# Logistic Regression Through the Veil of Imprecise Data

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Logistic regression is a popular supervised learning algorithm used to assess the probability of a variable having a binary label based on some predictive features. Standard methods can only deal with precisely known data; however, many datasets have uncertainties that traditional methods either reduce to a single point or completely disregard. This paper shows that it is possible to include these uncertainties by considering an imprecise logistic regression model using the set of possible models obtained from values within the intervals. This method can express the epistemic uncertainty neglected by traditional methods.

## 1 Introduction

Logistic regression is used to predict the probability of a binary outcome as a function of some predictive variable, where the time of the event is not essential. In medicine, for example, logistic regression can be used to predict the probability of an individual having a disease where the values of risk factors are known. While logistic regression is most commonly used for binary outcomes, it can be applied to any number of categorical outcomes (Menard, 2010, Chapter 1). However, many decisions and events are binary (yes/no, passed/failed, alive/dead, etc.), and for the sake of simplicity, we will restrict our discussion and examples to binary outcome logistic regression. Additionally, unlike discriminant function analysis, logistic regression does not require predictor variables to be normally distributed, linearly related or equal variance (Press & Wilson, 1978).

There are many practical applications for logistic regression across many different fields. For example, in the medical domain, risk factors in the form of continuous data–such as age–or categorical data–such as gender–may be used to fit a model to predict the probability of a patient surviving an operation (Bagley et al., 2001; Neary et al., 2003). In engineering systems, logistic regression can be used to determine whether a mineshaft is safe (Palei & Das, 2009); to predict the risk of lightning strikes (Lambert & Wheeler, 2005) or landslides (Ohlmacher & Davis, 2003). In the arts, it can explore how education impacts museum attendance or watching a performing arts performance (Kracman, 1996). It is also possible to predict the probability of sports results using logistic regression (Li et al., 2021). Due to its wide range of applications, logistic regression is considered a fundamental machine learning algorithm with many modern programming languages having packages for users to experiment with, such as *Scikit-learn* (Pedregosa et al., 2011) in Python, which has been used for the analysis within this paper.

Traditionally it has been assumed that all of the values of the features and labels used in logistic regression are *precisely* known. This assumption is valid when the sampling uncertainty or natural variability in the data is significant compared to the epistemic uncertainty or if values are missing at random (Ferson et al., 2007). However, in practice, there can be considerable imprecision in both the features and labels used in the regression analysis and the application of the regression model. Analysis using data from combined studies with inconsistent measurement methods can even result in datasets with varying degrees of uncertainty. Likewise, the outcome data can be uncertain if there is ambiguity in the classification scheme (good/bad). However, even relatively straightforward classifications (alive/dead) can yield uncertainty when a subject leaves a study, and the outcome is now unknown. Measurement uncertainty is sometimes best represented as intervals, sometimes called "censored data". In the case of continuous variables, the interval reflects the measurement uncertainty, while in the binary outcome, the interval is the vacuous $[0, 1]$ because the correct classification is unclear.

There are multiple methods of dealing with interval data with the features of a logistic regression model, but they often require approximations to be made to Simplify the process and allow the use of standard logistic regression techniques. One approach is to treat interval data as uniform distribution based on the "equidistribution hypothesis" that each possible value can be considered to be equally likely (Bertrand, 2000; Billard & Diday, 2000; Bock & Diday, 2000). This idea has its roots in the principle of insufficient reason, first described by both Bernoulli and Laplace, and more recently known as the principle of indifference (Keynes, 1921). Alternatively, the interval is commonly represented by the interval's midpoint, which represents the mean and median of a uniform distribution or a random value from within the interval (Osler et al., 2010). While these approaches are computationally expedient, they underrepresent the imprecision by presenting a single middle-of-the-road logistic regression.

Similar methods include performing a conjoint logistic regression using the interval endpoints or averaging separate regressions performed on the endpoints of the intervals (de Souza et al., 2008; 2011). While these various methods make different assumptions about the data within the interval ranges, ultimately, they still transform interval data such that the final results can be represented by a single binary logistic regression (de Souza et al., 2011). The use of least squares regression has also been used with interval data (Gioia et al., 2005; Fagundes et al., 2013), but primarily for linear regression models.

The approach proposed within this paper for dealing with interval data in logistic regressions is based on imprecise probabilities and considers the set of models rather than a single one (Walley, 1991; Manski, 2003; Ferson et al., 2007; Nguyen et al., 2012). This is similar to approaches proposed by Utkin & Coolen (2011), Wiencierz (2013) and Schollmeyer (2021) for dealing with interval uncertainty within linear regression. If separate logistic regressions are generated via maximum likelihood estimation from the interval data and displayed as cumulative distribution functions, the envelope of the extreme functions bound the true model. The primary benefit of such an approach is that it represents the existing epistemic uncertainty removed by traditional methods. Additionally, this method can also handle the case of uncertainty in discrete risk factors. The imprecise probabilities approach makes the fewest assumptions, but some statistics can be computationally challenging for large datasets (Ferson et al., 2007).

In the case of uncertainty in the outcome status used within logistic regression, traditionally, there is little that can be done but to discard these data points as they cannot be used as part of the analysis or to use a semi-supervised learning methodology (Amini & Gallinari, 2002; Chapelle et al., 2006; Chi et al., 2019). However, the proposed imprecise logistic regression technique can be used to include unlabeled examples within the dataset. Again the imprecise approach does not require making the assumptions required by other methods to fit a model.

This paper continues as follows: in Section 2 precise logistic regression is reviewed, Section 3 and Section 4 introduce the imprecise logistic regression for data with intervals within the features and the labels. In both these sections, a 1-dimensional synthetic dataset is used to demonstrate the methodology before it is compared to alternative methods on both the synthetic data and a real-world example. Finally, in Section 5 the method is used on a dataset which contains intervals in both the features and labels and is contrasted against a dataset from the literature.

## 2 Precise Logistic Regression

Let $\mathbf{x}$ be a $m$-dimensional covariate with a binary label, $y \in \{1, 0\}$. Logistic regression can be used to model the probability that $y = 1$ using:

$$\Pr(y = 1|\mathbf{x}) = \pi(\mathbf{x}) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m\right)\right)} \tag{1}$$

where $\beta_0, \beta_1, \ldots$ are a set of unknown regression coefficients. If dataset $D$ contains $n$ samples

$$D = \begin{Bmatrix} \left((x_1^{(1)} x_2^{(1)} \cdots x_m^{(1)}), y_1\right) \\ \left((x_1^{(2)} x_2^{(2)} \cdots x_m^{(2)}), y_2\right) \\ \vdots \\ \left((x_1^{(n)} x_2^{(n)} \cdots x_m^{(n)}), y_n\right) \end{Bmatrix}, \tag{2}$$

a logistic regression model can be trained on $D$, $\mathcal{LR}(D)$, by finding optimal values of $\beta_0, \beta_1, \ldots, \beta_m$ for the observed data. This is often done using *maximum likelihood estimation* (Menard, 2010; Myung, 2003), although other techniques exist, for instance through Bayesian analysis Jaakkola (1997); O'Brien & Dunson (2004).

A classification, $\hat{y}$, can be made from the logistic regression model by selecting a threshold value, $C$, and then defining

$$\hat{y} = \begin{cases} 1 & \text{if } \pi(\mathbf{x}) \geq C \\ 0 & \text{if } \pi(\mathbf{x}) < C \end{cases} \tag{3}$$

The simplest case is when $C = 0.5$, implying $\hat{y}$ is more likely to be true than false. However, this value could be different depending on the use of the model and the risk appetite of the analyst. For example, in medicine, a small threshold value may be used in order to produce a conservative classification and therefore reduce the number of false negative results. Where predictions are made within this paper, $C = 0.5$ unless otherwise stated.

### 2.1 Demonstration

To demonstrate, a synthetic 1-dimensional dataset ($X$) with a sample size of fifty was used to train a logistic regression model ($\mathcal{LR}(D)$) as shown in Figure 1. After training the model, it is useful to ask the question "how good is the model?" For logistic regression there are several ways in which that can be done, see Hosmer Jr et al. (2013, pp. 157–169) or Kleinbaum & Klein (2010, pp.318–326). For the analysis in this paper, we will consider the receiver operating characteristic graph and area under curve statistic, discriminatory performance visualisations (Royston & Altman, 2010) as well as the sensitivity and specificity of the classifications made by the algorithm.

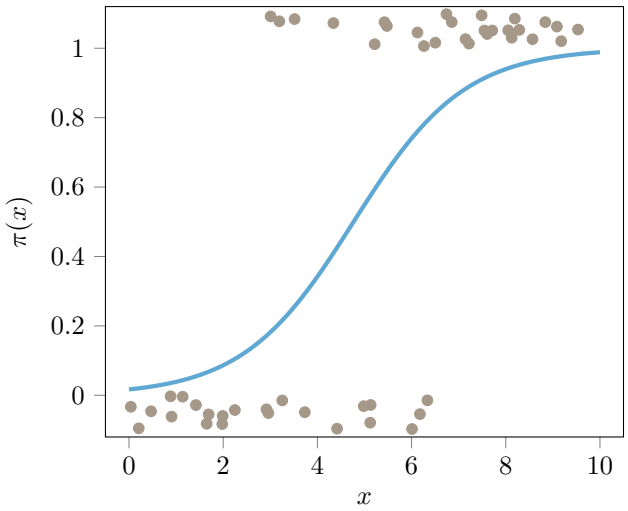
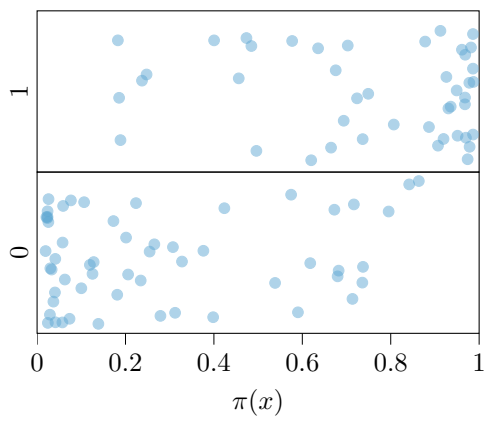
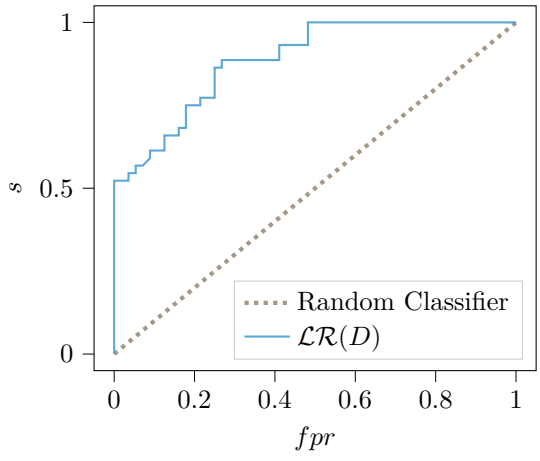



Figure 1: Logistic regression curve, $\mathcal{LR}(X)$, created by fitting the model on dataset $X$ (shown with scattered points).

