

# Binary Credal Classification under Sparsity Constraints <sup>★</sup>

Tathagata Basu<sup>1[0000-0002-6851-154X]</sup>, Matthias  
C. M. Troffaes<sup>1[0000-0002-1294-600X]</sup>, and Jochen Einbeck<sup>1[0000-0002-9457-2020]</sup>

Department of Mathematical Sciences, Durham University, UK

**Abstract.** Binary classification is a well known problem in statistics. Besides classical methods, several techniques such as the naive credal classifier (for categorical data) and imprecise logistic regression (for continuous data) have been proposed to handle sparse data. However, a convincing approach to the classification problem in high dimensional problems (i.e., when the number of attributes is larger than the number of observations) is yet to be explored in the context of imprecise probability. In this article, we propose a sensitivity analysis based on penalised logistic regression scheme that works as binary classifier for high dimensional cases. We use an approach based on a set of likelihood functions (i.e. an imprecise likelihood, if you like), that assigns a set of weights to the attributes, to ensure a robust selection of the important attributes, whilst training the model at the same time, all in one fell swoop. We do a sensitivity analysis on the weights of the penalty term resulting in a set of sparse constraints which helps to identify imprecision in the dataset.

**Keywords:** Classification · High dimensional data · Imprecise probability.

## 1 Introduction

Classification is a method for assigning a new object to a class or a group based on the observed features or attributes of the object. Classification is used in many applications such as pattern recognition for hand writing, disease treatment, facial recognition, chemical analysis, and so on. In general, a classifier can be seen as a function that maps a set of continuous or discrete variables into a categorical class variable. Constructing a classifier from random samples is an important problem in statistical inference. In our work, we will restrict ourselves to the case where there are only two classes to choose from, i.e. ‘binary classification’.

Let  $C$  be a random variable that takes values in  $\{0, 1\}$ . Let  $a$  be a  $p$ -dimensional vector that denotes the attributes of an object and let  $b = (b_1,$

---

<sup>★</sup> This work is funded by the European Commission’s H2020 programme, through the UTOPIAE Marie Curie Innovative Training Network, H2020-MSCA-ITN-2016, Grant Agreement number 722734.

$b_2, \dots, b_p)^T$  denote the vector of regression coefficients. In a regression setting, we construct a classifier through a generalised linear model (GLM) as follows:

$$E(C | a) = h(a^T b) \quad (1)$$

where  $h$  acts as a ‘link’ function and  $E$  stands for expectation. We define

$$\pi(a) := E(C | a) = P(C = 1 | a). \quad (2)$$

Logistic regression is a well-used special case of the GLM, which is suitable for classification with continuous attributes. Note that, for logistic regression,  $C$  follows a Bernoulli distribution. However, in the high dimensional case i.e. when the number of attributes is more than the number of observations ( $p > n$ ), the performance of logistic regression is often not satisfactory. Apart from overfitting, numerical optimisation methods often converge to local solutions because of multi-collinearity. Several techniques have been proposed to deal with this. Generally, a penalty term is introduced in the negative log-likelihood, leading to penalised logistic regression. A lasso-type penalty [16] is very popular because of its variable selection property [15, 21]. However, the lasso-type penalty can be inconsistent. To tackle this, Zou [22] introduced an adaptive version of the lasso for penalised logistic regression, which satisfies suitable asymptotic properties [8] for variable selection, and leads to consistent estimates.

Several works related to classification can be found in the imprecise probability literature. Zaffalon [19] introduced the idea of the naive credal classifier related to the imprecise Dirichlet model [18]. Bickis [3] introduced an imprecise logit-normal model for logistic regression. Corani and de Campos [5] proposed the tree augmented naive classifier based on imprecise Dirichlet model. Paton *et al.* [13, 14] used a near vacuous set of priors for multinomial logistic regression. Coz *et al.* [7] and Corani and Antonucci [4] investigated rejection based classifiers for attribute selection. However, high dimensional problems with automatic attribute selection are yet to be tackled in the context of imprecise probability.

In this study, we propose a novel imprecise likelihood based approach for high dimensional logistic regression problems. We use a set of sparsity constraints through weights in the penalty term. Working with a *set* of weights relaxes the assumption of preassigned weights and also helps to identify the behaviour of the attributes, whereas sparsity constraints help in variable selection which is essential for working with high dimensional problems. We use cross-validation for model validation using different performance measures [6].

The paper is organised as follows. We first discuss some properties of penalised logistic regression in Section 2. We discuss our sensitivity based classifier in Section 3. We discuss the model validation in Section 4, and we illustrate our results using two datasets in Section 5. We conclude in Section 6.

Throughout the paper, capital letters denote random variables or estimators that are dependent on any random quantity, and bold letters denote matrices.

## 2 Logistic Regressions for sparse problems

High dimensional regression is considered as sparse problem because of the small number of non-zero regression parameters. We often look for regularisation methods to achieve this sparsity/attribute selection. In this section, we discuss different penalised logistic regression schemes which are useful to attain a sparse model.

