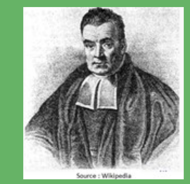
INM431 - Machine Learning Coursework, November 2015

Sanjoy Sarkar and Suzanne Fox



Comparison of Naïve Bayes and Decision Tree machine learning algorithms for UCI Student dataset

The dataset

Data Source : https://archive.ics.uci.edu/ml/datasets/Student+Performance

This is a social dataset collected from 2 secondary schools in Portugal during the 2005-2006 academic year. Grades in Maths and Portuguese for the 1st and 2nd terms, and the final grade are the target data. Attribute variables include demographic, social and school related features. Data were collected about students from school reports and self-completion questionnaires. There are 2 datasets, 1 for each academic subject, containing 30 attribute variables and 3 target variables.

Considerations :

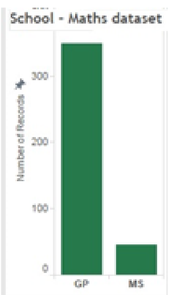


Figure 1. Maths dataset

Bias : The Maths dataset is highly skewed in favour of one school (see Fig. 1). As teaching standards might vary considerably between schools, we chose not to consider the Maths dataset.

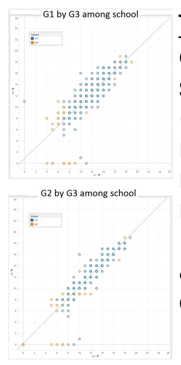


Figure 2. G1 x G3, G2 v G3

Target Variable : Logically enough, G1 and G2 are strongly correlated with G3 (see Fig. 2) and thus were not of interest. G3 is recorded as numerical data, ranged from 0-20. A score of 10 or more is a pass mark, anything less is a failing grade.

The dataset is composed of 17 Nominal variables (13 binary), 11 Ordinal and 2 Ratio. The target variable (G3) is supplied ranged from 0-20 (see Fig 3). For our classifier G3 is binned into a binary attribute with Pass and Fail criteria.

