Credal Classification of Automobile Risk

Ben Willis

January 29, 2017

Abstract

This project investigates how the credal classifier can be applied to the problem of determining auto mobile insurance risk. The goal is to show how the credal classifier can model uncertainty better than other classifiers and hence return more accurate classifications. This uncertainty may arise from small sample sets or missing data which are normally difficult to deal with.

Contents

Chapter 1

Introduction

1.1 Classification

Classification is the problem of identifying which class an object belongs to. Each object can be distinguished by a set of properties know as features and each object belongs to a single class. A classifier is an algorithm which, given previous observations and their classes, can determine which class a new observation belongs to [5]. There are many applications of classifiers including image recognition, sentiment analysis and medical diagnosis.

Classifier can be split into two categories, supervised and unsupervised. Unsupervised classifiers infer classes from the data. Supervised classifiers are constructed from a set of data for which the true classes are known and this is the type of classifier we will be exploring [3].

1.2 Auto mobile Insurance

Classifiers have many applications in the finance industry ranging from financial trading [1] to credit card fraud detection [2]. We will study the problem of classifying the risk to an insurer of a car and comparing this solution to the classification of an expert. We will then examine how both classifications compare to the normalised loss to the insurer.

The data set we will be analysing contains vehicular information about 205 auto mobiles. This features includes dimensions, engine specifications and vehicle characteristics. It also contains an experts assessed risk to the insurer of the vehicle on an integer scale of -2 to 3 with 3 being most risky and -2 being least risky. In addition to the technical information and the experts assessment the data set also contains the normalized loss to the insurer. This ranges from 65 to 256 and is normalized for all vehicles within a particular size classification (two-door small, station wagons, etc) and represents the

average loss per car per year [6].

Chapter 2

Naive Bayes Classifier

2.1 History

2.2 Theory

Formally, let us denote the class variable by C, taking values in the set $\mathcal{C}$. Also we measure k features $A_1, \ldots, A_k$ from the sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$. We denote observations of these variables as c and $a_1, \ldots, a_k$ respectively.

We are interested in $P(c \mid \mathbf{a})$. Using Bayes theorem we can rewrite this as:

$$P(c \mid \mathbf{a}) = \frac{P(\mathbf{a} \mid c)P(c)}{P(\mathbf{a})} \tag{2.1}$$

Moreover we can make use of the naivety assumption. The naivety assumptions states that each attribute is independent of one another. We can now write the probability of an object being in class c with attributes $a_1, \ldots, a_k$ as:

$$P(c \mid \mathbf{a}) = \frac{P(c)\prod_{i=1}^{k} P(a_i \mid c)}{P(\mathbf{a})} \tag{2.2}$$

To turn this into a classifier we need a way to make a decision for which class an object falls into based on the estimated probabilities. A common method is choosing the class that maximises $P(c \mid \mathbf{a})$. This is known as the maximum a posteriori (MAP) estimate. We also note that $P(\mathbf{a})$ is not dependent on C hence we can write our MAP estimate as:

$$c_{MAP} = argmax_{c \in \mathcal{C}} P(c) \prod_{i=1}^{k} P(a_i \mid c) \tag{2.3}$$

Now that we have our method for making our decision we need to estimate the required probabilities.

Firstly we parametrise these probabilities.We denote the unknown chances of observing an object with $C = c$ by θ_c and the change of observing an object with $C = c$ and $\mathbf{A} = \mathbf{a}$ by $\theta_{\mathbf{a},c}$. Similarly we denote the conditional chances of $A_i = a_i$ and $(A_1, \ldots, A_k) = (a_1, \ldots, a_k)$ given $C = c$ by $\theta_{a_i|c}$ and $\theta_{\mathbf{a}|c}$ respectively.

We can consider the likelihood function for the θ, the vector whose elements are the chances $\theta_{\mathbf{a},c}$ given data $\mathbf{n}$, the vector of all known frequencies.

The likelihood function can be expressed as:

$$l(\theta \mid \mathbf{n}) \propto \prod_{c \in \mathcal{C}} \left[\theta_c^{n(c)} \prod_{i=1}^{k} \prod_{a_i \in \mathcal{A}_i} \theta_{a_i|c}^{n(a_i,c)} \right] \tag{2.4}$$

A simple estimate for these parameters is the maximum likelihood estimate (MLE). To find the MLE first we take the log likelihood:

$$L(\theta \mid \mathbf{n}) \propto \sum_{c \in \mathcal{C}} n(c) log(\theta_c) + \sum_{c \in \mathcal{C}} \sum_{i=1}^{k} \sum_{a_i \in \mathcal{A}_i} n(a_i, c) log(\theta_{a_i|c}) \tag{2.5}$$

So to maximise the likelihood function we need to maximise the two parts of the log likelihood function.