Royston & Altman (2010) introduced visualisations to assess the discriminatory performance of the model by considering a scatter plot of the true outcome (jittered for clarity) vs the estimated probability. Such a plot is shown in Figure 2a. A perfectly discriminating model would have two singularities with all the points with outcome = 1 at (1,1) and all the points with outcome = 0 at (0,0). In general, the better the classifier, the more clustered the points would be towards these values, with the points on the upper band having larger probabilities and the points on the lower band having lower probabilities. From Figure 2a we can see that there is significant clustering towards the endpoints, showing that the model has excellent discriminatory performance.




(a) Scatter plot of jittered outcome vs estimated probability.




(b) ROC curve.

Figure 2: Two plots to show the discriminatory performance of the logistic regression model shown in Figure 1.

We can make and compare the predictions made from the logistic regression model using a larger dataset that has been generated using the same method as above. Tabulating these results in a confusion matrix for the base predictions is shown in Table 1.

| | 1 | 0 | Total |
|---|---|---|---|
| Predicted 1 | 34 | 14 | 48 |
| Predicted 0 | 10 | 42 | 52 |
| Total | 44 | 56 | 100 |

Table 1: Confusion matrix for 100 test data points using predictions from $\mathcal{LR}(X)$ shown in Figure 1.

There are numerous statistics than can be derived from confusion matrices to express the performance of classifiers. For this analysis, we shall consider the sensitivity $s$, the fraction of positive individuals correctly identified as such and the specificity $t$, the fraction of negative individuals correctly identified. Mathematically

$$s = \frac{\text{True Positive}}{\text{Total Number of Positives}} \tag{4}$$

and

$$t = \frac{\text{True Negative}}{\text{Total Number of Negatives}}. \tag{5}$$

From Table 1 we can calculate that $s = 0.773$ and $t = 0.750$. As confusion matrices and statistics are calculated from them depending on the cutoff value chosen ($C$ from Equation 3), a complete way of determining the classification performance of models is by considering the receiver operating characteristic (ROC) curve of the model (Kleinbaum & Klein, 2010; Hosmer Jr et al., 2013). ROC curves can be compared graphically

and by considering the area under the curve (AUC). The better the model is, the closer the AUC would be to 1. The worst possible AUC would be 0.5, as again, anything lower than that would be improved by simply switching the classification. For the ROC curve shown in Figure 2b, AUC = 0.887.

## 3 Uncertainty in Features

If there is interval uncertainty within dataset $D$,

$$D = \left\{ \begin{array}{c} \left(\left(\left[\underline{x_1^{(1)}}, \overline{x_1^{(1)}}\right] \left[\underline{x_2^{(1)}}, \overline{x_2^{(1)}}\right] \cdots \left[\underline{x_m^{(1)}}, \overline{x_m^{(1)}}\right]\right), y_1\right) \\ \left(\left(\left[\underline{x_1^{(2)}}, \overline{x_1^{(2)}}\right] \left[\underline{x_2^{(2)}}, \overline{x_2^{(2)}}\right] \cdots \left[\underline{x_m^{(2)}}, \overline{x_m^{(2)}}\right]\right), y_2\right) \\ \vdots \\ \left(\left(\left[\underline{x_1^{(n)}}, \overline{x_1^{(n)}}\right] \left[\underline{x_2^{(n)}}, \overline{x_2^{(n)}}\right] \cdots \left[\underline{x_m^{(n)}}, \overline{x_m^{(n)}}\right]\right), y_n\right) \end{array} \right\}, \tag{6}$$

and we have no more information about the true value, $x_j^{(i)\dagger}$, nor are willing to make further assumptions about the true value only that $x_j^{(i)\dagger} \in \left[\underline{x_j^{(i)}}, \overline{x_j^{(i)}}\right]$. Then it is only possible to partially identify an imprecise logistic regression model for the data, $\mathcal{ILR}(D)$:

$$\mathcal{ILR}(D) = \left\{ \mathcal{LR}\left(D'\right) : \forall D' \in \left\{ \left\{ \begin{array}{c} \left(\left(x_1'^{(1)} \cdots x_1'^{(m)}\right), y_1\right) \\ \vdots \\ \left(\left(x_n'^{(1)} \cdots x_n'^{(m)}\right), y_n\right) \end{array} \right\} \forall x_j'^{(i)} \in \left[\underline{x_j^{(i)}}, \overline{x_j^{(i)}}\right] \right\} \right\} \tag{7}$$

i.e. $\mathcal{ILR}(D)$ is the set of all possible logistic regression models that can be created from all possible datasets that can be constructed from the interval data; this ensures that the true logistic regression model, $\mathcal{LR}^\dagger$, is contained within the set. For continuous data, this set is infinitely large.

Predictions can be made by sampling all the possible models that are contained within $\mathcal{ILR}(D)$ and creating an interval containing the maximum and minimum values, $\pi(\mathbf{x}) = \left[\underline{\pi(\mathbf{x})}, \overline{\pi(\mathbf{x})}\right]$. As such, it is only necessary to find the logistic regression models that correspond to extreme values of $\pi(\mathbf{x})$ for all $\pi(\mathbf{x})$. Therefore, it is possible to reduce the number of models that need to be contained within the set to

$$\begin{aligned} \mathcal{ILR}(D) = &\left\{ \mathcal{LR}\left(D'_{\underline{\beta_i}}\right), \mathcal{LR}\left(D'_{\overline{\beta_i}}\right) \ \forall i = 0, 1, \ldots, m \right\} \\ &\cup \left\{ \mathcal{LR}\left(\underline{D}\right), \mathcal{LR}\left(\overline{D}\right) \right\} \end{aligned} \tag{8}$$

where $D'_{\underline{\beta_0}}$ is the dataset constructed from points within the intervals so that the value of $\beta_0$ is minimised, $D'_{\overline{\beta_0}}$ is the dataset constructed from points within the intervals so that the value of $\beta_0$ is maximised, and so on. $\underline{D}$ corresponds to the dataset created by taking the lower bound of every interval within $D$, similarly for $\overline{D}$. For a dataset with $m$ features, there are $2 + 2m$ models that are needed to find the bounds of the set. Algorithm 1 can be used to find this set.

**Algorithm 1:** Algorithm to find $\mathcal{ILR}(D)$ if $D$ has interval uncertainty within its features.

**Input:** $D = \left\{\left(\left(\left[\underline{x_j^{(i)}}, \overline{x_j^{(i)}}\right] \ \forall i = 0, \ldots, m\right), y_j\right) \ \ \forall j = 0, \ldots, n\right\}$
$\underline{D} = \left\{\left(\left(\underline{x_j^{(i)}} \ \forall i = 0, \ldots, m\right), y_j\right) \ \ \forall j = 0, \ldots, n\right\}$;
$\overline{D} = \left\{\left(\left(\overline{x_j^{(i)}} \ \forall i = 0, \ldots, m\right), y_j\right) \ \ \forall j = 0, \ldots, n\right\}$;
$\mathcal{ILR}(D) \leftarrow \left\{\mathcal{LR}\left(\underline{D}\right), \mathcal{LR}\left(\overline{D}\right)\right\}$;
**for** $i \leftarrow 0$ **to** $m$ **do**
  Using stochastic optimisation find $D'_{\underline{\beta_i}}$ such that $\mathcal{LR}\left(D'_{\underline{\beta_i}}\right)$ has the minimum value of $\beta_i$;
  $\mathcal{ILR}(D) \leftarrow \mathcal{ILR}(D) \cup \{\mathcal{LR}(D')\}$;
  Using stochastic optimisation find $D'_{\overline{\beta_i}}$ such that $\mathcal{LR}\left(D'_{\overline{\beta_i}}\right)$ has the maximum value of $\beta_i$;
  $\mathcal{ILR}(D) \leftarrow \mathcal{ILR}(D) \cup \{\mathcal{LR}(D')\}$;
**end**
**Output:** $\mathcal{ILR}(D)$

If the dataset contains a large number of intervals, then the optimisation itself may take a prohibitively long time to find $\underline{\beta_i}$, $\overline{\beta_i}$. In such a situation, Algorithm 2 can be used to find the imprecise model. This algorithm minimises and maximises the spread of the values around various points in order to find the minimum and maximum $\beta$ values.

**Algorithm 2:** Algorithm to estimate $\mathcal{ILR}(D)$ if $D$ has interval uncertainty within its features.

**Input:** $D = \left\{\left(\left(\left[\underline{x_j^{(i)}}, \overline{x_j^{(i)}}\right] \ \forall i = 0, \ldots, m\right), y_j\right) \ \ \forall j = 0, \ldots, n\right\}$
$\underline{D} = \left\{\left(\left(\underline{x_j^{(i)}} \ \forall i = 0, \ldots, m\right), y_j\right) \ \ \forall j = 0, \ldots, n\right\}$;
$\overline{D} = \left\{\left(\left(\overline{x_j^{(i)}} \ \forall i = 0, \ldots, m\right), y_j\right) \ \ \forall j = 0, \ldots, n\right\}$;
$\mathcal{ILR}(D) \leftarrow \left\{\mathcal{LR}\left(\underline{D}\right), \mathcal{LR}\left(\overline{D}\right)\right\}$;
**for** *many* $P$ **do**
  Find $T$ such that $\pi_{\mathcal{LR}(\underline{D})}(T) = P$;
  $d_{min} \leftarrow D$;
  $d_{max} \leftarrow D$;
  **for all** *Interval* $(I = \left[\underline{i}, \overline{i}\right])$ **in** $D$ **do**
    **if** $T \in P$ **then**
      $d_{min}(I) = T$;
      **if** $|\underline{i} - T| > |\overline{i} - T|$ **then** $d_{max}(I) = \underline{i}$ **else** $d_{max}(I) = \overline{i}$;
    **else if** $T < P$ **then**
      $d_{min}(I) = \underline{i}$;
      $d_{max}(I) = \overline{i}$;
    **else**
      $d_{min}(I) = \overline{i}$;
      $d_{max}(I) = \underline{i}$;
    **end**
  **end**
  $\mathcal{ILR}(D) \leftarrow \mathcal{ILR}(D) \cup \{\mathcal{LR}(cud_{min}), \mathcal{LR}(d_{max})\}$;
**end**
**Output:** $\mathcal{ILR}(D)$

When calculating the probability of a value being 1 under the imprecise model, there is an interval probability $\left[\underline{\pi(\mathrm{x})}, \overline{\pi(\mathrm{x})}\right]$. As such, when using the model to perform classifications, this interval means that Equation 3

becomes

$$\hat{y} = \begin{cases} 1 & \text{if } \underline{\pi}(\mathbf{x}) > C \\ 0 & \text{if } \overline{\pi}(\mathbf{x}) < C \\ [0,1] & \text{if } C \left[\underline{\pi(\mathbf{x})}, \overline{\pi(\mathbf{x})}\right] \end{cases} \tag{9}$$

The final line of this equation returns the *dunno* interval, meaning there is uncertainty in determining whether the datum should be predicted 0 or 1, and the model should therefore abstain from providing a classification. It is left up to the analysis to decide what to do with such a result.