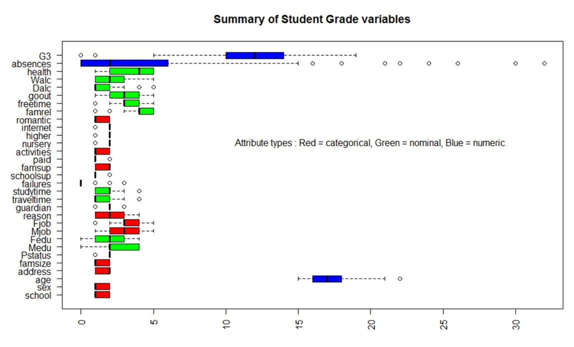


Figure 3. Visual summary of attribute variables

school	student's school	Categorical, Binary	"GP" - Gabriel Pereira "MS" - Mousinho da Silveira
sex	student's sex	Categorical, Binary	"F" / "M"
age	student's age	Numeric	15 to 22
address	home address type	Categorical, Binary	"U" – urban /"R" – rural
famsize	family size	Categorical, Binary	"LE3" - less or equal to 3 "GT3" - greater than 3
Pstatus	parent's cohabitation status	Categorical, Binary	"T" - living together "A" – apart
Medu	mother's education	Numeric	0 - none, 1 - 4th grade, 2 – 5th to 9th grade, 3 – secondary 4 – higher
Fedu	father's education	Numeric	0 - none, 1 - 4th grade, 2 – 5th to 9th grade, 3 – secondary 4 – higher
Mjob	mother's job	Numeric	"teacher", "health", civil "services" (e.g. admin or police), "at_home", "other"
Fjob	father's job	Numeric	"teacher", "health", civil "services" (e.g. admin or police), "at_home", "other"
reason	reason to choose this school	Numeric	close to "home", school "reputation", "course" preference or "other"
guardian	student's guardian	Numeric	"mother", "father" or "other"
traveltime	home to school travel time	Numeric	1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
studytime	weekly study time	Numeric	1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
failures	number of past failures	Numeric	n if 1<=n<3, else 4
schoolsup	extra educational support	Categorical, Binary	Yes/No
famsup	family educational support	Categorical, Binary	Yes/No
paid	extra paid classes	Categorical, Binary	Yes/No
activities	extra-curricular activities	Categorical, Binary	Yes/No
nursery	attended nursery school	Categorical, Binary	Yes/No
higher	wants higher education	Categorical, Binary	Yes/No
internet	Internet access at home	Categorical, Binary	Yes/No
romantic	with romantic relationship	Categorical, Binary	Yes/No
famrel	quality of family relationships	Numeric, Likert scale	1 - very bad to 5 – excellent
freetime	free time after school	Numeric, Likert scale	1 - very low to 5 - very high
goout	going out with friends	Numeric, Likert scale	1 - very low to 5 - very high
Dalc	workday alcohol consumption	Numeric, Likert scale	1 - very low to 5 - very high
Walc	weekend alcohol consumption	Numeric, Likert scale	1 - very low to 5 - very high
health	current health status	Numeric, Likert scale	1 - very bad to 5 - very good
absences	number of school absences	Numeric	0 to 93
G3	final grade	Numeric	0 to 20

Figure 4. Attribute variables

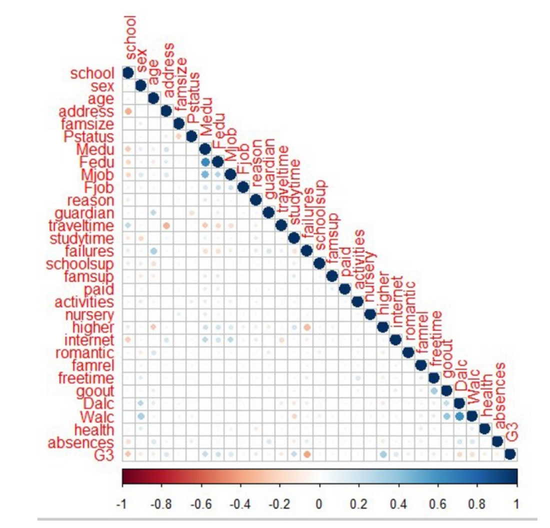


Figure 5. Cross-correlation matrix of attributes

Description and motivation of the problem

This coursework considers a supervised learning problem where we attempt to learn a classifier for grade results of secondary school students studying Portuguese. We consider the nature and provenance of the data and choose two suitable algorithms. We then make a hypothesis about which will model the data better and why. An experiment is constructed to determine the validity of hypothesis. If the hypothesis is valid, we explore under what conditions it holds true.

The choice of dataset and problem was motivated by an interest in social datasets. The domain we chose was education grades achieved by secondary school pupils. This dataset provides the opportunity to explore underlying factors contributing to academic achievement.

Summary of the ML methods used

We chose to use a Naïve Bayes classifier and a Decision Tree to learn this data. The basis for these choices were the nature of the problem - i.e. supervised classification - and the provenance of the data - i.e. mixed collection methods, mixed data types and likely interrelationships between variables. Different classification models and their metrics perform better or worse given the structure of the data [2] and we wanted to explore the difference between a probabilistic model (naïve Bayes) and a deterministic one (Decision Tree)

Naïve Bayes assumes predictor variables are independent, but it appears to work well when this condition does not exist [3] [4]. Naïve Bayes has a training step where the parameters of a probability distribution are estimated from the training data, and a testing step where each test sample is classified according to its largest posterior probability among the target classes.

Decision Trees is a deterministic algorithm where the data is modelled as a hierarchy of connected nodes. At each node the data is split into two or more sub-nodes according to the specific algorithm in operation and the model parameters. MATLAB uses the C4.5 algorithm, which calculates information gain provided by each data attribute to determine the best split for the data at that node. This process is iterated over the dataset. Trees which succeed at placing attributes with higher information gain early are the most useful.