### 2.1 Penalised Logistic Regression (PLR)

Consider the generalised model in Eq. (1). For logistic regression, we use the following link function:

$$h(x) := \frac{\exp(x)}{1 + \exp(x)}. \quad (3)$$

We define a vector  $C := (C_1, C_2, \dots, C_n)^T$  denoting  $n$  observed classes such that,  $C_i \in \{0, 1\}$ . The  $C_i$  are thus Bernoulli random variables. Let  $\mathbf{a} := [a_1, a_2, \dots, a_n]$ , with  $a_i \in \mathbb{R}^p$ , denote the observed attributes for  $n$  objects, so that  $\mathbf{a}^T$  corresponds to the design matrix in the terminology of classical statistical modelling. It is easy to see that the negative log likelihood of the data is:

$$-\log(L(C, \mathbf{a}; b)) = \sum_{i=1}^n \left( -C_i (a_i^T b) + \log(1 + \exp(a_i^T b)) \right). \quad (4)$$

Therefore, the maximum likelihood estimate of the unknown parameter  $b$  is:

$$\hat{B}_{\text{lr}} := \arg \min_b \{-\log(L(C, \mathbf{a}; b))\}. \quad (5)$$

Here, we denote estimates, such as  $\hat{B}_{\text{lr}}$ , with capital letters because they are random variables (as they depend on  $C$ , which is random). The matrix of observed attributes  $\mathbf{a}$  is denoted with a lower case letter, as it is customary to consider it as fixed and thereby non-random.

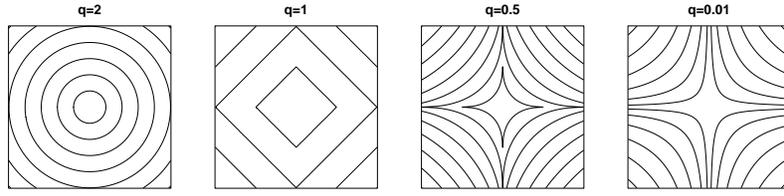
In high dimensional problems, we often seek for regularisation methods to avoid overfitting. We use penalised logistic regression (PLR) [15, 21] as a regularisation method which is defined by:

$$\hat{B}_{\text{plr}}(\lambda) := \arg \min_b \{-\log(L(C, \mathbf{a}; b)) + \lambda P(b)\}, \quad (6)$$

where  $P(b)$  is a penalty function. We get sparse estimate for  $b$  when:

$$P(b) := \sum_{j=1}^p |b_j|^q \quad (7)$$

with  $0 \leq q \leq 1$ . However, for  $q < 1$ , the problem is non-convex (see Fig. 1) and the optimisation is computationally expensive. In contrast, for  $q = 1$ , the penalty is a lasso-type penalty [16], which is convex and easy to solve numerically. The value of  $\lambda$  is chosen through cross-validation, where  $\lambda$  acts as a tuning parameter. In Fig. 1, we show contours of different  $\ell_q$  penalties for two variables..



**Fig. 1.** Contour plots of (7) for different  $\ell_q$  penalties.

## 2.2 Adaptive Penalised Logistic Regression (APLR)

The lasso type penalty in PLR can be inconsistent in variable selection and it is also not asymptotically unbiased. This issue can be resolved by assigning carefully chosen weights in the penalty term. This approach is known to be adaptive penalised logistic regression (APLR) [22, 2].

Let  $\hat{B} := (\hat{B}_1, \hat{B}_2, \dots, \hat{B}_p)$  be any root- $n$  consistent estimate for our logistic regression problem. Then, for any fixed  $\gamma > 0$ , the APLR [22] estimates are given by:

$$\hat{B}_{\text{aplr}}(\lambda, \gamma) := \arg \min_b \left( -\log(L(C, \mathbf{a}; b)) + \lambda \sum_{j=1}^p W_j(\gamma) |b_j| \right) \quad (8)$$

where

$$W_j(\gamma) := \frac{1}{|\hat{B}_j|^\gamma}. \quad (9)$$

Note that, for  $\gamma = 0$ , Eq. (8) becomes the regular penalised logistic regression with lasso penalty. Zou [22] showed that with these weights along with some mild regularity conditions, APLR follows desirable asymptotic properties for high dimensional problems [8].

**Computation** For  $\gamma > 0$ , the objective function of APLR is given by:

$$J(b) := \left( \sum_{i=1}^m [-C_i (a_i^T b) + \log(1 + \exp(a_i^T b))] + \lambda \sum_{j=1}^p W_j(\gamma) |b_j| \right), \quad (10)$$

where  $W_j(\gamma)$  is given by Eq. (9). Now, for optimality Eq. (10) must satisfy Karush-Kuhn-Tucker condition. Therefore, we have,

$$0 \in \sum_{i=1}^m \left[ -a_{ji} C_i + a_{ji} \frac{\exp(a_i^T b)}{1 + \exp(a_i^T b)} \right] + \lambda W_j(\gamma) \partial(|b_j|), \quad (11)$$

where,  $\partial|b_j|$  is defined [12] as

$$\partial(|b_j|) = \text{sign}(b_j), \quad (12)$$

with

$$\text{sign}(b_j) := \begin{cases} \{-1\} & \text{if } b_j < 0 \\ [-1, 1] & \text{if } b_j = 0 \\ \{1\} & \text{if } b_j > 0, \end{cases} \quad (13)$$

for  $j = 1, 2, \dots, p$ .

