Probabilistic approach over Decision Trees for problems with discrete data with a small number of instances 2017 — Norwich, UK

Luke M. Garrigan

Machine Learning, University of East Anglia, UK
l.garrigan@uea.ac.uk

Abstract

Bayesian classifiers are widely known for their optimality when attributes are independent given the class. This paper attempts to prove that small samples of discrete data with arbitrary dependencies are more accurately classified using a probabilistic approach over decision trees.

Index Terms: Probabilistic, Naive Bayes, Decision Trees

1 Introduction

In machine learning a learner algorithm is given a set of training instances with their corresponding class labels, it then produces a classifier. The classifier takes unlabelled testing instances and assigns it to a class. Choosing the best suited algorithm specific to the sample set is not a trivial process.

Probabilistic classification is the application of approximating a joint distribution with a product distribution. Bayes rule is used to approximate the conditional probability of a given class label. Approaches such as *naive* Bayes are among the most popular classifiers used in the machine learning community, derived from generative probability models they are generally easy to understand and the induction of these classifiers is extremely fast, requiring only a single pass through the data if all attributes are discrete [1]. The *naive* Bayes classifier is the simplest of models in this paper, it assumes that all attributes are independent of each other given the context of the class. Although the *naive* assumption of independence is not true in terms of most sample sets, many papers such as [2] have proven that *naive* Bayes classification accuracy is very competitive when compared with more complex state-of-the-art algorithms.

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

Decision trees classify instances by sorting them down the tree from the root to some leaf node which represents the classification of the given instance. Nodes specify a test of some attribute of the instance and each branch from that node corresponds to one of the possible values for this attribute. A given instance is classified moving down the tree, the attribute specific to that node is tested. Following down the branch corresponding to the value of the attribute in the given example, this is then repeated until a leaf node is reached and a classification is made.

2 Data Description

2.1 Discretization

Discretization concerns with the process of transferring continuous data into discrete counterparts. Numeric attributes were discretized into ten equal-length intervals, unless the number of uniquely observed values for an attribute was less than 10. This approach was compared in [3] with entropy-based and purity-based methods, which are supervised algorithms. An empirical evaluation showed that the *naive* Bayes algorithm significantly improved accuracy when features were discretized using an entropy-based method. However due to its simplicity the unsupervised binning discretization method was used.

2.2 Methods For Accuracy Estimation

2.2.1 K-Fold Cross-Validation

Cross-validation is a computationally expensive algorithm used to estimate performance, it uses all available instances as both training and testing sets. The dataset is split into k equally sized non-overlapping subsets S. Given a fold S_i a model is trained on $S \setminus S_i$, then S_i is used to create the accuracy estimation.

2.2.2 Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross-Validation is K-fold cross validation where K is equal to the number of instances in the dataset. The classifier is trained on all data expect the one instance being left out and the prediction is made. An average of the data is collected and used to evaluate the classifier.

2.2.3 Bias And Variance Tradeoff

As k increases the less bias the classification is in overestimating the true expected error because the folds will be closer to the total dataset. However in doing this it induces a significant amount of variance. To minimise the testing bias a large portion of the dataset must be used for training, meaning not much data is used for testing, this ensures that the model will be as close as possible to the the one that would be achieved from training using the entire dataset. Minimising the testing variance would mean quite the opposite, so a large amount of data would be used for testing, this ensures a more reliable estimate of error of the classifier.

Due to the small sample sizes 10-fold cross-validation was used as an estimator in an attempt to minimise the estimation variance.

Table 1: An example table.

Data Set	Instances	Attributes
Echo cardiogram	131	13
Teaching Assistant	150	6
Seeds	209	8
Planning Relax	180	13
Hepatitis	154	20
Breast Cancer	284	10
Ecoli	223	10
Glass	141	6
Haberman	203	4
Hayes Roth	158	5
Heart	169	14
Lymphography	156	19
Promoters	104	58
Shuttle Landing	252	7
Sonar	137	61
Thyroid	142	6

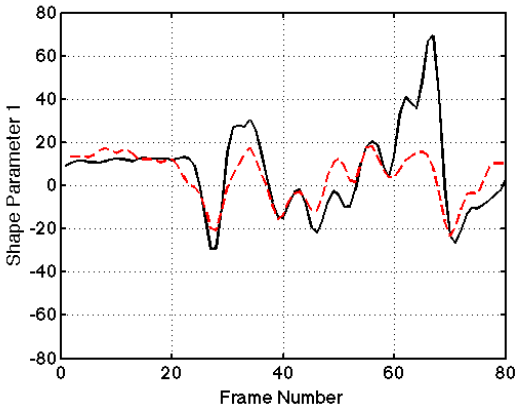

Figure 1: An example of a figure centred on a single column. The figure is not centred automatically. To centre the figure, use `\centering` within the `figure` environment.