	Pros	Cons
Naïve Bayes	Easy to use Works well in practice even when data has dependencies Computationally inexpensive Only needs a small amount of training data to calculate probabilities	Assumes variables are independent Difficult to explain results Can't learn interactions between attributes
Decision Tree	Highly tolerant of spurious data Easy to understand and explain/simple to interpret model Tolerant of dependencies Depth/performance of the tree controlled by parameters Performs implicit feature selection Data does not require normalising or excessive preparation Performance unaffected by non-linear relationships between data attributes	Simplistic Easier to overfit if parameter selection is not well chosen Might get stuck in local minima Quite complex C4.5 Computationally expensive

Figure 6. Pros and Cons of naïve Bayes and Decision Trees

Hypothesis statement

A Decision Tree will model this data more accurately than a Naïve Bayes Classifier. This hypothesis is based on the following features of the dataset –

- there are many dependencies between the variables, as seen from the cross-correlation matrix
- the data has mixed collection methods (school reports, questionnaires) and thus reliability and/or accuracy of the attributes is not uniform
- the Decision Tree has parameters which can be adjusted so we should be able to refine the model

Choice of training and evaluation methodology

Scope of the experiment

Scenario 1 - MATLAB classifiers with default parameters were created; first using Naïve Bayes (NB) algorithm, second using Decision Tree (DT) algorithm. The results were used to test hypothesis validity.

As NB is non-parametric, the default model was compared with various scenarios of optimized DT model.

Scenario 2 – An extensive grid search method was used to optimise both the DT parameters 'Minimum Leaf Size' (MLS and 'Minimum Parent Size' (MPS) by running 900 (30 x 30) models searching for the one with the lowest error rate. The results were used to test hypothesis validity.

Scenario 3 - As scenario 2 was likely to overfit the data, we looked for a more general model. The results were used to test hypothesis validity.

Evaluation : Accuracy and F (or F1) Measure calculated from confusion matrix stats was plotted as a moving average over a set of 20-Fold cross validation iterations. To compare model performance we plotted both measures for both models on a single graph. We also looked at the test and training set error from both models on the same basis.

Training experiment : We experimented with creating a 10 fold stratified sample, distributing attributes school and target evenly between folds. We plotted MATLAB default NB and DT with 10-fold cross validation. Comparing with use of the MATLAB cvpartition function to randomly create the folds shows both models perform equivalently, although the moving average for the stratified sample was smoother.

Training Method : We used 20 fold cross validation, giving a sample count of 32 for the testing phase of each iteration, which was the minimum required to give a reasonable result.

Choice of parameters and experimental results

Scenario 1

The initial experiment was to benchmark both models using the MATLAB default parameters for the DT and compare the results.

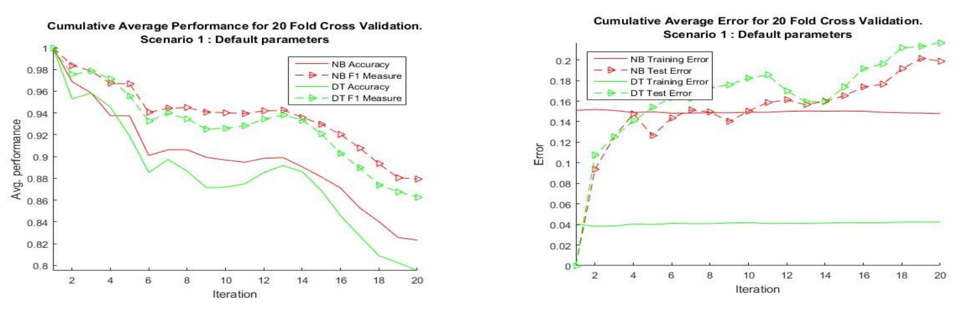


Figure 7. Comparative Performance and Error Graphs for Scenario 1

It can be seen that "out of the box" the NB classifier's generalisation performance is better in comparison to DT. Interesting to note from the error graph that though training error for DT is far less than that of NB, generalisation testing error is higher than NB after 4 folds. In this scenario the hypothesis fails.