Let  $S := (S_1, S_2, \dots, S_p)$  be subject to the constraint  $S \in \text{sign}(\hat{B}_{\text{aplr}})$ . Then,  $\hat{B}_{\text{aplr}}$  satisfies the following:

$$\sum_{i=1}^m \left[ -a_{ji}C_i + a_{ji} \frac{\exp(a_i^T \hat{B}_{\text{aplr}}(\lambda, \gamma))}{1 + \exp(a_i^T \hat{B}_{\text{aplr}}(\lambda, \gamma))} \right] = -\lambda W_j(\gamma) S_j \quad (14)$$

$$\sum_{i=1}^m a_{ji} \left[ C_i - \frac{\exp(a_i^T \hat{B}_{\text{aplr}}(\lambda, \gamma))}{1 + \exp(a_i^T \hat{B}_{\text{aplr}}(\lambda, \gamma))} \right] = \lambda W_j(\gamma) S_j. \quad (15)$$

Now, let  $h(\mathbf{a}^T \hat{B}) := \left( h(a_1^T \hat{B}), h(a_2^T \hat{B}), \dots, h(a_n^T \hat{B}) \right)^T$ , where  $h$  is the link function defined in Eq. (3). Then, we can write Eq. (15) as,

$$\mathbf{a} \left[ C - h(\mathbf{a}^T \hat{B}_{\text{aplr}}(\lambda, \gamma)) \right] = \lambda W(\gamma) \cdot S \quad (16)$$

where ‘ $\cdot$ ’ denotes component wise multiplication. Note that Eq. (16) is not analytically solvable for  $\hat{B}_{\text{aplr}}$ . However, any sub-gradient based numerical optimisation method can be applied to solve it. Once we have the estimate, we can then define, for any new object with known attributes  $a_* \in \mathbb{R}^p$  and unknown class  $C_*$ ,

$$\hat{\Pi}(a_*, \lambda, \gamma) := P \left( C_* = 1 \mid a_*; \hat{B}_{\text{aplr}}(\lambda, \gamma) \right) = h \left( a_*^T \hat{B}_{\text{aplr}}(\lambda, \gamma) \right). \quad (17)$$

We can then for instance classify the object as 0 if  $\hat{\Pi}(a_*, \lambda, \gamma) < 1/2$ , as 1 if  $\hat{\Pi}(a_*, \lambda, \gamma) > 1/2$ , and as either if  $\hat{\Pi}(a_*, \lambda, \gamma) = 1/2$ . The parameter  $\gamma$  is often simply fixed (usually taken to be equal to 1), and  $\lambda$  is chosen through cross-validation, as with PLR.

**Properties** For a sequence of  $n$  observations, where  $a_i$  is the attribute vector for the  $i$ -th observation, we now denote:

$$\mathbf{a}_n := \mathbf{a} = [a_1, \dots, a_n] \quad (18)$$

in order to make the dependence of this  $p \times n$  matrix on  $n$  explicit. Define by  $b^* := (b_1^*, \dots, b_p^*)$  the vector of true regression coefficients. Assume the true model to be sparse, then without loss of generality  $\mathcal{S} := \{j: b_j^* \neq 0\} = \{1, 2, \dots, p_0\}$ , where  $p_0 < p$ . Let  $\phi(x) := \log(1 + \exp(x))$ , then for any observation  $a_i \in \mathbb{R}^p$  ( $1 \leq i \leq n$ ), we define the Fisher information matrix by:

$$I(b) := \phi''(a_i^T b) a_i a_i^T = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \quad (19)$$

where,  $I_{11}$  is a  $p_0 \times p_0$  matrix.

*Regularity Conditions:* We define the following regularity conditions for asymptotic properties of APLR.

C.1 Let  $\lambda_n(\gamma)$  be a sequence such that, for  $\gamma > 0$

$$\lim_{n \rightarrow \infty} \frac{\lambda_n(\gamma)}{\sqrt{n}} = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \lambda_n(\gamma) \cdot n^{(\gamma-1)/2} = \infty. \quad (20)$$

For example, the above holds for  $\lambda_n(\gamma) = n^{1/2-\gamma/4}$ .

C.2 The Fisher information matrix is finite and positive definite.

C.3 Let there exist an open set  $\mathcal{B} \subseteq \mathbb{R}^p$ , such that  $b^* \in \mathcal{B}$ . Then for every  $b \in \mathcal{B}$  and observation  $a_i \in \mathbb{R}^p$  ( $1 \leq i \leq n$ ), there exists a function  $M$  so that

$$|\phi'''(a_i^T b)| \leq M(a_i) < \infty. \quad (21)$$

Let  $\mathcal{S}_n = \{j : \hat{B}_{\text{aplr}, j} \neq 0\}$ .

**Theorem 1.** *Under C.1-C.3, APLR estimates satisfy the following properties:*

*P.1 Consistency in variable selection, i.e.*

$$\lim_{n \rightarrow \infty} P(\mathcal{S}_n = \mathcal{S}) = 1 \quad (22)$$

*P.2 Asymptotic normality, i.e.*

$$\sqrt{n} \left( \hat{B}_{\text{aplr}, \mathcal{S}} - b_{\mathcal{S}}^* \right) \xrightarrow{d} N(0, I_{11}^{-1}) \quad (23)$$

Note, that here  $\hat{B}_{\text{aplr}, \mathcal{S}}$  is dependent on both  $\lambda_n(\gamma)$  and  $\gamma$  but we skip writing it for the sake of notation.

