A Comparison of Naïve Bayes and Random Forest Applied to the Adult Dataset

Adrian Ham and David Asboth

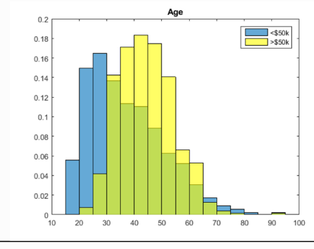
Description and motivation of the problem

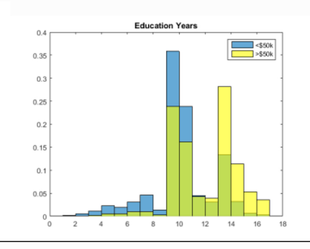
- Compare and contrast the performance of Naïve Bayes and Random Forest in a binary classification problem, predicting whether people earn more or less than $50,000 per year based on demographic data
- We will contrast our results to those obtained by a previous empirical study by Caruana & Nicolescu-Mizil (2006)[1].

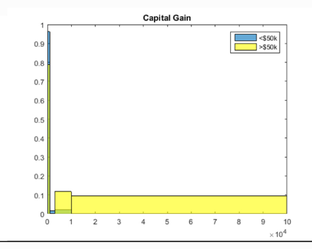
Initial analysis of the data set including basic statistics

- Dataset: Adult from UCI
- Training set: 5000 points, 25000 held out for final test
- The original dataset has 14 predictors - 6 numeric (ratio), 8 categorical (nominal)
- One duplicate and one redundant indicator were removed leaving 12 features
- Mean & standard deviation for the two classes for each numeric column were calculated (table opposite)
- Normalised histograms show noticeable difference in distributions of some variables between the two classes (e.g. age is left-skewed for <$50k class, but mostly normal for the >$50k class)
- Identified that CapGain and CapLoss were not good approximations of a normal distribution due to the predominant zero entries

Feature	mean<50k	mean>50k	std<50k	std>50k	skew<50k	skew>50k
Age	36.7	43.8	13.6	9.9	0.8	0.4
Education Yrs	9.7	11.5	2.4	2.4	-0.4	-0.3
Capital Gain ($)	146.8	3762.1	979.8	13914.0	16.6	6.1
Capital Loss ($)	60.1	207.5	324.5	604.5	5.4	2.6
Hours Worked	39.5	45.8	11.5	10.6	0.1	0.9







Two ML models with their pros and cons

Naïve Bayes (NB)

- Builds a model by obtaining probabilities of class membership for each predictor either based upon a normal distribution or categorical probability.
- A data point is assigned to a class by calculating the product of the probabilities for each predictor for each class and assigning to the class with the highest probability.

Pros

- Fast and scaling is linear with number of data points
- Understandable
- Surprisingly effective given simplicity

Cons

- Generally it predicts less accurately than many other models, such as Random Forest.
- Assumes independence between predictors (but surprisingly accurate even when predictors are not independent, potentially based upon the cancelling of dependencies according to Zhang, 2004[2]).
- Known to produce poor probabilities but empirically often gives the correct classification[3]

Random Forest (RF)

- Builds an ensemble of decision trees
- Each of them using a bagged subset of the training data
- Each tree uses a subset of features chosen randomly at each node to reduce correlation between trees.
- Each data point sent down each tree - a majority vote is taken on which class the point should belong to.

Pros

- Often outperforms other classification techniques in multiple domains e.g. detecting seismic activity[4] or detecting email forgery[5].
- Less likely to overfit to the data due to the random nature of each individual tree.
- Ranks features according to their importance, which some other techniques cannot.
- The algorithm is also parallelisable, making it potentially more feasible than other techniques on large datasets

Cons

- Some research has shown that training on larger datasets Random Forests can be computationally more expensive, even with parallelisation (Chen et al, 2012)[6]

Hypothesis Statement

- We expect both to produce worthwhile results; RF to produce more accurate results than the NB.
- This is due to the superior performance of RF over Naïve Bayes on the adult data set with 5000 training items reported by Caruana & Nicolescu-Mizil, 2006[1]
- They report a scorecard of 8 performance measures. NB was reported at 0.843 and RF as 0.930.
- Both models were significantly better than a random prediction but RF showed clear superior performance to NB.