Choice of parameters

We looked for optimal parameters by running various analyses of the DT statistics among several models. We looked at both graphical and numerical representations of the output.

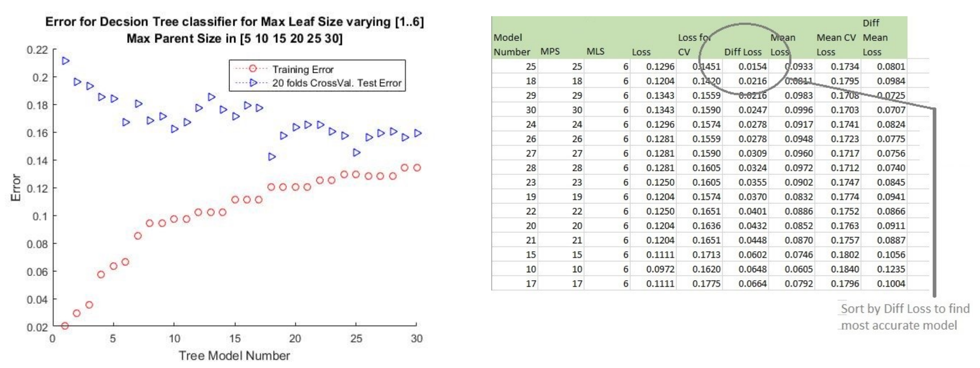


Figure 8. Example of searching for optimum DT parameters, error graph and numerical output

Scenario 2

After some iterations, we used MLS=6, MPS=25 as the optimum parameters which generalised well, without overfitting the data.

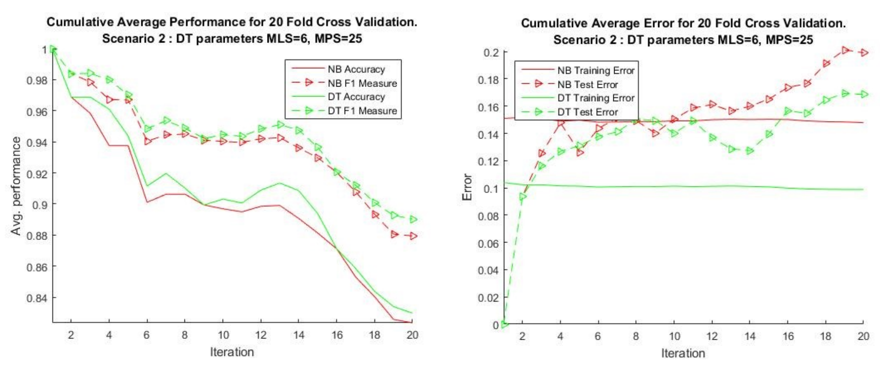


Figure 9. Comparative Performance and Error Graphs for Scenario 2

We came to this conclusion by considering the accuracy and error of the model, and by examining the DT built by the model. Some DTs were deep and unbalanced. For this combination of parameters the accuracy is greater than that of NB, the DT models the data in a shallow, balanced fashion. For this scenario the hypothesis holds true.

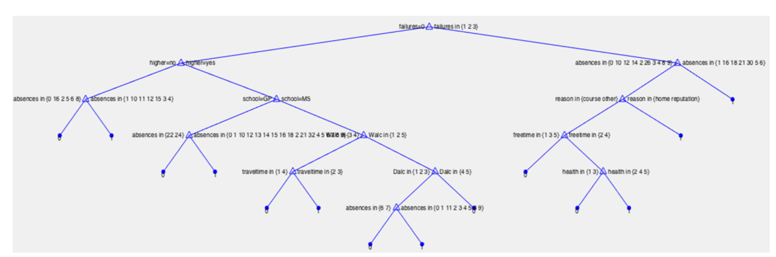


Figure 10. Decision Tree generated by Optimum model

Scenario 3

Using a more expansive grid search to generate 900 models (30 x 30 grid of MLS x MPS) allows the most accurate model to be determined. The tree produced by this model was fragmented and deep.