P.1 and P.2 are well known results for high dimensional problems and the proofs can be found in [22].

### 3 Imprecise Adaptive Penalised Logistic Regression

The use of data-driven weights in APLR makes APLR consistent in attribute selection, where the parameter  $\gamma$  is pre-assigned (usually equal to 1) or is estimated through cross-validation. However, high dimensional problems are sparse in nature, i.e. we have to deal with very limited information and therefore a single vector of weights is often proved to be sensitive and leads to misclassification, especially when the variability of the attributes is negligible with respect to each other. Sometimes, APLR may also perform poorly during model validation as, a single value of  $\gamma$  can provide two very different vectors of weights for two different parts of a single dataset. For instance, fixing  $\gamma = 1$ , essentially gives us the inverse of the absolute values of our estimates, which are generally sensitive to the data in sparse regime. So, we propose a sensitivity analysis of APLR over an interval of  $\gamma$  and obtain a non-determinate classifier. We call this method as imprecise adaptive penalised logistic regression or simply IAPLR. This allows

the weights to vary in the order of  $\gamma$  providing us a set of sparse constraints of the form  $\sum_{j=1}^p |b_j|/|\hat{B}_j|^\gamma$ . This set of weight vectors allows the model to be flexible but consistent as we only rely on the data-driven weights.

The sensitivity analysis gives us a set of APLR estimates as a function of  $\gamma$ . We use this set of APLR estimates to obtain a set of estimated probabilities which are used for the decision making.

### 3.1 Decision rule

Consider the APLR estimates defined by Eq. (8) and Eq. (9). As we described earlier, we perform a sensitivity analysis on the parameter  $\gamma$ . This gives us a set of estimated probabilities dependent on  $\gamma$ , such that  $\gamma \in [\underline{\gamma}, \bar{\gamma}]$ . We use the notion of credal dominance [19] for the decision criteria.

We can then for instance classify a new object with attributes  $a_* \in \mathbb{R}^p$  as  $\{0\}$  if  $\hat{\Pi}(a_*, \lambda, \gamma) < 1/2$  for all  $\gamma \in [\underline{\gamma}, \bar{\gamma}]$ , as  $\{1\}$  if  $\hat{\Pi}(a_*, \lambda, \gamma) \geq 1/2$  for all  $\gamma \in [\underline{\gamma}, \bar{\gamma}]$ , and as  $\{0, 1\}$  (i.e. indeterminate) otherwise. Note that our classifier now returns non-empty subsets of  $\{0, 1\}$  rather than elements of  $\{0, 1\}$ , to allow indeterminate classifications to be expressed.

### 3.2 Prediction Consistency

We define the following:

$$a_{*,S} := [a_{*,j}]_{j \in S}, \quad (24)$$

i.e.,  $a_{*,S}$  is a  $p_0$ -dimensional vector.

**Theorem 2.** *Let  $a_* \in \mathbb{R}^p$  such that  $a_{*,S}^T a_{*,S} > 0$ . Then for  $\gamma > 0$  and under C.1 -C.3, we have the following:*

$$\sqrt{n} \left( \hat{\Pi}(a_*, \lambda_n(\gamma), \gamma) - \pi(a_*) \right) \xrightarrow{d} N \left( 0, [\pi(a_*) (1 - \pi(a_*))]^2 a_{*,S}^T I_{11}^{-1} a_{*,S} \right) \quad (25)$$

where,  $I_{11}$  is the leading block matrix of the Fisher information matrix defined in Eq. (19).

*Proof.* We know that, under C.1-C.3 APLR estimates satisfies P.1. Therefore, as  $n \rightarrow \infty$ ,

$$a_*^T \hat{B}_{\text{aplr}} = a_{*,S_n}^T \hat{B}_{\text{aplr}, S_n} = a_{*,S}^T \hat{B}_{\text{aplr}, S}. \quad (26)$$

Then from Eq. (26), we have,

$$\hat{\Pi}(a_*, \lambda_n(\gamma), \gamma) = h \left( a_{*,S}^T \hat{B}_{\text{aplr}, S} \right). \quad (27)$$

Now, by P.2, we know that  $\hat{B}_{\text{aplr}, S}$  is root- $n$  consistent. Therefore,

$$\left( \hat{B}_{\text{aplr}, S} - b_S^* \right) = O_p(n^{-1/2}). \quad (28)$$

Following the approach of [1] for logistic regression problems, we apply Taylor's series expansion in Eq. (27) with respect to the true parameter  $b_S^*$ . Then we have,

$$\hat{\Pi}(a_*, \lambda_n(\gamma), \gamma) = h(a_{*,S}^T b_S^*) + \left(\hat{B}_{\text{aplr}, S} - b_S^*\right)^T \frac{\partial h(a_{*,S}^T b_S^*)}{\partial b_S^*} + o_p(n^{-1/2}) \quad (29)$$

$$= \pi(a_*) + \left(\hat{B}_{\text{aplr}, S} - b_S^*\right)^T \frac{\partial h(a_{*,S}^T b_S^*)}{\partial b_S^*} + o_p(n^{-1/2}). \quad (30)$$