Under such a scenario, there are two approaches to characterising the classifier using a confusion matrix. The first is to consider the intervals directly within the confusion matrix. Calculating statistics derived from the confusion matrix requires careful handling for interval calculations (Gray et al., 2022).

To demonstrate, let $a$ be the number of true positives, $b$ be the number of false positives, $c$ the number of false negatives and $d$ the number of true negatives, where $a$, $b$, $c$ and $d$ are all intervals ($a = [\underline{a}, \overline{a}]$ etc). Since the number of positive individuals is fixed, $a$ and $c$ are oppositely dependent on each other (as $a \to \overline{a}$ then $c \to \underline{c}$), as are $b$ and $d$. These dependencies imply that care needs to be taken when calculating the sensitivity and specificity. Sensitivity needs to be calculated as

$$s = \frac{a}{\underline{a} + \overline{c}}, \tag{10}$$

and specificity as

$$t = \frac{d}{\underline{b} + \overline{d}}. \tag{11}$$

In this instance, the $s$ and $t$ would themselves both be intervals ($s = [\underline{s}, \overline{s}]$ and $t = \left[\underline{t}, \overline{t}\right]$). $\overline{s}$ and $\overline{t}$ would be the sensitivity and specificity of a system that first perfectly classified all the uncertain predictions.

Off course, a whole host of alternative statistics could be used to describe the classifier's performance, some of which require special care when using imprecise numbers. For instance, *precision* and *recall* are often used within the machine learning literature to assess classifiers. Precision is the fraction of positive predictions that are true positives,

$$\text{precision} = \frac{\text{True Positive}}{\text{Total Number of Positive Predictions}}, \tag{12}$$

and recall is analogous to sensitivity. Often quoted alongside these statistics is the $F_1$ score, which is the harmonic mean of these values,

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{13}$$

With intervals, precision needs to be calculated using a single-use expression to ensure that there is no artifactual uncertainty within the calculation

$$\text{precision} = \frac{1}{1 + \frac{\underline{b}}{\underline{a}}} \tag{14}$$

and recall calculated using Equation 10.

The $F_1$ score is again best calculated through a single-use expression

$$F_1 = 2\frac{1}{\frac{\underline{TP}+\underline{b}}{\underline{a}} + 1} \tag{15}$$

where $TP$ is the total number of positive predictions ($TP = \underline{a} + \overline{c}$).

An alternative approach to constructing a confusion matrix with uncertain predictions is to tabulate the uncertain predictions in a separate row. If the model returned $u$ true positives, $v$ false positives, $w$ false negatives and $x$ true negatives but did not make a prediction for $y$ positives and $z$ negatives, then the confusion matrix shown in Table 2 can be created.

From this confusion matrix, some useful statistics can be calculated to account for the uncertainty produced by these uncertain classifications. The first is to consider that the traditional definitions of sensitivity and specificity can be could be re-imagined by defining what the *predictive sensitivity* $s'$ as the sensitivity out of the points for which a prediction was made

$$s' = \frac{u}{u+w} \tag{16}$$

and similarly, the *predictive specificity* $t'$ as the specificity for which a prediction was made

$$t' = \frac{x}{v+x}. \tag{17}$$

| | 1 | 0 | Total |
|---|---|---|---|
| Predicted 1 | $u$ | $v$ | $P_+$ |
| Predicted 0 | $w$ | $x$ | $P_-$ |
| No Prediction | $y$ | $z$ | $P_\times$ |
| Total | $T_+$ | $T_-$ | $N$ |

Table 2: Alternative confusions matrix where uncertain predictions are tabulated separately.

Two other statistics are useful to describe the data in Table 2. We can define the *positive incertitude* $\sigma$ to be the fraction of positive cases for which the model could not make a prediction

$$\sigma = \frac{y}{u+w+y}. \tag{18}$$

Similarly, the *negative incertitude* $\tau$ can be defined as the total number of negative cases for which the model could not make a prediction

$$\tau = \frac{z}{v+x+z}. \tag{19}$$

### 3.1 Alternative Methods

Several authors have suggested different approaches to compute logistic regression models with interval uncertainty. This section will consider three methods to compare with the method presented within this paper.

#### Midpoint

The most straightforward approach to dealing with interval data is to produce a precise dataset by replacing the intervals with their midpoints and then fitting a dataset on this midpoint data. i.e.

$$D_m = \left\{ \left( \left( \frac{\underline{x_j^{(i)}} + \overline{x_j^{(i)}}}{2} \right), y_j \right) \ \forall \left[ \underline{x_j^{(i)}}, \overline{x_j^{(i)}} \right] \in D \right\} \tag{20}$$

then

$$\mathcal{LR}_m(D) = \mathcal{LR}(D_m). \tag{21}$$

This model can then be used as a precise logistic regression model described in Section 2.

#### de Souza

de Souza et al. (2008; 2011) present a method for characterising the uncertainty with . Their method performs two separate logistic regressions on the lower and upper bounds of the intervals and averages the posterior probabilities to obtain a pooled posterior probability. They find

$$\mathcal{LR}_{dS}(D) = \left\{ \mathcal{LR}(\underline{D}), \mathcal{LR}\left(\overline{D}\right) \right\} \tag{22}$$

and then reduces this to a single logistic regression model based upon the average of the outputted probabilities.

$$\pi_{\mathcal{LR}_{dS}(D)}(\mathbf{x}) = \frac{\pi_{\mathcal{LR}(\underline{D})}(\mathbf{x}) + \pi_{\mathcal{LR}(\overline{D})}(\mathbf{x})}{2} \tag{23}$$

#### Billard-Diday

Billard & Diday (2000) propose a method, based upon Bertrand (2000), for characterising interval uncertainties within linear regression that can be easily extended to logistic regression. Their method assumes that each value from within the interval is equally likely, and therefore constructs consider the logistic regression models as the uniform mixture of $N$ logistic regression models that are fitted from random samples,

$$\mathcal{LR}_{BD}(D) = \left\{ \mathcal{LR}(D_k) : D_k = \left\{ \left\{ \left( (r_j^{(i)}), y_j \right) \right\} \; r_j'^{(i)} \in \left[ \underline{x_j^{(i)}}, \overline{x_j^{(i)}} \right] \right\} \; k = 0, \ldots, N \right\} \tag{24}$$

like de Souza they then average the probability from all models

$$\pi_{\mathcal{LR}_{BD}(D)}(\mathbf{x}) = \frac{1}{N} \sum_{\forall l \in \mathcal{LR}_{BD}(D)} \pi_l(\mathbf{x}). \tag{25}$$

### 3.2 Demonstration

Dataset $X$ from 2 has been intervalised into dataset $Y$ using the following transformation $x_i' = [m - \epsilon, m + \epsilon]$ where $m$ is a number drawn from the triangular distribution $\mathrm{T}(x_i - \epsilon, x_i + \frac{\epsilon}{6}, x_i + \epsilon)$ with $\epsilon = 0.375$ for all $x_i \in X$. With this dataset we can use Algorithm 1 to construct $\mathcal{ILR}(Y)$, as is shown in Figure 3. In this plot, the bounds on the model are shown by the black dashed lines. The coloured lines represent the six extreme logistic regression models within the set shown in Equation 8. This plot also shows 60 logistic regression models generated by fitting on random samples from within the intervals. This plot demonstrates that Equation 8 can be used to estimate the extreme values of the set of models.

For comparison, logistic regression models have been fitted using the methods described in Section 3; these models are shown in Figure 4. Whilst there are subtle differences between the different approaches, it is clear that they are all approximately equal. This is unsurprising as they all implicitly make the equidistribution assumption that the interval data points can be represented by a uniform distribution.

It is also possible to consider situations where data is intervalised differently. Figure 5 shows four different intervalisations of dataset $X$ as described in Table 3. In plot (a), the intervalisation has occurred by placing the true value at the left edge of the interval; similarly, in (b), the true value is the right edge of the interval. In (c), the value of $x$ impacts the width of the interval. In (d), the label impacts the intervalisation.

In Figure 5, an imprecise model has been fitted on the dataset alongside the 'true' model (from Figure 1) and the midpoint model. The de Souza and Billard-Diday methods are not shown due to the similarity with the midpoint model. Looking at all these figures, we see that the imprecise model always bounds the base model. As a result, any interval regression analysis performed is guaranteed to bound the true model. There can be significant differences between the base and midpoint model. The alternative approaches provide a good approximation of the true value if the equidistribution hypothesis can be justified, as in plot (c). If the intervalisation depends on the outcome, then the approaches appear inadequate, as is shown in plot (d). This implies that the alternative approaches should only be used in cases within which one can assume that the data has been intervalised independent of either the true underlying value or the outcome status and each value within the interval is equally likely. In many real-world datasets, this will not be the case, and in those scenarios, only the imprecise method guarantees coverage of the true model.

