Information Processing and Management 50 (2014) 653–674

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A new decision to take for cost-sensitive Naïve Bayes classifiers

Giorgio Maria Di Nunzio *

Department of Information Engineering, University of Padua, Padova, Italy

ARTICLE INFO

Article history:
Received 22 August 2013
Received in revised form 17 April 2014
Accepted 28 April 2014
Available online 2 June 2014

Keywords:
Cost sensitive learning
Bayesian decision theory
Binary classification
Classical probabilistic models

ABSTRACT

Practical classification problems often involve some kind of trade-off between the decisions a classifier may take. Indeed, it may be the case that decisions are not equally good or costly; therefore, it is important for the classifier to be able to predict the risk associated with each classification decision. Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. The objective is to quantify the trade-off between various classification decisions using probability and the costs that accompany such decisions. Within this framework, a loss function measures the rates of the costs and the risk in taking one decision over another.

In this paper, we give a formal justification for a decision function under the Bayesian decision framework that comprises (i) the minimisation of Bayesian risk and (ii) an empirical decision function found by Domingos and Pazzani (1997). This new decision function has a very intuitive geometrical interpretation that can be explored on a Cartesian plane. We use this graphical interpretation to analyse different approaches to find the best decision on four different Naïve Bayes (NB) classifiers: Gaussian, Bernoulli, Multinomial, and Poisson, on different standard collections. We show that the graphical interpretation significantly improves the understanding of the models and opens new perspectives for new research studies.

1. Introduction

The task of data classification is today commonly applied in many contexts, ranging from customer target marketing to medical diagnosis, from biological data analysis to document categorisation. Practical classification problems often involve some kinds of constraints with respect to the effectiveness of the classifier, which is usually measured in terms of false-positive and/or false-negative rates. In some domains, these two rates may not be equally important. For example, for a spam classification system, a mis-classified legitimate email is generally considered unacceptable, while a spam message classified as non-spam is less serious (Kolcz, 2005).

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification (Duda, Hart, & Stork, 2001). The objective is to quantify the trade-off between various classification decisions using probability and the costs that accompany such decisions. In this cost-sensitive framework, we can choose a learning algorithm and optimise its performance by tuning not only its parameters but also its misclassification costs. For example, in binary classification problems, i.e. when we have only two categories c_1 and c_2, the Bayes decision rule can be interpreted as calling for deciding one category for an object o if the likelihood ratio (which is the ratio between the two probabilities $P(o|c_1)$ and $P(o|c_2)$)

* Tel.: +39 049 8277613.
E-mail address: dinunzio@dei.unipd.it
URL: http://www.dei.unipd.it/~dinunzio

http://dx.doi.org/10.1016/j.ipm.2014.04.008
0306-4573/

exceeds a threshold value t. This threshold is independent of the observation o and can be tuned by means of the misclassification costs (the details of this formulation are presented in Section 3):

$$t < \frac{P(o|c_1)}{P(o|c_2)} \tag{1}$$

where t is the threshold that depends on the misclassification costs and on the priors of the two categories, $P(c_1)$ and $P(c_2)$. This way of optimising classifiers is very effective for unbalanced binary classification tasks (Almeida, Almeida, & Yamakami, 2011; Metsis, Androutsopoulos, & Paliouras, 2006).

In the literature of pattern classification (Duda et al., 2001), NB classifiers have been shown to be one of the most efficient and effective inductive learning algorithms for classification tasks, despite the strong unrealistic assumptions (see Appendix A). The work by Domingos and Pazzani (1997) and the further study by Zhang (2005) define the conditions under which the NB classifier is an optimal classifier. An important consideration to take into account when working with NB classifiers is that there are simple linearly separable cases where the Bayesian classifier fails to predict the correct class; however, quoting (Domingos & Pazzani, 1997) "a simple modification of the Bayesian classifier will allow it to perfectly discriminate all positive examples from negatives: adding a constant to the discriminant function for the concept, or subtracting the same constant from the discriminant function for its negation". This decision can be written in the following way:

$$P(o|c_2)P(c_2) < P(o|c_1)P(c_1) + l \tag{2}$$

This constant l, which has an empirical justification, hides a more complex interpretation of the costs in the context of cost-sensitive learning. In fact, it cannot be directly derived from the definition of costs in the cost-sensitive learning context. To the best of our knowledge, this simple step, which has been empirically shown to be very effective, has never been formally proven.

The main contributions of this paper are:

- A formal justification for a decision function under the Bayesian decision framework that comprises both (i) the minimisation of Bayesian risk and (ii) an empirical decision function found by Domingos and Pazzani (1997). The decision have the following linear form:

  $$P(o|c_2) < mP(o|c_1) + q$$

  where m and q depend on the mis-classification costs and can be seen as the angular coefficient and the intercept of a linear function. Note that for $q = 0$ we can derive Eq. (1) with $m = \frac{P(c_2)}{P(c_1)}t$, while for $m = 1$ we obtain Eq. (2) with $q = l$.
- Since this new decision function has a very intuitive geometrical interpretation that can be explored on a Cartesian plane, we present an adaptation of the Angular Region algorithm (Di Nunzio & Micarelli, 2004) that can efficiently find a (sub)optimal decision on a two-dimensional space.
- We use this graphical interpretation to analyse different approaches to find the best decision on four different Naïve Bayes (NB) classifiers: Gaussian, Bernoulli, Multinomial, and Poisson, on different standard collection. We show that the graphical interpretation significantly improves the understanding of the models and opens new perspectives for new research studies.

The paper is organised as follows: in Section 2, we define the task of binary classification for NB classifiers. In Section 3, we present the Bayesian decision theory framework that is at the base of our formulation of the problem. Section 4 defines the new conditions and costs under which an optimal decision function which merges both Eqs. (1) and (2) can be found. In Sections 5 and 6 we present the experimental analysis and the discussion of the results, respectively. In Section 7, we suggest some of the related works, while in Section 8, we give our final remarks.

2. Binary classification

Binary classification is the task of classifying objects into two classes on the basis of some properties of the objects. The usual notation to indicate these two classes is: c for the class of 'positive' examples, and $\bar{c}$ for the class of 'negative' examples. Often, real world classification problems have more than two classes, for example a set of classes $C = \{c_1, \ldots, c_i, \ldots, c_n\}$. In these cases, a common approach in machine learning is to define n binary classification problems, one for each class in the set C. Given an object o and a set of categories C, if we want to decide whether o should be assigned to category $c_i \in C$, we can build a simple probabilistic classifier that checks the following statement:

$$P(\bar{c}_i|o) < P(c_i|o) \tag{3}$$

where $\bar{c}_i = C \setminus c_i$. Therefore, if the probability of the class c_i is greater than the probability of its complement $\bar{c}_i$ we can assign the object to c_i.[1] Since we do not know the value of $P(c_i|o)$ of unseen objects (unless an 'oracle' tells us what the value of $P(c_i|o)$), in order to predict the probability $P(c_i|o)$ we need to reverse it by using the Bayes rule:

[1] You may have noticed an inverted use of the inequality, which is usually written as $P(c_i|o) > P(\bar{c}_i|o)$. This will help us maintain the same order (less than) when presenting the Bayesian conditional risk in Section 3.