Here,  $o_p(n^{-1/2})$  comes from the condition mentioned in Eq. (28). Now, rearranging the terms we get,

$$\hat{\Pi}(a_*, \lambda_n(\gamma), \gamma) - \pi(a_*) = \left(\hat{B}_{\text{aplr}, S} - b_S^*\right)^T \frac{\partial h(a_{*,S}^T b_S^*)}{\partial b_S^*} + o_p(n^{-1/2}). \quad (31)$$

Now, from P.2 we have,

$$\sqrt{n} \left(\hat{B}_{\text{aplr}, S} - b_S^*\right) \xrightarrow{d} N(0, I_{11}^{-1}). \quad (32)$$

Then, applying Eq. (32) in Eq. (30), we get

$$\sqrt{n} \left(\hat{\Pi}(a_*, \lambda_n(\gamma), \gamma) - \pi(a_*)\right) \xrightarrow{d} N\left(0, \left[\frac{\partial h(a_{*,S}^T b_S^*)}{\partial b_S^*}\right]^T I_{11}^{-1} \frac{\partial h(a_{*,S}^T b_S^*)}{\partial b_S^*}\right). \quad (33)$$

Now,

$$\frac{\partial h(a_{*,S}^T b_S^*)}{\partial b_S^*} = \left[ \frac{\exp(a_{*,S}^T b_S^*) (1 + \exp(a_{*,S}^T b_S^*)) - \exp(a_{*,S}^T b_S^*)^2}{(1 + \exp(a_{*,S}^T b_S^*))^2} \right] a_{*,S} \quad (34)$$

$$= \left[ \frac{\exp(a_{*,S}^T b_S^*)}{(1 + \exp(a_{*,S}^T b_S^*))^2} \right] a_{*,S} \quad (35)$$

$$= h(a_{*,S}^T b_S^*) \left[ 1 - \frac{\exp(a_{*,S}^T b_S^*)}{1 + \exp(a_{*,S}^T b_S^*)} \right] a_{*,S} \quad (36)$$

$$= h(a_{*,S}^T b_S^*) (1 - h(a_{*,S}^T b_S^*)) a_{*,S} \quad (37)$$

$$= h(a_{*,S}^T b_S^*) (1 - h(a_{*,S}^T b_S^*)) a_{*,S} \quad (38)$$

$$= \pi(a_*) (1 - \pi(a_*)) a_{*,S}. \quad (39)$$

Therefore, using Eq. (39) in Eq. (33), we have

$$\sqrt{n} \left(\hat{\Pi}(a_*, \lambda_n(\gamma), \gamma) - \pi(a_*)\right) \xrightarrow{d} N\left(0, [\pi(a_*) (1 - \pi(a_*))]^2 a_{*,S}^T I_{11}^{-1} a_{*,S}\right) \quad (40)$$

The result shows that for infinite data stream, the estimated probabilities will be equal to the true probability  $\pi(a_*)$ .

## 4 Model Validation

In our method, we perform a sensitivity analysis over  $\gamma$ . This gives us a set of estimated probabilities for each fixed value of  $\lambda$ . Depending on these values in this set, the predicted class will be either unique or both ‘0’ and ‘1’. Therefore, the classical measures of accuracy will not be applicable in this context. So we use the following performance measures, proposed by Corani and Zaffalon [6] for Naive Credal Classifier (NCC).

### 4.1 Measures of Accuracy

We use cross-validation for model validation where  $\lambda$  is used as a tuning parameter. We consider the following performance measures [6, 14] for credal classification.

**Definition 1 (Determinacy).** *Determinacy is the performance measure that counts the percentage of classifications with unique output.*

**Definition 2 (Single accuracy).** *Single accuracy is accuracy of the classifications when the output is determinate.*

There are two other performance measures called *indeterminate output size* and *set accuracy*. However, in the context of binary credal classification, indeterminate output size is always equal to 2 and set accuracy is always equal to 1.

The above mentioned performance measures will be used for model validation but for the model selection, we first need to choose an optimal  $\lambda$ , i.e. a value of  $\lambda$  that maximises the performance of our model. For this purpose, we need to use a trade-off between determinacy and single accuracy. We use  $u_{65}$  utility on the discounted accuracy, as proposed by Zaffalon *et al.* [20]. We display  $u_{65}$  on the discounted accuracy measure in Table 1, where each row stands for predicted class and each column stands for the actual class.

	{0}	{1}
{0}	1	0
{1}	0	1
{0, 1}	0.65	0.65

**Table 1.** Discounted utility ( $u_{65}$ ) table for binary credal classification

Note that, for binary credal classification, we can formulate this unified  $u_{65}$  accuracy measure in the following way:

$$\text{Accuracy} = \text{Determinacy} \times \text{Single accuracy} + 0.65 \times (1 - \text{Determinacy}) \quad (41)$$

## 4.2 Model Selection and validation

We use nested loop cross-validation for model selection and validation. We first split the dataset  $\mathcal{D}$  in 2 equal parts  $\mathcal{D}_1$  and  $\mathcal{D}_2$ . We take  $\mathcal{D}_1$  and split it in 5 equal parts. We use 4 of them to train our IAPLR model and use the remaining part for the selection of  $\lambda$ . We do this for each of the 5 parts to get an optimal  $\lambda$  based on the averaged performance measure. After obtaining the optimal  $\lambda$  through cross-validation, we validate our model with  $\mathcal{D}_2$  for model validation.