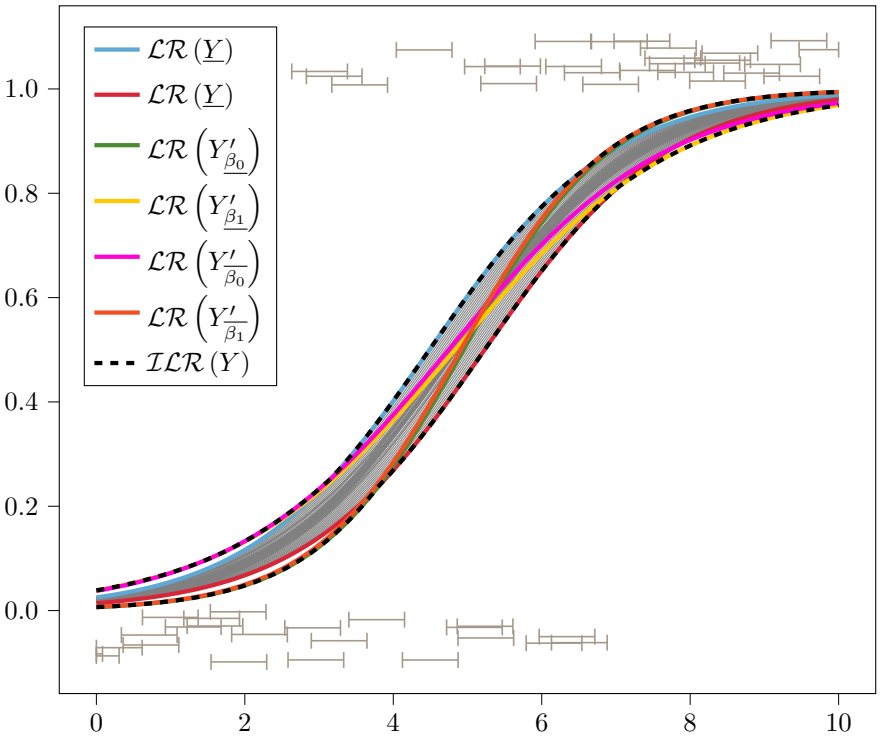



Figure 3: Imprecise logistic regression model (dashed lines) for the interval data (jittered for clarity). The colour lines represent the six different models that have been fitted using Algorithm 1, whereas the black bounds have been fitted using Algorithm 2. The grey lines represent numerous Monte Carlo samples generated by taking random values from within the intervals and fitting logistic regression models.

| Plot | Intervalisation |
|---|---|
| (a) | $Y_a = \{([x_i, x_i + 2], y_i) \ \forall (x_i, y_i) \in X\}$ |
| (b) | $Y_b = \{([x_i - 2, x_i], y_i) \ \forall (x_i, y_i) \in X\}$ |
| (c) | $Y_c = \left\{ \left( \begin{cases} x_i \text{ if } x_i < 2.5 \\ [m - 0.25, m + 0.25], \ m \in U(x_i - 0.25, x_i + 0.25) \text{ if } 2.5 \leq x_i < 5.0 \\ [m - 0.75, m + 0.75], \ m \in U(x_i - 0.75, x_i + 0.75) \text{ if } 5.0 \leq x_i < 7.5 \\ [x_i, x_i + 1.5], \text{ if } 7.5 < x_i \end{cases}, y_i \right) \forall (x_i, y_i) \in X \right\}$ |
| (d) | $Y_d = \left\{ \left( \begin{cases} [x_i, x_i + 1.5] \text{ if } y_i = 1 \\ [x_i - 1.5, x_i] \text{ otherwise} \end{cases}, y_i \right) \forall (x_i, y_i) \in X \right\}$ |

Table 3: Intervalisations used in Figure 5

### 3.3 Red Wine Example

In order to demonstrate the methodology on a real-world dataset, we can use the red wine dataset from Cortez et al. (2009). This dataset contains 11 covariates[1] that can be used to predict the quality of the wine sample based on a scale from 0 to 10. In order to provide a binary classification, we define wine as good if it has quality $\geq 6$. The dataset contains 1599 samples, of which 855 have been classified as good wine. The dataset with added uncertainties is denoted as $R$.

In order to fit the logistic regression model, the dataset has been split into training and test subsets containing half the data in each. To intervalise the data, the covariates have been turned into intervals based upon the

[1]fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol

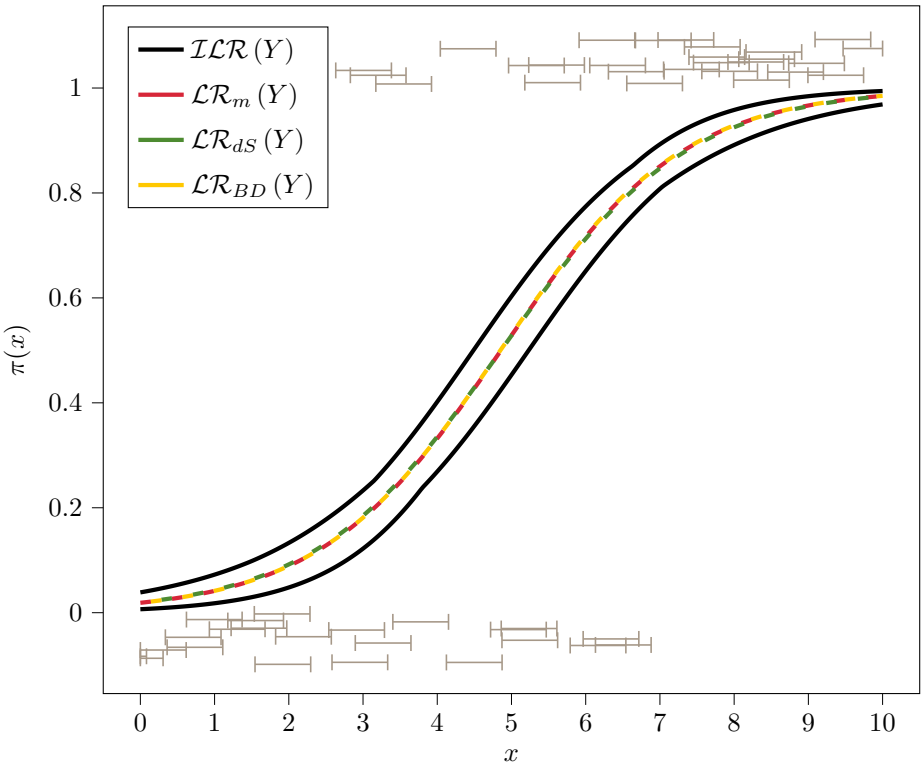



Figure 4: Imprecise logistic regression model (black lines) for the interval dataset (jittered for clarity). The midpoint approach is shown using the red line, the de Souza approach with the green line and the Billard-Diday approach with the orange line.

number of significant figures within the model (for example, 0.5 would become $[0.45, 0.55]$). An imprecise model can then be fitted on the dataset using Algorithm 2.

It is helpful to consider visualisations when discussing the classifier's performance (Figure 6). The simplest of these is the scatter plots shown in Figure 6a. From the plot, the majority of the wines rated as good were given a high $\pi$ value, and vice versa for the bad wines - although no wine was given a very low *pi*. There is, however, substantial overlap between the two groups. The plot also shows that some wines have a wide interval $\pi$. The plot also shows the size of the intervals for $\pi$. In this instance, all of the intervals are reasonably consistent and not overly wide.

We can also construct ROC plots and calculate their AUCs, as shown in Figure 6b. In this plot, we can see that the $\mathcal{LR}\left(R^{\dagger}\right)$ and $\mathcal{LR}_m(R)$ curves are only subtly distinguishable from each other. The same is also true of $\mathcal{LR}_{dS}(R)$ and $\mathcal{LR}_{BD}(R)$. The imprecise model bounds all these models. Additionally, for the imprecise model, we can plot a ROC curve for when the model abstains from making a prediction. In this instance $s'$ is plotted against $fpr'(= 1 - t')$. This ROC curve outperforms the others.

The AUC for the curves are $AUC\left[\mathcal{ILR}(R)\right] = [0.742, 0.872]$, $AUC\left[\mathcal{ILR}(R)(\mathit{Abstain})\right] = 0.834$, $AUC[\mathcal{LR}\left(R^{\dagger}\right)] \approx AUC\left[\mathcal{LR}_m(R)\right] = 0.818$ and $AUC[\mathcal{LR}_{dS}(R)] \approx AUC[\mathcal{LR}_{BD}(R)] = 0.813$.

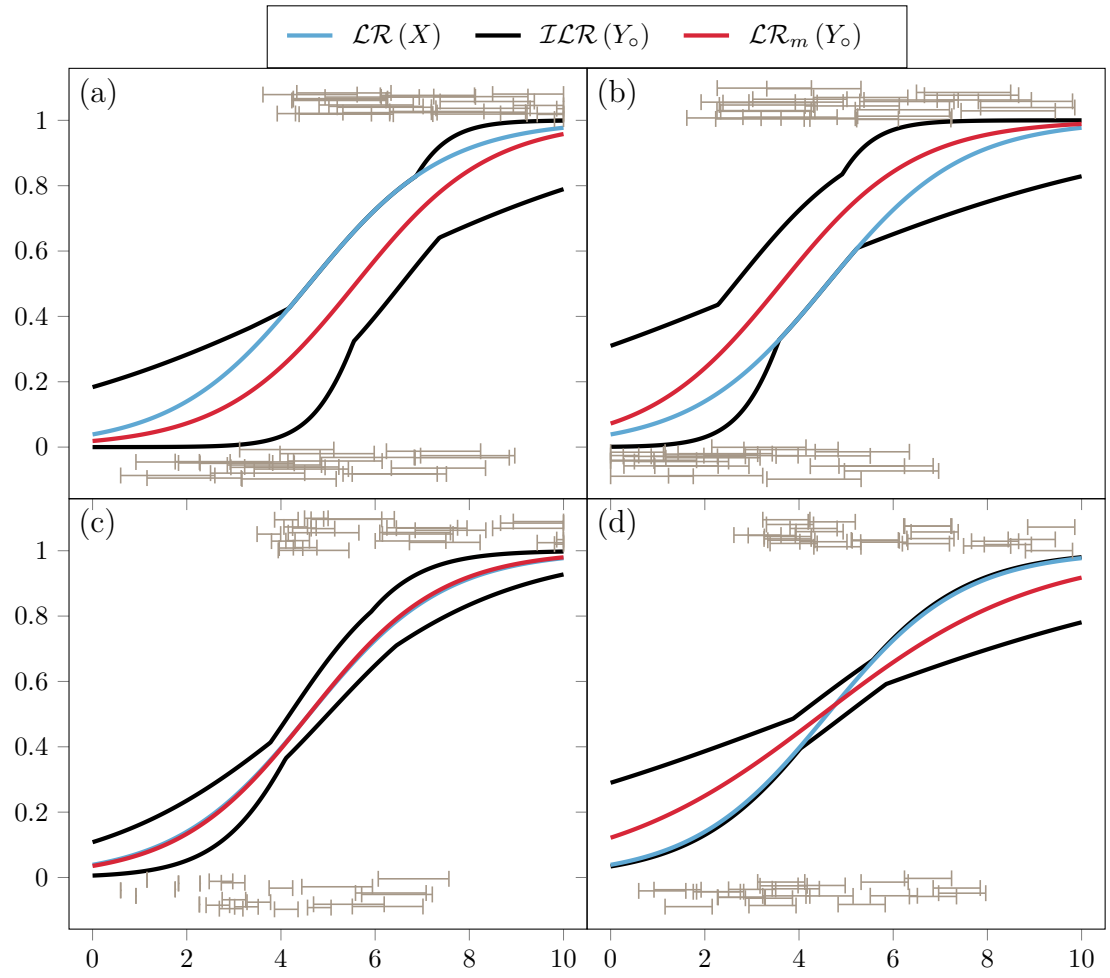
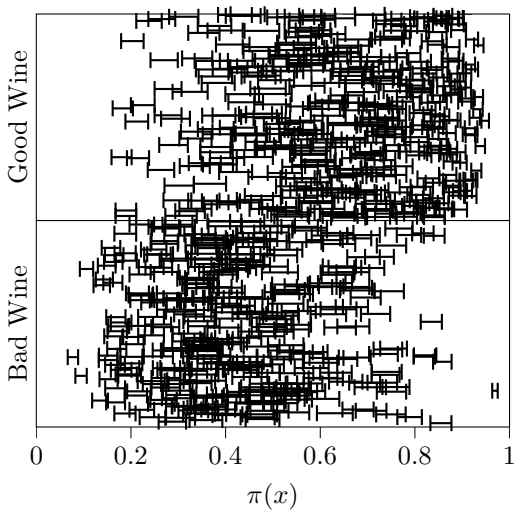
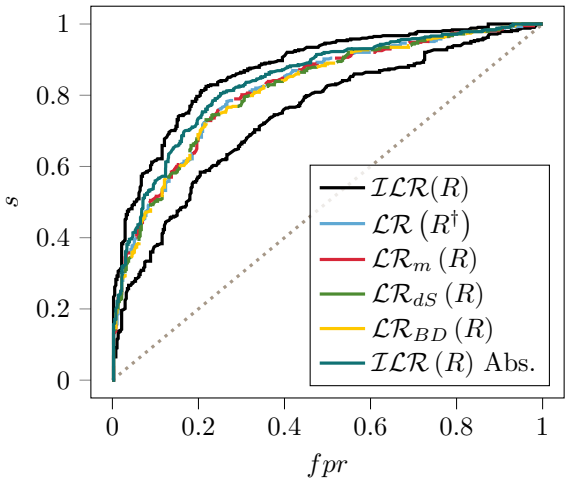



Figure 5: Logistic Regression models fitted on interval data that has been intervalised in the biased ways shown in Table 3. In each case $\mathcal{LR}(D)$ is the precise model shown in Figure 1.




