Credal Classification of Automobile Risk

Ben Willis

January 24, 2017

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 [3]. There are many applications of classifiers, ranging from image recognition to sentiment analysis.

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.

1.2 Auto mobile Insurance

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 [4].

Initially we will discard objects with missing values and discretize all continuous attributes.

Chapter 2

MLE estimate for NBC

First we look at the maximum likelihood estimate for the naive Bayes classifier.

2.1 Notation

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 will also 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.

Finally after making observations of the attributes and class of N objects. We denote the frequency of those in class c by $n(c)$ and those in class c with attribute a_i by $n(a_i, c)$. We have the following structural constraints:

$$0 \leq n(a_i \mid c) \leq n(c) \tag{2.1}$$

$$\sum_{a_i \in \mathcal{A}_i} n(a_i|c) = n(c) \tag{2.2}$$

$$\sum_{c \in \mathcal{C}} n(c) = N \tag{2.3}$$

2.2 Assumptions

Both the NCC and the NBC share the naivety assumption [5]. This is the assumption that the features of an object are independent [2]. Hence:

$$\theta_{\mathbf{a}|c} = \prod_{i=1}^{k} \theta_{a_i|c} \tag{2.4}$$

This assumption greatly simplifies the problem.

They also make use of Bayes' theorem which allows us to rewrite the probability of an object belonging to a class like so:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2.5}$$

2.3 Likelihood Function

Using eqs. (2.4) and (2.5) we can derive 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:

$$l(\theta \mid \mathbf{n}) \propto \prod_{c \in \mathcal{C}} \left[\theta_c^{n(c)} \prod_{i=1}^{k} \prod_{a_i \in \mathcal{A}_i} \theta_{a_i|c}^{n(c,a_i)} \right] \tag{2.6}$$

2.4 Maximum Likelihood Estimate

We can derive the maximum likelihood estimate from this function.

First we take the log likelihood:

$$L(\theta \mid \mathbf{n}) \propto \sum_{c \in \mathcal{C}} n(c) log(\theta_c) + \sum_{c \in \mathcal{C}} \sum_{i=1}^{k} \sum_{a_i \in \mathcal{A}_i} n(c, a_i) log(\theta_{a_i|c}) \tag{2.7}$$

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

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

$$g(\theta, \mathbf{n}) = \sum_{c \in \mathcal{C}} \theta_c - 1 \tag{2.9}$$

This gives us our Lagrangian:

$$\mathcal{L}(\theta, \mathbf{n}, \lambda) = \sum_{c \in \mathcal{C}} n(c) log(\theta_c) - \lambda(\sum_{c \in \mathcal{C}} \theta_c - 1) \tag{2.10}$$

Differentiating with respect to θ_c we have:

$$\nabla_{\theta_c} \mathcal{L}(\theta, \mathbf{n}, \lambda) = \frac{n(c)}{\theta_c} - \lambda \tag{2.11}$$

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.

2.5 Classification

We can use the maximum likelihood estimates for these chances to create a naive Bayes Classifier. We can estimate $P(c|\mathbf{a})$ with our maximum likelihood estimates for our theta chances.

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* [1].

In this case we set $k = 10$ and the estimated accuracy was %66.54.

This is pretty bad, can we improve?

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

[2] I. Rish. An empirical study of the nave bayes classifier. 2001.

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

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

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