs sampling.

[6] D. Mesa S. Kim, R. Ma and T. P. Coleman. Efficient bayesian inference methods via convex optimization and optimal transport. *IEEE International Symposium on Information Theory*.

[7] T. P. Coleman D. A. Mesa, S. Kim. A scalable framework to transform samples from one continuous distribution to another. *IEEE International Symposium on Information Theory (ISIT)*.

[8] K. E. Stephan J. M. Buhmann K. H. Brodersen, C. Ong. The balanced accuracy and its posterior distribution. *2010 International Conference on Pattern Recognition*, 2010.