(a) Scatter plots of probability vs outcome for $\mathcal{ILR}(R)$




(b) Receiver operating characteristic curve for the simple example with added uncertain classifications.

Figure 6: Plots to show the discriminatory performance of the logistic regression models for the red wine example.

Classifications about whether a wine is good or not can be made. Selecting a threshold value of 0.5 gives the confusion matrices shown in Table 4. There are two possible interpretations of how the intervals can be expressed within the confusion matrices shown in a the intervals are plotted directly within the matrix giving the following statistics: $s = [0.652, 0.769]$ and $t = [0.707, 0.817]$. Allowing the model to abstain as in b implies that there are 91 wines for which a prediction could not be made solely as a result of the imprecision of the model. The summary statistics from this confusion matrix are: $s' = 0.738$, $t' = 0.795$, $\sigma = 0.117$ and $\tau = 0.110$.

| | Good Wine | Bad Wine | Total |
|---|---|---|---|
| Predicted Good | [279,329] | [68,109] | [347,438] |
| Predicted Bad | [99,149] | [263,304] | [362,453] |
| Total | 428 | 372 | 800 |

(a) Tabulating inconclusive results as [0,1] intervals.

| | Good Wine | Bad Wine | Total |
|---|---|---|---|
| Predicted Good | 279 | 68 | 347 |
| Predicted Bad | 99 | 263 | 362 |
| No Prediction | 50 | 41 | 91 |
| Total | 428 | 372 | 800 |

(b) Tabulating inconclusive results separately.

Table 4: Two possible confusion matrices for 100 test samples from the imprecise logistic regression model shown in Figure 4.

## 4 Uncertainty in Labels

This set-based approach can be extended to the situation where there is uncertainty about the outcome status meaning there are some points for which we do not know the binary classification and can be represented as the dunno interval $[0, 1]$. In this situation the dataset $\mathcal{D}$ contains $p$ variables with corresponding labels $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_p, y_p)$ but also $q$ variables for which the label is unknown $(\mathbf{x}_{p+1}, \ ), (\mathbf{x}_{p+2}, \ ), \cdots, (\mathbf{x}_{p+q}, \ )$. For simplicity, we shall refer to the points that have labels as being in set $d$ and those without labels in set $u$, $D = d \cup u$.

Traditional analysis may ignore all the points in $u$. However, they can be included within the analysis by considering the set of possible logistic regression models trained on all possible datasets that could be possible based upon the uncertainty. This set of datasets can be created by giving all unlabeled values the value 0, all unlabeled values the value 1 and all combinations thereof, i.e.

$$\mathcal{ILR} = \left\{ \mathcal{LR}(D') \ \forall D' \in \left\{ \begin{array}{c} d \cup \{(\mathbf{x}_{p+1}, 0), \cdots, (\mathbf{x}_{p+q}, 0)\} \\ d \cup \{(\mathbf{x}_{p+1}, 0), \cdots, (\mathbf{x}_{p+q}, 1)\} \\ \vdots \\ d \cup \{(\mathbf{x}_{p+1}, 1), \cdots, (\mathbf{x}_{p+q}, 0)\} \\ d \cup \{(\mathbf{x}_{p+1}, 1), \cdots, (\mathbf{x}_{p+q}, 1)\} \end{array} \right\} \right\} \tag{26}$$

This leads to $2^q$ possible logistic regression models. An imprecise logistic regression model can then be created by finding the envelope of the set, as shown in Algorithm 3. As the computational time for this algorithm increases as $\mathcal{O}(2^q)$, then as $q$ increases finding the bounds by calculating the envelope for all possible combinations can become computationally expensive.

**Algorithm 3:** Algorithm to find $\mathcal{ILR}(D)$ if $D$ has interval uncertainty within its labels.

**Input:** $d = \{(\mathbf{x}_i, y_i)\ \forall i = 0, \ldots, p\}$; $u = \{(\mathbf{x}_j, [0,1])\ \forall j = 0, \ldots, q\}$; $D = d \cup u$
$\mathcal{ILR}(D) \leftarrow \{\}$;
**for all combinations** $C \in \{(0, \ldots, 0), (0, \ldots, 1), (1, \ldots, 0), (1, \ldots, 1), \cdots\}$ **do**
  $D' \leftarrow d \cup \{(\mathbf{x}_j, C_j)\ \forall \mathbf{x}_j \in u\}$;
  $\mathcal{ILR}(D) \leftarrow \mathcal{ILR}(D) \cup \mathcal{LR}(D')$;
**end**
**Output:** $\mathcal{ILR}(D)$

Algorithm 4 reduces the number of iterations models that are needed to be fitted in order to find an *esitmate* for the imprecise bounds. This algorithm first finds the logistic regression model assuming all uncertain labels are 0 and the logistic regression model assuming all uncertain labels are 1. The uncertain points are split into three groups: $G_1$ contains points for which have a low $\pi$ value with both models, $G_2$ contains points for which have a high $\pi$ value with both models and all other points in $G_3$. The algorithm assumes that the most extreme models can be found by giving all the points in $G_1$ the same label, all the points in $G_2$ the same label and only finding all possible combinations of labels for the points within $G_3$. This reduces the number of logistic regression models fitted to $2 + 2^{2+q'}$ where $q'$ is the number of points in $G_3$.

**Algorithm 4:** Algorithm to find $\mathcal{ILR}(D)$ if $D$ has interval uncertainty within its labels using heuristics to reduce the number of iterations needed.

**Input:** $d = \{(\mathbf{x}_i, y_i)\ \forall i = 0, \ldots, p\}$; $u = \{(\mathbf{x}_j, y_j = [0,1])\ \forall j = 0, \ldots, q\}$; $D = d \cup u$
$D_{(1,\ldots,1)} \leftarrow d \cup \{(\mathbf{x}_j, 1)\ \forall \mathbf{x}_j \in u\}$;
$D_{(0,\ldots,0)} \leftarrow d \cup \{(\mathbf{x}_j, 0)\ \forall \mathbf{x}_j \in u\}$;
Find $\mathcal{LR}\left(D_{(1,\ldots,1)}\right)$ and $\mathcal{LR}\left(D_{(0,\ldots,0)}\right)$;
$G_1 \leftarrow \{\}$; $G_2 \leftarrow \{\}$; $G_3 \leftarrow \{\}$;
**for all** $u_j = (\mathbf{x}_j, y_j) \in u$ **do**
  **if** $\pi_{\mathcal{LR}(D_{1,\ldots,1})}(\mathbf{x}_i) < 0.5$ *and* $\pi_{\mathcal{LR}(D_{0,\ldots,0})}(\mathbf{x}_i) < 0.5$ **then**
    $G_1 \leftarrow G_1 \cup \{u_j\}$
  **else if** $\pi_{\mathcal{LR}(D_{1,\ldots,1})}(\mathbf{x}_i) > 0.5$ *and* $\pi_{\mathcal{LR}(D_{0,\ldots,0})}(\mathbf{x}_i) > 0.5$ **then**
    $G_2 \leftarrow G_2 \cup \{u_j\}$
  **else**
    $G_3 \leftarrow G_3 \cup \{u_j\}$
  **end**
**end**
**for all** $A$ *in* $\{0,1\}$ **do**
  **for all** $B$ *in* $\{0,1\}$ **do**
    **for all combinations** $C \in \{(0, \ldots, 0), (0, \ldots, 1), (1, \ldots, 0), (1, \ldots, 1), \cdots\}$ **do**
      $D' \leftarrow d \cup \{(\mathbf{x}_j, A)\ \forall (\mathbf{x}_j, y_j) \in G_1\} \cup \{(\mathbf{x}_j, B)\ \forall (\mathbf{x}_j, y_j) \in G_2\} \cup \{(\mathbf{x}_j, C_j)\ \forall (\mathbf{x}_j, y_j) \in G_3\}$;
      $\mathcal{ILR}(D) \leftarrow \mathcal{ILR}(D) \cup \mathcal{LR}(D')$;
    **end**
  **end**
**end**
**Output:** $\mathcal{ILR}(D)$

### 4.1 Alternative Methods

#### Exclude Uncertain Results

The most straightforward approach to dealing with this is to remove the uncertain results from $D$ to produce $D_\times$ and a precise logistic regression model $\mathcal{LR}_\times(D)$. This approach may be valid if the missing data is small compared to the total size of the data and if it is missing at random or completely at random.

### Semi-Supervised Logistic Regression

Semi-supervised learning methods extend supervised learning techniques to cope with additional unlabeled data. Numerous authors present semi-supervised logistic regression methods based upon a variety of different methods: Amini & Gallinari (2002) use Classification Expectation Maximisation; Krishnapuram et al. (2004) and Chi et al. (2019) use Bayesian methods; Bzdok et al. (2015) use an autoencoder and a factored logistic regression model; Chai et al. (2018) combine active learning and semi-supervised learning to " achieve better performances compared to the widely used semi-supervised learning and active learning methods." In all cases, it is important that the smoothness, clustering and manifold assumptions are valid to use semi-supervised learning techniques (Chapelle et al., 2006).

For this analysis, we have used the scikit-learns semi-supervised learning algorithm, which uses Yarowsky's algorithm to enable the logistic regression to learn from the unlabeled data (Yarowsky, 1995; Pedregosa et al., 2011).