We repeat the same for  $\mathcal{D}_2$ , we use  $\mathcal{D}_2$  to obtain an optimal  $\lambda$  for model selection and then validate it using  $\mathcal{D}_1$ . By this way, we use each observation exactly once for testing. This also gives a comparison between these two models and gives us an idea of interactions between the observations.

## 5 Illustration

We use two different datasets for illustration. The Sonar dataset is a regular logistic regressional data while the LSVT dataset is high dimensional. In both cases, we normalise the attributes to avoid scaling issues and split the datasets in two equal parts  $\mathcal{D}_{S,1}$   $\mathcal{D}_{S,2}$  (Sonar) and  $\mathcal{D}_{L,1}$ ,  $\mathcal{D}_{L,2}$  (LSVT). We first select our model using  $\mathcal{D}_{S,1}$  ( $\mathcal{D}_{L,1}$ ). We vary our set of weights through 20 different  $\gamma$ 's ranging from 0.01 to 1. We take a grid of 50  $\lambda$  values. We find optimal  $\lambda$  by 5-fold cross validation. We use this optimal  $\lambda$  for model selection.

We compare our results with the naive credal classifier (NCC) [19]. For this, we first categorise the attributes in 5 factors. We train our model in a grid of the concentration parameter  $s$  with 50 entries ranging from 0.04 to 2. We run a 5-fold cross-validation the choice of optimal  $s$  and this value of  $s$  for model selection. We also compare our result with naive Bayes classifier (NBC) [11] and APLR [22, 2]. For APLR select the value of optimal  $\lambda$  through a 5-fold cross-validation. We use `glmnet` [9] for training APLR and IAPLR model. We validate our model using  $\mathcal{D}_{S,2}$  ( $\mathcal{D}_{L,2}$ ). We then select our model using  $\mathcal{D}_{S,2}$  ( $\mathcal{D}_{L,2}$ ) and validate using  $\mathcal{D}_{S,1}$  ( $\mathcal{D}_{L,1}$ ) to capture interaction between the observations.

We show a summary of our results in Table 2. The left most column denotes the training set. We show determinacy in the second column. In third and fourth column, we display the single accuracy and utility based ( $u_{65}$ ) accuracy, respectively and in the right most column we display range of active attributes.

### 5.1 Sonar Dataset

We use the Sonar dataset [10] for the illustration of our method. The dataset consists of 208 observations on 60 attributes in the range of 0 to 1. Sonar signals are reflected by either a metallic cylinder or a roughly cylindrical rock, and the attributes represent the energy of the reflected signal within a particular frequency band integrated over time. We use these attributes to classify the types of the reflectors.

In the top row of Fig. 2, we show the cross validation plots with respect to  $\lambda$ . For  $\mathcal{D}_{S,1}$ , the optimal  $\lambda$  is found to be 0.039 and for  $\mathcal{D}_{S,2}$  the value is equal

to 0.087. We observe from Table 2, that IAPLR outperforms the rests in terms of derermincay and  $u_{65}$  utility measure. It also has a good agreement in model validation with respect to the datasets unlike NCC or NBC which are sensitive with respect to the training dataset. It performs an automatic variable selection like APLR. We show the selected variables in the left most column. For IAPLR, we have a range of active attributes unlike APLR, which is computed using  $\gamma = 1$ . We observe that for  $\mathcal{D}_{S,1}$ , the sparsity of the model is more sensitive than the sparsity of the model trained by  $\mathcal{D}_{S,2}$ . In the top row of Fig. 3, show the sensitivity of sparsity with respect to  $\gamma$  for the optimal value of  $\lambda$  obtained through cross-validation. We observe that for both partitions the method selects more attributes as the value of  $\gamma$  increases.

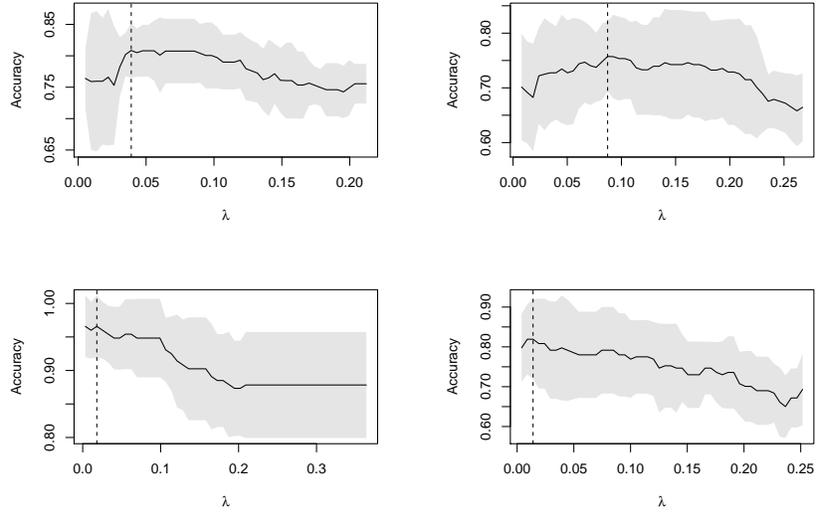
## 5.2 LSVT Dataset

We use the LSVT dataset [17] for the illustration with high dimensional data. The dataset consists of 126 observations on 310 attributes. The attributes are 310 different biomedical signal processing algorithms which are obtained through 126 voice recording signals of 14 different persons diagnosed with Parkinson’s disease. The responses denote acceptable (1) vs unacceptable (2) phonation during LSVT rehabilitation.