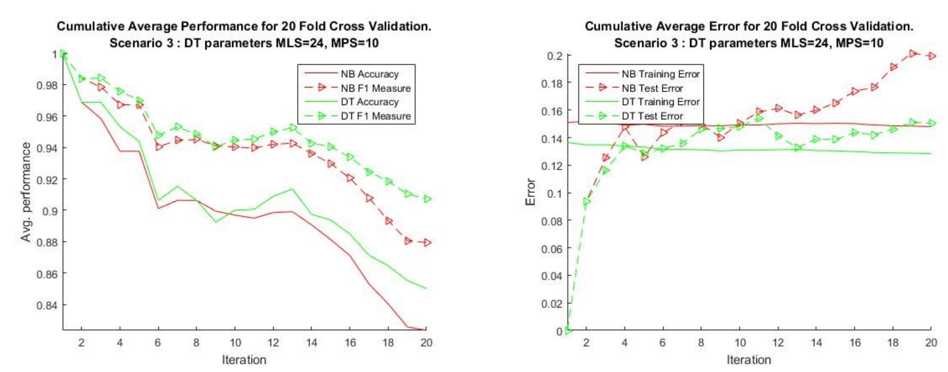


Figure 11. Decision Tree parameters generated by overfitted model

Results

For this dataset, the default Naïve Bayes MATLAB model classifies the data more accurately than the default Decision Tree and the hypothesis fails. However, when the Decision Tree parameters are optimised the hypothesis holds true and the Decision Tree is the more accurate model. It becomes easy to overfit the DT model when only considering accuracy and F1.

Analysis and critical evaluation of results

Scenario	DT v NB - Accuracy	DT v NB - F1
1	-2.00%	-2.00%
2	0.50%	1.00%
3	4.00%	3.00%

Figure 12. Comparative Accuracy and F1 for NB and DT across all 3 scenarios

Figure 12 shows how the comparative performance of Accuracy and F1 for the DT classifier and NB changed as the DT model parameters were altered. The accuracy of the DT improves from 2% less than NB, to 4% better (6% change overall). For F1 the improvement is 5% overall. Without looking at the tree that the DT model produces this comparison is not that all that meaningful. We discussed whether we could be confident that a higher accuracy meant a better, more generalised model, and concluded we could not rely on these statistics alone. Additionally looking at the trees produced by the various iterations of our models was necessary to interpret the results more analytically.

We selected Scenario 2 as our optimum model not only because it gave an improved performance over the NB model, but also because the tree was not too deep and it give results which were understandable and explainable in terms of the data attributes. The overfitted model produced for scenario 3, despite greater accuracy measures, gave a deeper, more unbalanced tree.

Several of the Pros and Cons of Decision Trees listed in Figure 6 can be demonstrated by the optimum tree we produced.

It was stated that one advantage of the DT model is that its "performance can be controlled by parameters", and this is clearly demonstrated by the results shown in Figure 12. A "con" is that DTs are easy to overfit and we saw this also in scenario 3.

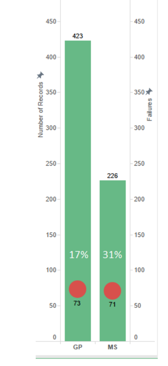


Figure 13. Failure rate

It was stated that one advantage of the DT algorithm is that it can "Perform feature selection" and to investigate this we looked at the five data attributes at the root of the optimised tree. These attributes were – failures, absences, reason (for choosing school), higher (students want to go onto higher education) and school. This feature selection enabled us to look more closely at the data, and there were interesting findings.

