Teaching Machine Learning: A Geometric View of Naïve Bayes

Giorgio Maria Di Nunzio

Deptartment of Information Engineering
University of Padua, Italy
dinunzio@dei.unipd.it
http://www.dei.unipd.it/~dinunzio/

Abstract. In this demo, we present two applications which allow users to 'see' a geometric interpretation of the Bayes' rule and interact with a Naïve Bayes text classifier on a real dataset, namely the Reuters-21578 newswire collection. The main objective of this demo is to show how the pattern recognition capabilities of the human increase the effectiveness of the classifier even when technical details are not known in advance or the user is not an expert in the field. These two applications were developed with the R package Shiny; they have been deployed online and they are freely accessible from the links indicated in the paper.

1 Introduction

When we want to quantify the uncertainty of the outcome of an experiment, we can use Bayesian modelling to build the mathematical model of the experiment. In this context, Bayes' rule is used to compute the posterior probability of a variable given some observed data. Posterior probabilities can be hard to compute; therefore, a "naïve" solution is to make some assumptions that allow for a factorisation of the posterior probability into simple conditional probabilities which are easy to compute [3]. Naïve Bayes (NB) classifiers have been widely used in the literature of Data Mining and Machine Learning since they are easy to train and reach satisfactory results which are comparable to the results of more complex state-of-the-art classification algorithms. However, the optimisation of the parameters of NB classifiers is often not adequate, if not missing at all. Based on the idea of Likelihood Spaces, a two-dimensional representation of probabilities [2], we have developed two Web applications which provide an adequate data and knowledge visualization to teach in a real machine learning setting how parameter optimisation and misclassification costs affect the performance of the classifier. In this demo, we show a geometric interpretation of the Bayes' rule which can be used to teach to non-experts how NB works and how to optimise the parameters of these classifiers in a very intuitive way. In addition, we present a real text classification problem based on the Reuters 21578 collection.[1]

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

2 Mathematical Background

Bayes' rule gives a simple but powerful link between prior and posterior probabilities of events. For example, assume that we have two classes c_1 and c_2 and we want to classify objects according to some measurable features of the objects. The probability that an object o belongs to c_1 is:

$$\overbrace{P(c_1|o)}^{posterior} = \frac{\overbrace{P(o|c_1)}^{likelihood}\overbrace{P(c_1)}^{prior}}{P(o)} \tag{1}$$

Bayes' rule tells us that, starting from a prior probability on c_1, we may change our idea about the probability of class c_1 after observing the object o, according the 'likelihood' of that object. If o is represented by a set of features $F = \{f_1, f_2, ..., f_j\}$, a Naïve Bayes approach factorises $P(o|c_1)$ as:

$$P(o|c_1) = \prod_j P(f_j|c_1) \tag{2}$$

where features are independent from each other given the class, that is the conditional independence assumption.

3 Demo

The first part of the demo[2] shows the implications of the visual interpretation of the Bayes' rule on a two-dimensional space. In the second part of the demo,[3] we show how this two-dimensional space can be used in a real scenario of text classification using the Reuters-21678 newswire collection.

Bayes' rule If we imagine $P(c_1|o)$ and $P(c_2|o)$ as two coordinates of a cartesian space, we can draw objects on the segment with endpoints $(1, 0), (0, 1)$ (since $P(c_1|o) = 1 - P(c_2|o)$). In Fig. 1, objects are represented by three binary features f_1, f_2, and f_3 (this is called a multivariate Bernoulli NB classifier). Each point in the plot represents an object (three binary features, $2^3 = 8$ objects), when a point is below the line that passes through the origin (i.e. $P(c_1|o) > P(c_2|o)$), it is classified under c_1. The user can change the values of the conditional probability for each feature and class and see the effect in terms of the position of the points.

The demo allows users to study situations that are more realistic in terms of real machine learning problems. First, it is often the case that the normalisation factor $P(o)$ is not computed when classifying the object because it is a constant that does not change the classification. By de-selecting it from the interface, we see how the coordinates change accordingly ($P(c_1|o) \propto P(o|c_1)P(c_1)$). Users can also adjust the prior $P(c_1)$ and see the effect of an "unbalanced" class situation.