## 4.2 Demonstration

Dataset $Z$ has been created from dataset $X$ by replacing five labels from the dataset with the $[0, 1]$ interval. The labels that have been changed are around the point at which the data goes from 0 to 1. This dataset is shown in Figure 7, with the uncertain labels plotted as the vertical lines. The figure shows the all logistic regression models that has been fitted using both Algorithms 3 (grey lines) and 4 (coloured). $\mathcal{ILR}(Z)$ is the envelope of these sets and is shown with the black dashed lines. Since the black lines always correspond to the extremum of the colour lines, Algorithm 4 has correctly estimated the imprecise bounds and any interval $\pi$ value found from imprecise models is guaranteed to contain the true value.

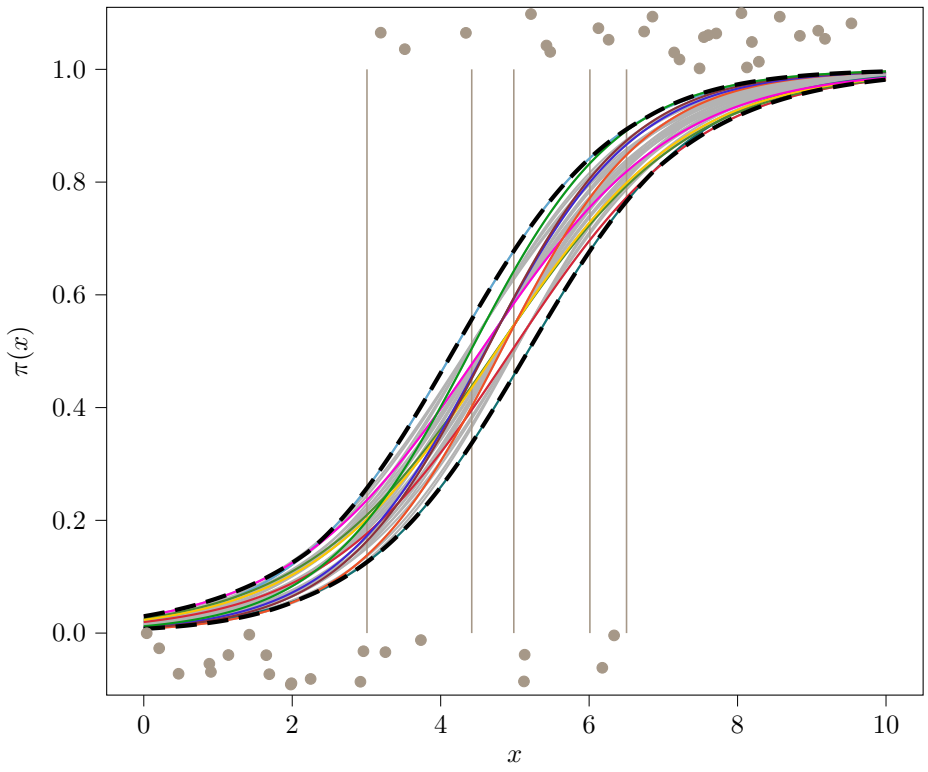



Figure 7: Imprecise logistic regression model with the uncertain labels represented by the vertical lines. The grey lines represent all possible models found using Algorithm 3. The coloured lines represent those found using Algorithm 4. The black dashed lines are the bounds of the model.

Figure 8 shows the imprecise logistic regression model that is trained on this uncertain dataset, and, for comparison, the model trained on the dataset with the dunno labels removed, $\mathcal{LR}_{\times}(Z)$, and the semi-supervised model, $\mathcal{LR}_{ss}(Z)$. From the figure, it is also notable that $\mathcal{LR}_{\times}(Z)$ and $\mathcal{LR}_{ss}(Z)$ are similar.

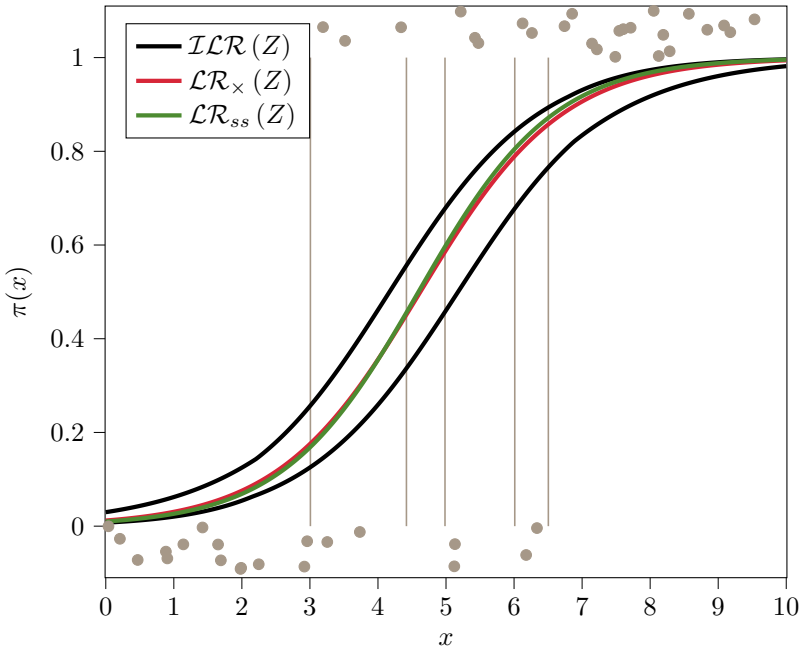



Figure 8: Bounds for the imprecise logistic regression or all 50 points with 5 points made uncertain (shown with vertical lines with the true values shown with black diamonds). Compared with $\mathcal{LR}_{\times}(Z)$ and $\mathcal{LR}_{ss}(Z)$.

As in Section 3.2, it is helpful to consider different scenarios within which the labels have been removed. Figure 9 shows four different scenarios within which the data has been made uncertain, and the imprecise logistic regression models have been fitted using Algorithm 4. Using this algorithm allowed plot (b) to be computed since Algorithm 3 would have required $2^{20}$ models to be fitted and have been computationally prohibitive. In all four plots, the imprecise model bounds the base model. It is also notable that the semi-supervised and discarded data approaches are similar in all the plots.

This plot demonstrates that if it can be assumed that the data is missing at random as in (a) and (b), the two alternative approaches are reasonably close to the true model. However, if this is not the case, then significant differences between the approaches and the imprecise method must be used to obtain a model that is guaranteed to bound the true model.

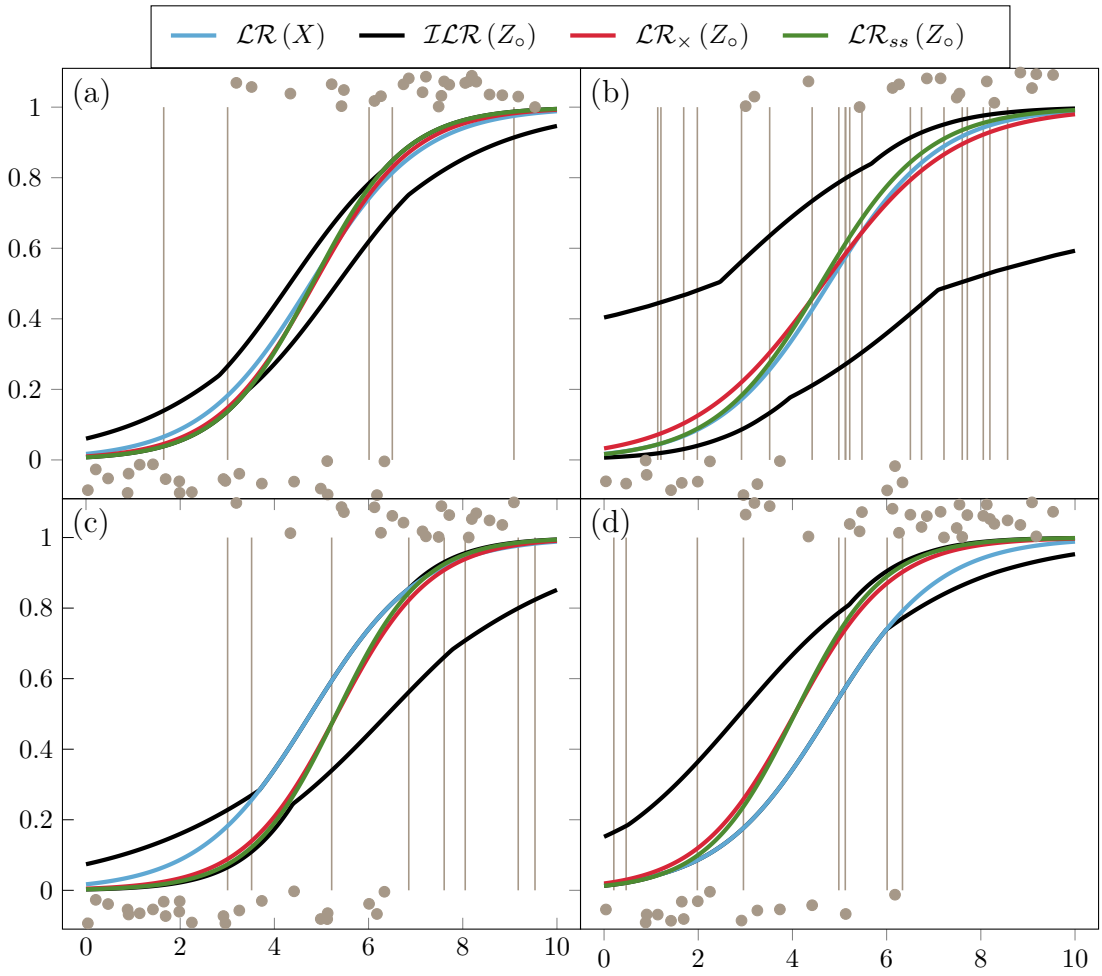
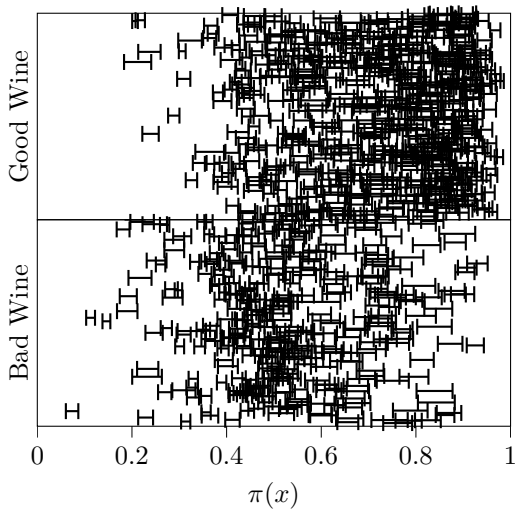



Figure 9: 4 different scenarios within which dataset $X$ has had labels removed. In (a) 5 labels have been removed at random, in (b) 20 labels have been removed at random, in (c) 8 labels have been removed from 1 labels only, and in (d) 8 labels have been removed from 0 labels only.