The failure rate for school MS is 31% compared with 17% for GP (Fig 13). Looking at the reasons for choosing schools, 13% of people choose MS based on its reputation compared with 27% for GP (Fig 14). A picture is building up of the difference between the two schools. MS has almost twice the failure rate of GP, and a worse reputation.

	School	
Reason	GP	MS
course	39.48%	52.21%
home	27.19%	15.04%
other	6.38%	19.91%
reputation	26.95%	12.83%
Grand Total	100.00%	100.00%

Figure 14. Reasons for choosing school

Looking at the "higher" attribute, 92% of students at GP want to go on to further education, compared to 84% at MS. The only attribute which, on the face of it, seems worse for GP is the level of absences, which averages 4.2 per pupil for GP vs 2.6 for MS. A possible explanation for this could be that older students have more flexibility over their schedule, and sure enough looking at the spread of ages between schools it becomes immediately obvious that MS only has students up to the age of 12, whereas GP has students up to the age of 20. The objective of the original data collection was to investigate what attributes contribute to scholastic performance, and, in that context, the findings of this analysis offer very little information. One school outperforms the other. Whether this is due to social attributes of the students is difficult to say. Looking at the data from this point of view suggests that an omission from the original dataset is any information which enables a comparison of the teaching standards in both schools. Similarly it would be interesting to understand the sampling methodology of the original study as the data does not appear to be very comparative.

This chain of investigation follows directly from the results of the DT, and would have not been prompted by looking at the NB classifier results, no matter how accurate. This clearly demonstrates the advantages of a model which is easy to understand and explain.

Conclusion

The optimised DT outperforms the NB classifier on performance statistics, but the margins are not that big. The real power of the DT, and the thing that makes it a far more useful algorithm in this example, is that it has given a useful insight into the nature of the data which enabled further exploration and understanding.

The original hypothesis is proven. The optimised Decision Tree models the data more accurately.

Lessons learned and future work

Deploying, optimising and comparing multiple models has been a very interesting and useful exercise. The models we chose worked well together. The NB gave us a good benchmark against which we were able to optimise the DT model. A lesson learned is that deploying multiple models is a very useful approach.

Future work would include extending the number of models used, and some initial analysis of the data using SVM was carried out. SVM performance was better than NB, but worse than DT. We also put the data into IBM Watson which predicted "failure" and "higher" as being two highly significant attributes. Examining the performance of Random Forests would be a logical extension to the DT classifier used in this coursework.

Several different software systems were needed to explore the data and produce the charts and graphs — R, Tableau, Excel, MATLAB. A lesson learned was the importance of having a range of tools at your disposal. Future work would include developing further skill in all of these areas, and continuing to add knowledge of other software to our personal Data Science toolbox.

A further lesson learned was that it is important to consider and compare different performance measures for each model. Figure 12 shows the change in performance between accuracy and F1 as we benchmark, optimise and overfit the DT model. The measures change in the same general direction, but not in the same exact manner. [1] suggest that Receiver Operating Characteristic (ROC) curve as measurement of performance gives better results for NB and DT classifiers and future work would be to extend the number of performance measures collected. It is also suggested [1] that there is a performance difference between the C4.4 and C4.5 algorithm for DT classifiers and it would be interesting to do further work in that area.

References :

[1] J. Huang, J. Lu and C. Ling. Comparing Naïve Bayes, Decision Trees, and SVM with AUC and Accuracy. In *The Third IEEE International Conference on Data Mining 2003*

[2] R. Caruana, A. Niculescu-Mizil, An Empirical Comparison of Supervised Learning Algorithms. In *Proceedings of the 23rd International Conference on Machine Learning. 2006*

[3] I. Rish, J. Hellersteing, J. Thathachar, An analysis of data characteristics that affect naïve Bayes performance. *From IBM Watson Research Center, In IJCAI 2001 workshop on empirical methods in artificial intelligence*

[4] MATLAB Help library under Naïve Bayes Classification