[2] `https://gmdn.shinyapps.io/bayes2d/`
[3] `https://gmdn.shinyapps.io/shinyK/`

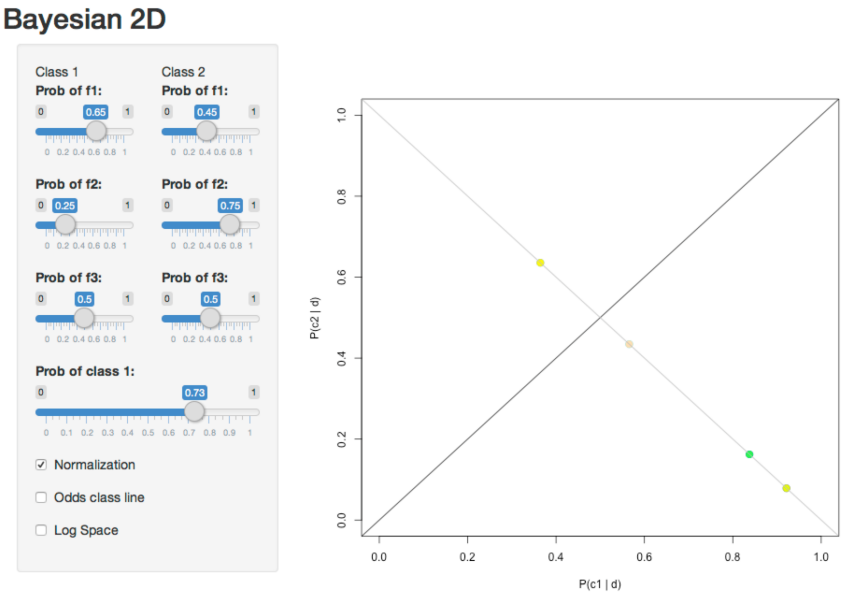

Fig. 1. Teaching how Bayes' rule work on a two-dimensional space.

When normalisation is not active, we can even choose whether to shift points or rotate the classification line by selecting 'Odds class line' ($P(c_1|o) \propto P(o|c_1)$). In addition, it is possibile to show the coordinates of the objects as logarithms of probabilities by selecting 'Log space' on the interface. Therefore we can study the problem in terms of log-likelihood, $\log(\prod P(f_j|c_1)) = \sum \log(P(f_j|c_1))$.

Newswires classification task In the second part of the demo, we show a real text classification problem with a multivariate NB classifier. The interface allows users to: choose one out of ten categories, select the number of training/validation folds and features for training the classifier, smooth probabilities to avoid zero probabilities [5], adjust misclassification costs [4]. In real cases like this one, the logarithm transformation is necessary. Starting from default parameters, the user can interact with the model and find the parameters that produce a good separation between the two classes with little effort and without necessarily being an expert in the field [1].

4 Conclusions

The objective of this demo was two-fold: to introduce a geometric interpretation of Bayes' rule which can be used to teach how NB classifier works in an intuitive way; to show how the pattern recognition capabilities of the human can improve

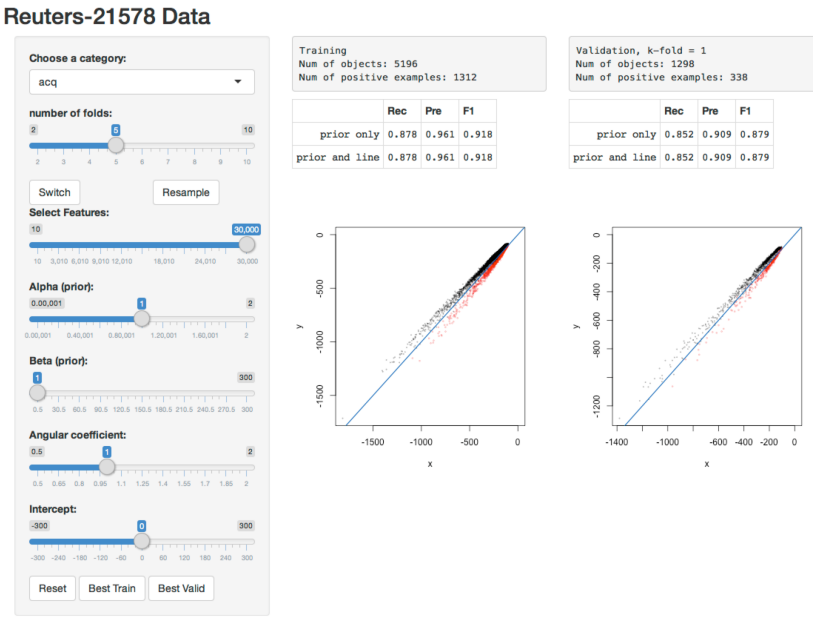

Fig. 2. Optimizing a NB classifier on a real text classification problem.

the effectiveness of the default NB classifier even when technical details are not known in advance. There are still some interesting open questions about the meaning of the parameters that are found by visual inspection compared to other solutions found by automatic optimization approaches.

References

1. Mihael Ankerst, Martin Ester, and Hans-Peter Kriegel. Towards an effective cooperation of the user and the computer for classification. In *Proceedings of the sixth ACM SIGKDD 2000*, pages 179–188, 2000.
2. Giorgio Maria Di Nunzio. A new decision to take for cost-sensitive naïve bayes classifiers. *Information Processing & Management*, 50(5):653 – 674, 2014.
3. Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Mach. Learn.*, 29(2-3):103–130, November 1997.
4. Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, pages 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
5. Quan Yuan, Gao Cong, and Nadia Magnenat Thalmann. Enhancing naive bayes with various smoothing methods for short text classification. In *Proceedings of the Int. Conf. WWW'12*, pages 645–646, New York, NY, USA, 2012. ACM.