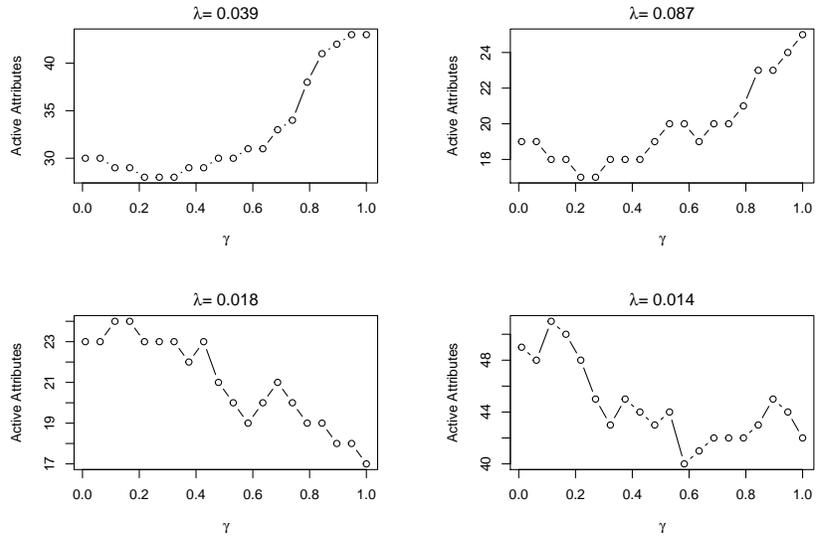
In the bottom row of Fig. 2, we show the cross validation plots with respect to  $\lambda$ . For  $\mathcal{D}_{L,1}$ , the optimal  $\lambda$  is found to be 0.018 and for  $\mathcal{D}_{L,2}$  the value is equal to 0.014. We observe from Table 2, that IAPLR outperforms the rests. It also has a good agreement in model validation with respect to the datasets unlike NCC, NBC and APLR. We notice that the sparsity levels are significantly different for different partitions of the dataset. We show the sparsity level in the bottom row of Fig. 3. We observe that for both partitions the method rejects more attributes as the value of  $\gamma$  increases.

## 6 Conclusion

In this article, we introduce a novel binary credal classifier for high dimensional problems. We exploit the notion of adaptive penalised logistic regression and use an imprecise likelihood based approach for the classifier. We illustrate our result using two different datasets One involving sonar signals bounced from two hard objects and the other involving LSVT rehabilitation of patients diagnosed with Parkinson’s disease. We compare our result with naive credal classifier, naive Bayes classifier and adaptive penalised logistic regression. We observe that our method is in good agreement with NCC in terms of single accuracy but outperforms NCC in terms of the determinacy and  $u_{65}$  utility measure. We notice that for both Sonar and LSVT dataset, NCC performs better than any other method for the second partition of the datasets. We observe that, IAPLR or APLR performs relatively better than the other methods as it does not rely on the factorisation. Our method also does an automatic attribute selection. We notice that the sensitivity of the attribute selection is almost monotone with respect to the parameter  $\gamma$ .



**Fig. 2.** Cross-validation curve with respect to the tuning parameter  $\lambda$ . The top row represents the results obtained for  $\mathcal{D}_{S,1}$  (left),  $\mathcal{D}_{S,2}$  (right) and the bottom row represents that of  $\mathcal{D}_{L,1}$  (left),  $\mathcal{D}_{L,2}$  (right).



**Fig. 3.** Sensitivity of sparsity with respect to  $\gamma$ . The top row represents the results obtained for  $\mathcal{D}_{S,1}$  (left),  $\mathcal{D}_{S,2}$  (right) and the bottom row represents that of  $\mathcal{D}_{L,1}$  (left),  $\mathcal{D}_{L,2}$  (right).

Method	Training	Deter.	Single Acc.	$u_{65}$	Active
IAPLR ( $\lambda = 0.039$ )	$\mathcal{D}_{S,1}$	0.87	0.73	0.72	28–43
IAPLR ( $\lambda = 0.087$ )	$\mathcal{D}_{S,2}$	0.87	0.77	0.75	17–25
NCC ( $s = 0.02$ )	$\mathcal{D}_{S,1}$	0.77	0.68	0.67	–
NCC ( $s = 0.56$ )	$\mathcal{D}_{S,2}$	0.49	0.78	0.72	–
NBC	$\mathcal{D}_{S,1}$	–	–	0.59	–
NBC	$\mathcal{D}_{S,2}$	–	–	0.74	–
APLR ( $\lambda = 0.104$ )	$\mathcal{D}_{S,1}$	–	–	0.71	12
APLR ( $\lambda = 0.189$ )	$\mathcal{D}_{S,2}$	–	–	0.72	9
<hr/>					
IAPLR ( $\lambda = 0.018$ )	$\mathcal{D}_{L,1}$	0.98	0.82	0.82	17–24
IAPLR ( $\lambda = 0.014$ )	$\mathcal{D}_{L,2}$	0.83	0.85	0.81	40–51
NCC ( $s = 0.08$ )	$\mathcal{D}_{L,1}$	0.14	0.78	0.67	–
NCC ( $s = 0.04$ )	$\mathcal{D}_{L,2}$	0.25	0.88	0.71	–
NBC	$\mathcal{D}_{L,1}$	–	–	0.51	–
NBC	$\mathcal{D}_{L,2}$	–	–	0.40	–
APLR ( $\lambda = 0.052$ )	$\mathcal{D}_{L,1}$	–	–	0.81	11
APLR ( $\lambda = 0.285$ )	$\mathcal{D}_{L,2}$	–	–	0.76	11