Description of choice of training and evaluation methodology

- Train on 5000 data points to allow comparability with our reference paper.
- Within each model, we vary hyperparameters to find optimum values
- Use ten fold cross validation to estimate the generalisation error and provide information to seek to prevent overfitting.
- The evaluation criteria chosen was to minimise the error in accuracy of the prediction of the > $50k category

Choice of parameters and experimental results

Naïve Bayes

Parameters

- Used +1 smoothing
- Binned numeric variables by splitting into equal interval bins 10 different ways
- Manipulating the prior to measure effect on accuracy, recall and F Score

Main Experimental Results

The best model identified was with Capital Loss binned in 4 bins and the other continuous fields as normals.

Further Results

- Manipulating the prior showed that accuracy could be increased at the expense of recall and F1 Score.
- Modelled training vs. validation error for a number of different dataset to gain further insight into the risk of overfitting based on having a 'small' number of data points (i.e. empirically what appears to be 'small' for this dataset)

Accuracy Error

Naïve Bayes		Random Forest
23.6%	Train	12.8%
23.9%	Val.	18.3%
24.2%	**Final Test**	**18.5%**

Random Forests

Parameters

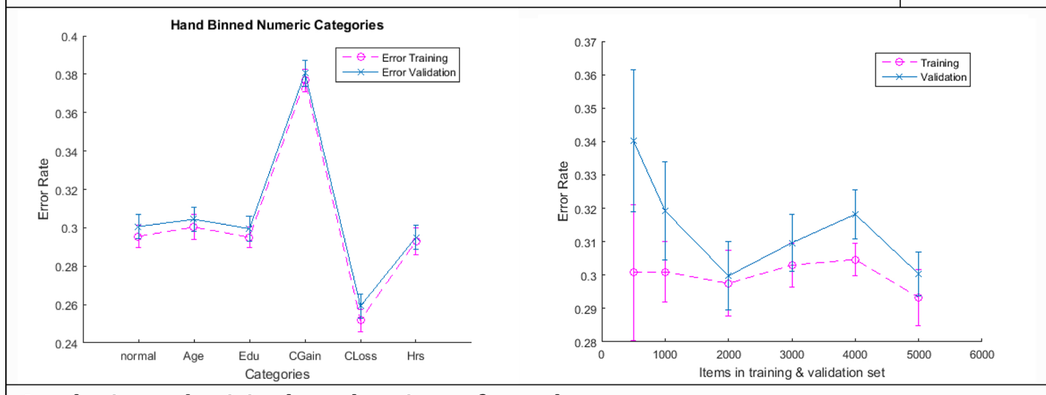
Used a grid search algorithm to optimise two hyperparameters, namely the number of trees in the forest and the number of features to sample at random at each node.
Breiman - fewer sampled features and more trees can yield the optimum hyperparameters

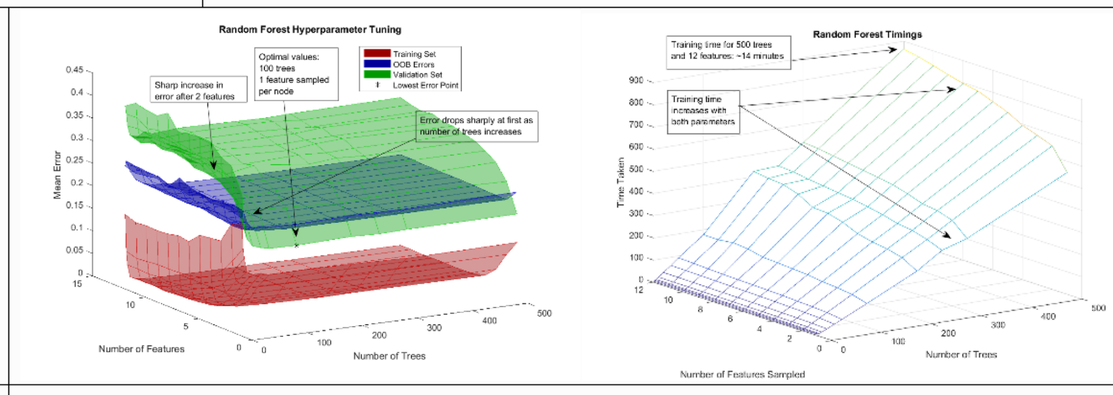
Main Experimental Results

The optimum value for hyperparameters was consistently shown to be around 100 trees, and only 1 feature sampled per node, which is consistent with Breiman's research.

