Credal Classification of Automobile Risk

Ben Willis

January 6, 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 [2]. 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 dataset we're analysing contains 24 attributes that give information about the type of vehicle. These range from number of doors to engine location. The data set also contains a risk assigned by an expert. This risk is initially assigned by price and then, if it is more (or less) risky, shifted up (or down). This is known as symboling [3].

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 [4]. This is the assumption that the features of an object are independent [1]. 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:

$$\theta_{c|\mathbf{a}} = \frac{\theta_c \theta_{\mathbf{a}|c}}{\theta_{\mathbf{a}}} \tag{2.5}$$

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 lagrange multipliers. These allow us to solve problems in the form max f(x,y) s.t. g(x,y) = 0.

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.

Bibliography

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

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

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

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