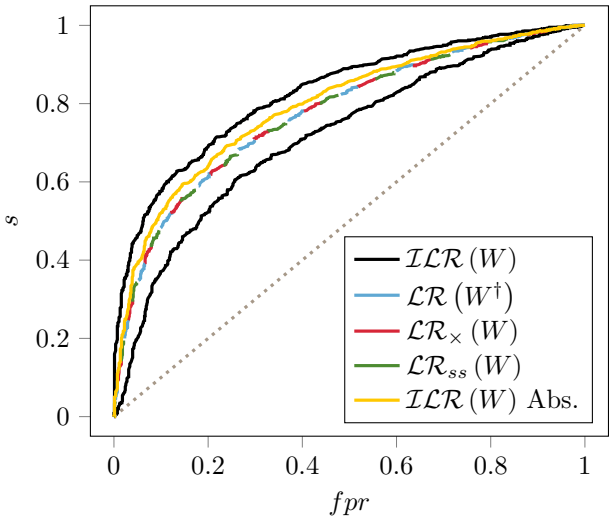
### 4.3 White Wine Example

As in Section 3.3 it is useful to consider this methodology on a real dataset. In this instance, we will use the white wine dataset from Cortez et al. (2009). This dataset contains the same covariates as the red wine dataset used in Section 3.3 but contains many more samples (4898). Again good wine has been defined as having a quality $\geq 6$. This data has again been split into training and test samples with 1618 and 3281 samples in each. In order to simulate sommeliers being unsure about the classification of marginally good wine, 100 samples with quality 6 have had their labels removed. Let $W$ be the uncertain dataset. Algorithm 4 can be used to fit the imprecise logistic regression model on this dataset. For comparison, $\mathcal{LR}_{ss}(W)$ and $\mathcal{LR}_\times(W)$ have also be found.

The discrimination plots for these models are shown in Figure 10. Figure 10a shows that very few points have been given a low probability of being bad wine. Most of the bad wine has $\pi \approx 0.5$ whereas good wine have been given a high probability ($\pi \approx 0.9$). This plot suggests that when making classifications from the model selecting a threshold value of $C = 0.7$ would be an appropriate choice to distinguish between the two classes. ROC curves can also be plotted. As with the previous examples, the precise models all have very similar curves and AUC values which the abstaining model 'beats'. (In this case $AUC[\mathcal{ILR}(W)] = [0.716, 0.826]$, $AUC[\mathcal{ILR}(W)(Abstain)] = 0.794$, $AUC[\mathcal{LR}_{ss}(W)] \approx AUC[\mathcal{LR}_\times(W)] \approx AUC[\mathcal{LR}(W^\dagger)] = 0.774$)




(a) Scatter plots of probability vs outcome for $\mathcal{ILR}(W)$




(b) Receiver operating characteristic curve for the simple example with added uncertain classifications.

Figure 10: Plots to show the discriminatory performance of the logistic regression models for the white wine example.

## 5 Uncertainty in Both Features and Labels

The imprecise approach can be used when there is uncertainty about both the features and the labels. Such situations are present in numerous real-world datasets. An imprecise logistic regression model can be found in this scenario through a combination of the algorithms in Sections 1 and 4.

### 5.1 Example

Osler et al. (2010) use a logistic regression model to predict the probability of death for a patient after a burn injury. The model they use is based upon a subset of data from the American Burn Association's National Burn Database[2]. The dataset has a mix of discrete (gender, race, flame involved in injury, inhalation injury) and continuous variables (age, percentage burn surface area) that can be used to model the probability that a person dies (outcome 1) after suffering a burn injury. Osler et al. exclude some patients from the dataset before training their model. They remove patients if their age or 'presence of inhalation injury' was not recorded. Additionally, as patients older than 89 were assigned to a single age category in the original dataset, they gave them a random age between 90 and 100 years.

Osler et al. did not need to exclude these patients merely because of epistemic uncertainty about the values. The proposed approach can be used with the original data. For instance, patients for which the outcome was unknown could have been included within their analysis as described in Section 4. Similarly, patients whose inhalation injury or age was unknown could have been included with the method described in Section 3. Patients with unknown inhalation injury could have been included as the $[0, 1]$ interval. Patients whose age was completely unknown could have been replaced by an interval between the minimum and maximum age, whereas if there was uncertainty because they were over 90 years old, then they could be intervalised as $[90, 100]$.

Other interval uncertainties may be present within the dataset. It is unlikely to be the case that all the people used within the study fit neatly into the discrete variables given. For instance the variable race is valued at 0 for "non-whites" and 1 for "whites". However, it goes without saying that the diversity of humanity does

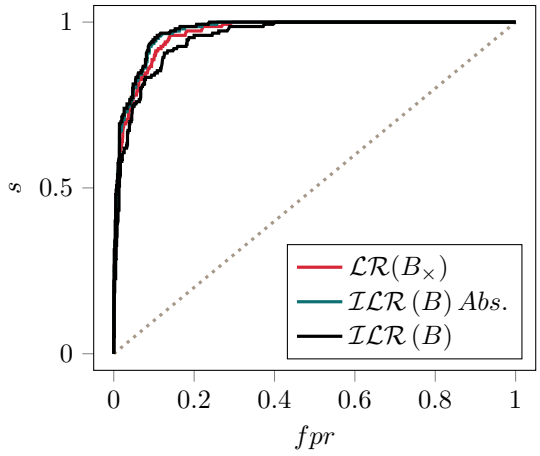
[2] `http://ameriburn.org/research/burn-dataset/`

not simply fall into such overly simplified categories; there are likely to be many people who could not be given a value of 0 or 1 and should instead have a $[0, 1]$ value. The same is true for gender. Not everyone can be defined as male or female. Also, there is almost certainly some measurement uncertainty associated with calculating the burn surface area that may also be best expressed as intervals. For simplicity, these uncertainties have not been addressed below.

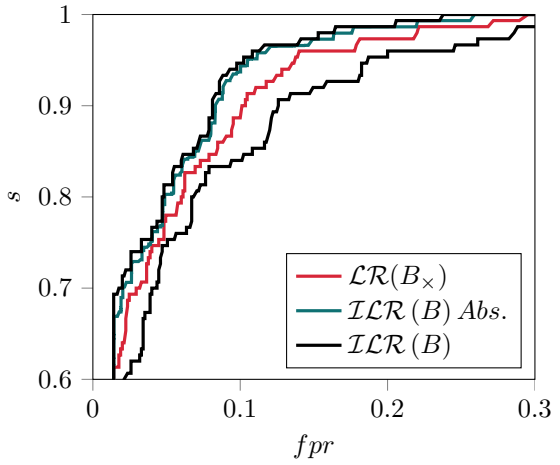
For this analysis, the subsample of the dataset used by Osler et al. made available by Hosmer Jr et al. (2013, p. 27) has been used. This version of the dataset includes 1000 patients from the 40,000 within the entire study and has a much higher prevalence of death than the original dataset. Because access to the original data is prohibitively expensive, the values in this dataset have been re-intervalised to replicate some of the removed uncertainty to create a hypothetical dataset, $B$, for this exposition. As there are no individuals older than 90 within the dataset, that particular re-intervalisation has not been possible, so all patients older than 80 have had their ages intervalised as [80,90]. Similarly, for 20 patients, the censored inhalation injury has been restored to dunno interval. Ten patients, who had been dropped because their outcome status was unknown, have been restored with status represented as [0,1].

There are two possible routes for an analyst to proceed when faced with such a dataset. They could follow the original methodology of Osler et al. and randomly assign patients with interval ages a precise value and then discard all other patients for which there is some uncertainty. Alternatively, the analyst could include the uncertainty within the model by creating an imprecise logistic regression model. As there is uncertainty within both the features and the labels, the model can be estimated by finding the values within the intervals that correspond to the minimum and maximum for $\beta_0$, $\beta_1$, etc. $\mathcal{ILR}(B)$ is the imprecise logistic regression fitted from this burn data. For comparison, $\mathcal{LR}(B_\times)$ has also been fitted based on removing the uncertainty in $B$ using the same methodology as Osler.

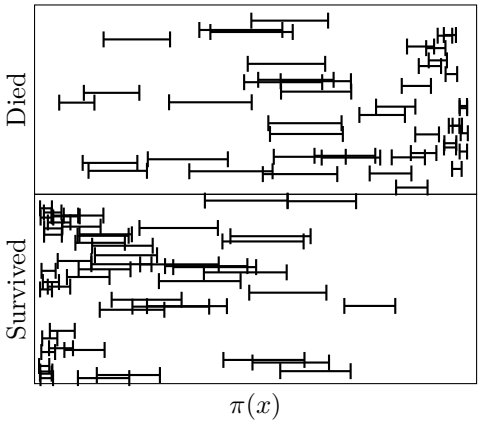
Regarding the performance of the two models, we can again turn to visualisations, as shown in Figure 11. Firstly, looking at Figure 11c we can see that the vast majority of patients who were given a low probability of death ($\pi$) did indeed survive, and patients who were given a high probability of death did sadly die. The ROC plots are shown in Figure 11a, Figure 11b shows the upper right corner of the plot in more detail. $\mathcal{ILR}(B)$ has a $AUC = [0.955, 0.974]$, the no prediction model has $AUC = 0.972$ and $\mathcal{LR}(B)$ has $AUC = 0.966$.




(a) Receiver operating characteristic curves for the burn example.




(b) Receiver operating characteristic curves for the burn example.




(c) Scatter plots of probability vs outcome for the model where uncertain values have been excluded (red) and the model including the uncertain values (blue). The two outcomes have been separated into different plots for clarity.

Figure 11: Plots to show the discriminatory performance of the various logistic regression models for the burn survivability example

It is pertinent to consider how a model is likely to be used and how uncertainty about the predicted probability of death impacts the classification. One method of dealing with this uncertainty that arises in Sections 4 and 3 is not making a prediction when the interval for $\pi$ straddles $C$. This method may not be appropriate in this example. What should happen when the model is unable to make a prediction should depend on what the result of deciding a patient has a high risk of death means clinically. If the model was being used to triage patients that need to go to a major trauma centre because the probability of death is considered high, then–out of an abundance of caution–one might prefer that if any part of the interval probability was greater than some threshold, the patient should be considered high risk. Equivalent to taking the probabilities from the upper bound of the range,

$$\text{high risk} = \begin{cases} 1, & \text{if } \overline{\pi} \geq C \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

However, if patients who are considered high risk then undergo some life-altering treatment that is perhaps only preferable to death, then under the foundational medical aphorism of "first do no harm", it may be preferable to consider a patient high-risk only if the whole interval is greater than the decision threshold, this is equivalent of taking the probabilities from the lower bound of the range,

$$\text{high risk} = \begin{cases} 1, & \text{if } \underline{\pi} \geq C \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Using the imprecise model in these scenarios would lead to better outcomes as the epistemic uncertainty would not be ignored. It is also the case that a patient that has a wide interval (as is the case for some in Figure 11c), implying that there is large epistemic uncertainty about the prediction, the medics would be aware of the uncertainty associated with the model and therefore may prefer to decide another way.