Further Results

- Modelling the grid search as an error surface shows the error decreasing and converging as the number of trees increases, and as the number of features sampled decreases.
- Measuring the training time shows the high cost of more complex forests.
- Measuring feature importance showed that Capital Gain was the most important.





Analysis and critical evaluation of results

- The Naïve Bayes model could be considered to be a high bias, low variance model. In our results the high bias leads to a high error in validation set compared to RF. There is a very small difference in test/train error indicating the low variance.
- The power of RF could be described as its ability to manage the bias/variance trade off to produce strong results. This is achieved using various techniques such as bagging and random feature selection which seek to achieve low bias without undue variance. In our results there is higher variance for the random forest shown by an appreciable test/train error difference. However, the prediction error is much lower, the low bias far outweighing the variance.
- In terms of the optimality of the RF Breiman[7] describes generalization in terms of minimising c/s^2 where c = correlation between trees and s = the strength of the trees. Our best solution split on 1 feature. The implication of this is that where trees splitting on more than 1 feature were produced, the additional correlation between trees was higher than the gain in strength^2.
- The work on binning implied that the binning of the data had a significant effect on the accuracy of the model because small changes in binning did make significant changes to the predictions. It also implied information as to the importance of a given feature.
- Our results showed that RF can be more computationally involved – the longest training run on NB was c. 10 minutes whereas RF was c. 8 hours. However, given the size of the dataset, the number of features and the circumstances of the project this did not prove an obstacle.
- NB can overfit or exhibit high variance when training set is small and/or features are high. Our experimental result showed that for 500 training items there was a significant test/train error hence variance/overfitting but by 5000 items the variance was negligible.

Lessons learned and future work

- Optimising Naïve Bayes involves more data manipulation, whereas Random Forests lend themselves more to hyperparameter optimisation
- Future work on Naïve Bayes – binning based upon equal frequency and other approaches such as entropy which can yield significant performance improvements according to Dougherty et al, 1995[8].
- Future work on Random Forest – investigate tree depth as another hyperparameter and the impact of reducing the number of training features to see if excluding the less important ones yields better results.
- Investigate the combination of the two techniques, as Random Forest's ability to perform feature selection can be combined with the speed of Naïve Bayes (as shown by Chihab et al[9] and Lou et al.[10])

---

[1] Rich Caruana and Alexandru Niculescu-Mizil, 'An Empirical Comparison of Supervised Learning Algorithms', in *Proceedings of the 23rd International Conference on Machine Learning* (ACM, 2006), 161–68.
[2] Harry Zhang, 'The Optimality of Naive Bayes', *AA* 1, no. 2 (2004): 3.
[3] Manning, Raghavan, and Schutze, 'Introduction to Information Retrieval', n.d., Ch 13.
[4] Dong, L., Li, X., Xie, G., 2014. Nonlinear Methodologies for Identifying Seismic Event and Nuclear Explosion Using Random Forest, Support Vector Machine, and Naive Bayes Classification. Abstract and Applied Analysis.
[5] Abdallah, E.E., Otoom, A.F., ArwaSaqer, Abu-Aisheh, O., Omari, D., Salem, G., 2012. Detecting Email Forgery using Random Forests and Naïve Bayes Classifiers, in: Proceedings of World Academy of Science, Engineering and Technology. World Academy of Science, Engineering and Technology (WASET), Çanakkale, Italy, pp. 276–280.
[6] Chen, B., Sheridan, R.P., Hornak, V., Voigt, J.H., 2012. Comparison of Random Forest and Pipeline Pilot Naïve Bayes in Prospective QSAR Predictions. J. Chem. Inf. Model. 52, 792–803. doi:10.1021/ci200615h
[7] Leo Breiman, 'Random Forests', *Machine Learning* 45, no. 1 (2001): 5–32.
[8] James Dougherty, Ron Kohavi, and Mehran Sahami, 'Supervised and Unsupervised Discretization of Continuous Features', in *Machine Learning: Proceedings of the Twelfth International Conference*, vol. 12, 1995, 194–202.
[9] Chihab, Y., Ouhman, A.A., Erritali, M., El Ouahidi, B., n.d. Detection & Classification of Internet Intrusion Based on the Combination of Random Forest and Naïve Bayes.
[10] Lou, W., Wang, X., Chen, F., Chen, Y., Jiang, B., Zhang, H., 2014. Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes: e86703. PLoS One 9.