2.3 Sections

Section headings should be centred on the line, be in bold typeface, and only the first letter should be capitalised. Sub-headings are also in bold face, but appear flush left and are typeset in the base font size. Sub-sub-headings appear like sub-headings, except they are in italics and are not boldface. No more than 3 levels of headings are allowed.

2.4 Headers and Footers

All headers and footers must be left empty. Your document should not contain page numbers etc. These will be added later.

2.5 Lists

Itemised lists can be included in your document, but please check the indentation if you are not using LaTeX. An example itemized list with the correct formatting should look like the following:

- First list item.
- Second list item.
- Third list item.

2.5.1 List Depth

Please try not to use hierarchical lists, these look cluttered in two-column format. Keep lists to a single level of depth.

2.6 Figures

All figures must be centred on the column (or page, if the figure spans both columns). A figure caption should follow the figure and be formatted like the example in Figure 1.

Figures spanning multiple columns should appear either at the top or the bottom of the page. You can use the \figure* command to span a figure across both columns. An example is shown in Figure 2.

Figures should preferably be line drawings. The proceedings will not be produced in colour, so please do not rely on colour to distinguish between curves on a graphs, etc. You should check to ensure the figures print well on a good quality printer, and that there are no issues when colour figures are printed in grey-scale.

Before including any figures in your document you should ensure you have the necessary copyright permission.

2.7 Tables

Tables should be centred on a column if possible. There is no strict requirement on the style as this will largely depend on the content to be displayed. An example table is shown in Table 2, but this is provided for illustrative purposes only.

Table 2: An example table.

Trial	Score
1	10
2	12
3	11
4	9
5	11
6	10

Note, for tables the caption should be above the table, as shown in Table 2.

2.8 Equations

Equations should appear on a separate line, they should be centred and they should be numbered. Some examples are:

$$y = mx + c, \tag{1}$$

which obviously is the equation of a straight-line of gradient m and intersecting the vertical axis at c. Another famous equation with m and c is

$$E = mc^2. \tag{2}$$

2.9 Fonts

You should use 9 point Times or Times Roman for the main text. All fonts should be embedded in the final PDF document.

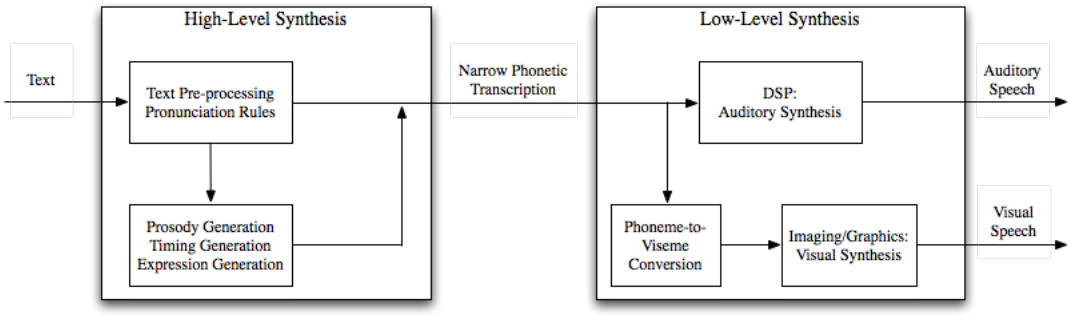

Figure 2: An example of a figure spanning both columns and centred on a page. Again, the figure is not centred automatically.

2.10 Hyperlinks

Hyperlinks should be written in full, e.g., `http://www.cmp.uea.ac.uk`, and must be coloured black. For ease of readability, authors are advised to use a different font family from the main text.

2.11 Supplementary Material

Authors may submit supplementary material with their paper. However, this should not be included in place of a technical description of your work. Reviewers are not obliged to watch video sequences, and your submission will be reviewed on the strength of the paper only. If you are submitting a multimedia file, please use widely accepted formats/codecs. The conference proceedings will not include media players.

3 Conclusions

The page limit is 4–6 pages. Please, please use LaTeX to typeset your document. This will minimise any formatting headaches!

3.1 References

You must reference any papers you have had accepted or are under review here. Example references [**?**], audio-visual speech synthesis [**?**, **?**], and audio-visual speech recognition [**?**, **?**].

See the references section for the formatting of the references from different sources (conferences, journals, and books). The formatting of references follows the standard IEEE format, LaTeX users should download the `IEEEtran` bibliography format. References should be listed in order of citation.

References

[1] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." in *KDD*, vol. 96. Citeseer, 1996, pp. 202–207.

[2] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2, pp. 103–130, 1997.

[3] J. Dougherty, R. Kohavi, M. Sahami *et al.*, "Supervised and unsupervised discretization of continuous features," in *Machine learning: proceedings of the twelfth international conference*, vol. 12, 1995, pp. 194–202.