## 6 Conclusion

Analysts face uncertainties in the measurement values of their data. Often this uncertainty is either assumed away or just completely ignored. However, it may be better to compute with what we know rather than to make assumptions that may need to be revised later. Many uncertainties are naturally expressed as intervals arising from measurement error and missing or censored values. In the case of logistic regression, when faced with interval uncertainties, samples are often dropped from analyses—assuming that they are missing at random—or reduced down to a single value. This uncertainty should not simply be thrown away to make subsequent calculations easier.

Interval uncertainties can be included within a logistic regression model by considering the set of possible models as an imprecise structure. This even includes situations where there is uncertainty about the binary outcome status. The present analysis showed that, it is not reasonable to throw away data when the status is unknown if the reason the data has gone missing is dependent on the value or status of the missing samples. The case studies showed that previous methods used to handle interval uncertainties are ill-suited to situations where the narrow assumptions that they rely upon are unjustified or untenable. The methods based upon imprecise probabilities described in this paper work whenever there are interval uncertainties in the data regardless of how they happened to arise.

When using the new approach to classify, each new sample gets an interval probability of belonging to one of the binary classifications. Therefore, there are likely to be samples for which a definitive prediction cannot be made. If an analyst is happy to accept a *don't know* result, then the regressions performance can be improved for the samples for which a prediction is made. It may seem counterproductive or unhelpful for a model to return a don't know result. However, this can be desirable behaviour; saying "I don't know" is perfectly valid in situations where the uncertainty is large enough that a different decision could have been reached. Uncertainty in the output can allow for decisions made by algorithms to be more humane by requiring further interrogation to make a classification. Alternatively, depending on the use case, other ways of making decisions based on uncertain predictions could be made, such as being conservative or cost-minimising.

It should be noted that within many active learning systems, the machine learning algorithm is already forced to abstain from making a prediction for samples for which the probability is close to the decision boundary ($\pi(x) \approx C$), so that a human can provide a classification (Schein & Ungar, 2007; Chai et al., 2018). If the imprecise logistic regression model presented in this paper was used in such a system, it would have the advantage of clearly providing a criteria for which abstentions are prefered as opposed to a post-hoc decision based upon a arbitary definition of how close to the boundary is too close. Additionally, this method exposes samples for which the model returns wide interval probabilities even if their center is not close to the boundary and would have been considered a clear decision previously. If an abstention region is provided then any interval probability produced from the model that straddles the region would be considered indeterminate.

We have shown that it is possible to include interval uncertainty in both outcome status and predictor variables within logistic regression analysis by considering the set of possible models as an imprecise structure.

Such a method clearly can express the epistemic uncertainties within the dataset that are removed by traditional methods. The method described is guarentee to give rigorous bounds on the true probabilities, although it is computationally expensive. Future work in this area should be invested in finding improved algorithms to make them practical for larger-scale datasets.

## References


Massih-Reza Amini and Patrick Gallinari. Semi-Supervised Logistic Regression. In *15th European Conference on Artificial Intelligence*, pp. 5, Lyon, France, 2002.

Steven C Bagley, Halbert White, and Beatrice A Golomb. Logistic regression in the medical literature : Standards for use and reporting , with particular attention to one medical domain. *Journal of Clinical Epidemiology*, 54:979–985, 2001.

Patrice Bertrand. Descriptive Statistics for Symbolic Data. In Hans-Hermann Bock and Edwin Diday (eds.), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, pp. 106–124. Springer Berlin Heidelberg, 2000.

L Billard and E Diday. Regression Analysis for Interval-Valued Data. In Henk A. L. Kiers, Jean-Paul Rasson, Patrick J. F. Groenen, and Martin Schader (eds.), *Data Analysis, Classification, and Related Methods*, pp. 369–374. Springer Berlin Heidelberg, 2000.

HH Bock and E Diday. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Edited by H.-H. Bock and E. Diday.* Springer Berlin Heidelberg, Berlin, Germany, 2000.

Danilo Bzdok, Michael Eickenberg, Olivier Grisel, Bertrand Thirion, and Gael Varoquaux. Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 9, Montreal, Canada, 2015.

Hua Chai, Yong Liang, Sai Wang, and Hai-wei Shen. A novel logistic regression model combining semi-supervised learning and active learning for disease classification. *Scientific Reports*, 8(1):13009, December 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-31395-5.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA, ebook edition, 2006.

Shengqiang Chi, Xinhang Li, Yu Tian, Jun Li, Xiangxing Kong, Kefeng Ding, Chunhua Weng, and Jingsong Li. Semi-supervised learning to improve generalizability of risk prediction models. *Journal of Biomedical Informatics*, 92:103117, April 2019. ISSN 15320464. doi: 10.1016/j.jbi.2019.103117.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. doi: 10.1016/j.dss.2009.05.016.

Renata M C R de Souza, Francisco Jos, A Cysneiros, Diego C F Queiroz, and Roberta A De A Fagundes. A Multi-Class Logistic Regression Model for Interval Data. In *IEEE International Conference on Systems, Man and Cybernetics*, pp. 1253–1258, Singapore,Singapore, 2008. doi: 10.1109/ICSMC.2008.4811455.

Renata M C R de Souza, Diego C F Queiroz, and Francisco Jose. Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications*, 14(3):273–282, 2011. doi: 10.1007/s10044-011-0222-1.

Roberta A A Fagundes, Renata M C R De Souza, and Francisco Jose. Robust regression with application to symbolic interval data. *Engineering Applications of Artificial Intelligence*, 26(1):564–573, 2013. doi: 10.1016/j.engappai.2012.05.004.

Scott Ferson, Vladik Kreinovich, Janos Hajagos, William Oberkampf, and Lev Ginzburg. Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty. Technical report, Sandia National Laboratories, Albuquerque, NM, USA, 2007.

Federica Gioia, Carlo N Lauro, and Napoli Federico. Basic statistical methods for interval data. *Statistica Applicata*, 17:1–29, 2005.

Nicholas Gray, Scott Ferson, Marco De Angelis, Ander Gray, and Francis Baumont de Oliveira. Probability bounds analysis for Python. *Software Impacts*, 12:100246, May 2022. ISSN 26659638. doi: 10.1016/j.simpa.2022.100246.

David W. Hosmer Jr, Stanley Lameshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Ltd, Hoboken, NJ, USA, 3rd edition, 2013.

Tommi S Jaakkola. A variational approach to Bayesian logistic regression models and their extensions. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, pp. 12, Fort Lauderdale, FL, USA, 1997. PMLR.

JM Keynes. *A Treatise on Probability*. Macmillian and Co., London, UK, 1921.

David G. Kleinbaum and Mitchel Klein. *Logistic Regression: A Self-Learning Text*. Springer, New York, NY USA, 3rd edition, 2010.

Kimberly Kracman. The effect of school-based arts instruction on attendance at museums and the performing arts. *Poetics*, 24:203–218, 1996.

Balaji Krishnapuram, David Williams, Ya Xue, Lawrence Carin, Mário Figueiredo, and Alexander J Hartemink. On Semi-Supervised Classification. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pp. 8, Vancouver/Whistler, Canada, 2004.

Winifred Lambert and Mark Wheeler. Objective Lightning Probability Forecasting for Kennedy Space Center and Cape Canaveral Air Force Station. Technical report, National Air and Space Administration, Hannover, MD, USA, 2005.

Yongjun Li, Lizheng Wang, and Feng Li. A data-driven prediction approach for sports team performance and its application to National Basketball Association R. *Omega*, 98:102123–102123, 2021. doi: 10.1016/j.omega.2019.102123.

CF Manski. *Partial Identification of Probability Distributions*. Springer, New York, NY USA, 2003. ISBN 0-387-00454-8.

Scott Menard. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications, Inc, Thousand Oaks, California, 2010. ISBN 978-1-4129-7483-7. doi: 10.4135/9781483348964.

In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1): 90–100, 2003. doi: 10.1016/S0022-2496(02)00028-7.

W D Neary, B P Heather, and J J Earnshaw. The Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM). *British Journal of Surgery*, 30(2):157–165, 2003. doi: 10.1002/bjs.4041.

Hung T. Nguyen, Vladik Kreinovich, Berlin Wu, and Gang Xiang. *Computing Statistics under Interval and Fuzzy Uncertainty*. Springer, Heidelberg, Germany, 2012. ISBN 978-3-642-22829-2.

Sean M. O'Brien and David B. Dunson. Bayesian Multivariate Logistic Regression. *Biometrics*, 60(3): 739–746, September 2004. ISSN 0006341X. doi: 10.1111/j.0006-341X.2004.00224.x.

Gregory C Ohlmacher and John C Davis. Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. *Engineering Geology*, 69:331–343, 2003. doi: 10.1016/S0013-7952(03)00069-3.

Turner Osler, Laurent G Glance, and David W Hosmer. Simplified Estimates of the Probability of Death After Burn Injuries : Extending and Updating the Baux Score. *Journal of Trauma Injury, Infection and Critical Care*, 68(3), 2010. doi: 10.1097/TA.0b013e3181c453b3.

Sanjay Kumar Palei and Samir Kumar Das. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines : An approach. *Safety Science*, 47(1):88–96, 2009. doi: 10.1016/j.ssci.2008.01.002.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

S James Press and Sandra Wilson. Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, 73(364):699–705, 1978.

Patrick Royston and Douglas G Altman. Visualizing and assessing discrimination in the logistic regression model. *Statistics in Medicine*, 29, 2010. doi: 10.1002/sim.3994.

Andrew I. Schein and Lyle H. Ungar. Active learning for logistic regression: An evaluation. *Machine Learning*, 68(3):235–265, August 2007. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-007-5019-5.

Georg Schollmeyer. Computing Simple Bounds for Regression Estimates for Linear Regression with Interval-valued Covariates. In *Proceedings of the Twelth International Symposium on Imprecise Probabilities: Theories and Applications*, pp. 7, Virtual Conference, 2021. Proceedings of Machine Learning Research.

Lev V Utkin and Frank P A Coolen. Interval-valued regression and classification models in the framework of machine learning. In *7th International Symposium on Imprecise Probability: Theories and Applications*, pp. 11, Innsbruck, Austria, 2011.

Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, UK, 1991.

Andrea Wiencierz. *Regression analysis with imprecise data*. PhD thesis, Ludwig-Maximilians-Universitat Munchen, 2013.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics* -, pp. 189–196, Cambridge, Massachusetts, 1995. Association for Computational Linguistics. doi: 10.3115/981658.981684.