To do so we use the method of Lagrange multipliers. This is a strategy for finding local maxima and minima of a function subject to constraints.

For the first equation we have

$$f(\theta, \mathbf{n}) = \sum_{c \in \mathcal{C}} n(c) log(\theta_c) \tag{2.6}$$

$$g(\theta, \mathbf{n}) = \sum_{c \in \mathcal{C}} \theta_c - 1 \tag{2.7}$$

This gives us our Lagrangian:

$$\mathcal{L}(\theta, \mathbf{n}, \lambda) = \sum_{c \in \mathcal{C}} n(c) log(\theta_c) - \lambda(\sum_{c \in \mathcal{C}} \theta_c - 1) \tag{2.8}$$

Differentiating with respect to θ_c we have:

$$\nabla_{\theta_c} \mathcal{L}(\theta, \mathbf{n}, \lambda) = \frac{n(c)}{\theta_c} - \lambda \tag{2.9}$$

Hence the maximum is achieved giving an mle of $\hat{\theta_c} = \frac{n(c)}{N}$. Intuitively this is just the relative frequency of observations that fall into that class.

We now have our naive Bayes classifier. We estimate $P(c)$ by $\frac{n(c)}{N}$ and $P(a_i \mid c)$ by $\frac{n(a_i,c)}{n(c)}$, the relative frequencies. Then we choose the class c which maximises $P(c) \prod_{i=1}^{k} P(a_i \mid c)$.

2.3 Applications

2.3.1 Forest Type Data Set

First we will apply this classifier to a data set from a remote sensing study. The study measured spectral information in the green, red and infrared wavelengths on three separate dates of different forest types in Japan. In total we have nine continuous attributes and four possible classes: Sugi forest, Hinoki forest, Mixed deciduous forest and other non-forest land.

To make our data appropriate for this method we discrete the continuous variables into n bins with an equal frequency.

2.4 Diagnostics

Applying this to our data set and using a technique known as k-fold cross validation to evaluate accuracy. In k-fold cross validation we split our dataset into k equally sized groups. Then for each group we train the classifier on all the other groups and test it on that group. We then average all these accuracy to return an (unbiased?) estimate for the accuracy of our classifier.

The choice of k leads to different types of cross validation. A standard choice is $k = 10$. A special case of cross validation is when $k = n$ (the number of observations). This is knowns as *Leave-one-out cross validation* [4].

Chapter 3

Corrected NBC with Dirichlet Prior

We can use the Dirichlet distribution as a conjugate prior to our likelihood function.

The Dirichlet distribution is the multinomial extension on the gamma distribution for $x_1, \ldots, x_k$ where $x_i \in (0, 1)$ and $\sum_{i=1}^{k} x_i = 1$ with probability density function:

$$f(x_1, \ldots, x_k \mid \alpha_1, \ldots, \alpha_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \tag{3.1}$$

where Γ is the gamma function and $\alpha_i > 0$.

We can rewrite the prior density of our Dirichlet distribution in a similar manner to our likelihood function. By setting $x_i = \theta_{c,\mathbf{a}}$ the prior distributions become:

$$f(\theta \mid \mathbf{t}, s) \propto \prod_{x \in \mathcal{C}} \left[\theta_c^{st(c)-1} \prod_{i=1}^{k} \prod_{a_i \in \mathcal{A}_i} \theta_{a_i|c}^{st(c,a_i)-1} \right] \tag{3.2}$$

where $t(\cdot)$ corresponds to $n(\cdot)$. This prior Dirichlet distribution [7] has the following constraints:

$$\sum_{c \in \mathcal{C}} t(c) = 1 \tag{3.3}$$

$$\sum_{a_i \in \mathcal{A}_i} t(a_i, c) = t(c) \tag{3.4}$$

$$t(a_i, c) > 0 \tag{3.5}$$

For all (i, a_i, c).

When we multiply our likelihood by this prior density get a posterior in the same form.

Bibliography

[1]

[2] PhD thesis.

[3] D. J. Spiegelhalter D. Michie and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. 1994.

[4] Paul E. Keller Kevin L. Priddy. *Artificial Neural Networks: An Introduction*. SPIE Press, 2005.

[5] Konstantinos Koutroumbas S. Theodoridis. *Pattern Recognition*. Elsevier Science, 2003.

[6] J. Schlimmer. *Automobile Data Set*, 1987 (accessed November 8, 2016). `https://archive.ics.uci.edu/ml/datasets/Automobile`.

[7] M. Zaffalon. Statistical inference of the naive credal classifier. 2001.