**Table 2.** Summary of model selection and validation

## References

1. Agresti, A.: Categorical Data Analysis. Wiley Series in Probability and Statistics, Wiley (2013), <https://books.google.co.uk/books?id=UOrr47-2oisC>
2. Algamal, Z.Y., Lee, M.H.: Penalized logistic regression with the adaptive lasso for gene selection in high-dimensional cancer classification. *Expert Systems with Applications* **42**(23), 9326 – 9332 (2015). <https://doi.org/https://doi.org/10.1016/j.eswa.2015.08.016>
3. Bickis, M.: The imprecise logit-normal model and its application to estimating hazard functions. *Journal of Statistical Theory and Practice* **3**(1), 183–195 (Mar 2009). <https://doi.org/10.1080/15598608.2009.10411919>
4. Corani, G., Antonucci, A.: Credal ensembles of classifiers. *Computational Statistics & Data Analysis* **71**, 818 – 831 (2014). <https://doi.org/https://doi.org/10.1016/j.csda.2012.11.010>
5. Corani, G., de Campos, C.P.: A tree augmented classifier based on extreme imprecise dirichlet model. *International Journal of Approximate Reasoning* **51**(9), 1053 – 1068 (2010). <https://doi.org/https://doi.org/10.1016/j.ijar.2010.08.007>
6. Corani, G., Zaffalon, M.: Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *J. Mach. Learn. Res.* **9**, 581–621 (2008)
7. José del Coz, J., Díez, J., Bahamonde, A.: Learning nondeterministic classifiers. *J. Mach. Learn. Res.* **10**, 2273–2293 (Dec 2009)
8. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456), 1348–1360 (2001), <http://www.jstor.org/stable/3085904>
9. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1), 1–22 (2010), <http://www.jstatsoft.org/v33/i01/>

10. Gorman, R.P., Sejnowski, T.J.: Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks* **1**(1), 75–89 (1988). [https://doi.org/10.1016/0893-6080\(88\)90023-8](https://doi.org/10.1016/0893-6080(88)90023-8)
11. Maron, M.E.: Automatic indexing: An experimental inquiry. *J. ACM* **8**(3), 404–417 (Jul 1961), <https://doi.org/10.1145/321075.321084>
12. Nesterov, Y.: *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1st edn. (2014)
13. Paton, L., Troffaes, M.C.M., Boatman, N., Hussein, M., Hart, A.: Multinomial logistic regression on markov chains for crop rotation modelling. In: Laurent, A., Strauss, O., Bouchon-Meunier, B., Yager, R.R. (eds.) *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. pp. 476–485. Springer International Publishing, Cham (2014). [https://doi.org/10.1007/978-3-319-08852-5\\_49](https://doi.org/10.1007/978-3-319-08852-5_49)
14. Paton, L., Troffaes, M.C.M., Boatman, N., Hussein, M., Hart, A.: A robust bayesian analysis of the impact of policy decisions on crop rotations. In: Augustin, T., Doria, S., Miranda, E., Quaeghebeur, E. (eds.) *ISIPTA '15: proceedings of the 9th International Symposium on Imprecise Probability : Theories and Applications*, 20-24 July 2015, Pescara, Italy, pp. 217–226. SIPTA (July 2015), <http://dro.dur.ac.uk/15736/>
15. Shevade, S., Keerthi, S.: A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* **19**(17), 2246–2253 (2003). <https://doi.org/10.1093/bioinformatics/btg308>
16. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1), 267–288 (1996), <http://www.jstor.org/stable/2346178>
17. Tsanas, A., Little, M.A., Fox, C., Ramig, L.O.: Objective automatic assessment of rehabilitative speech treatment in parkinson’s disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **22**(1), 181–190 (2014)
18. Walley, P.: *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis (1991), <https://books.google.co.uk/books?id=-hbvAAAAMAAJ>
19. Zaffalon, M.: The naive credal classifier. *Journal of Statistical Planning and Inference* **105**(1), 5 – 21 (2002). [https://doi.org/https://doi.org/10.1016/S0378-3758\(01\)00201-4](https://doi.org/https://doi.org/10.1016/S0378-3758(01)00201-4), *imprecise Probability Models and their Applications*
20. Zaffalon, M., Corani, G., Mauá, D.: Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning* **53**(8), 1282 – 1301 (2012). <https://doi.org/https://doi.org/10.1016/j.ijar.2012.06.022>, *imprecise Probability: Theories and Applications (ISIPTA'11)*
21. Zhu, J., Hastie, T.: Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5**(3), 427–443 (2004). <https://doi.org/10.1093/biostatistics/kxg046>
22. Zou, H.: The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**(476), 1418–1429 (2006). <https://doi.org/10.1198/016214506000000735>