STAT3006 Assignment 3, 2020

Classification –

Weighting 30% - due 26/10/2020

This assignment involves constructing and assessing classifiers. We will first consider the now familiar Iris dataset, collected by Anderson (1936) and first statistically analysed by Fisher (1936), which requires one to solve the species problem, that is to predict which (known) species a given specimen belongs to.

The second dataset to be used to train classifiers is the Modified National Institute of Standards and Technology (MNIST) database of handwritten digits. This contains 70,000 images, of which 10,000 were reserved for testing. In each case you will use the given labelled data and attempt to construct classifiers which can accurately classify unlabelled observations.

You should select four classifiers, with at least one classifier based on a probability model, and one which is not, with each preferably mentioned in this course. Probability-based classifiers discussed include linear, quadratic, mixture and kernel density discriminant analysis. Classifiers discussed which are not based on a probability model include k nearest neighbours, classification trees, neural networks and support vector machines. All of these are implemented via various packages in R. If you wish to use a different method, please check with the coordinator. You cannot use a (classifier, dataset) combination that you have used or are using for an assignment in another course.

Tasks:

1. For one probability-based classifier and one non-probability-based classifier that you will use to answer later questions, describe the method mathematically. Include a mathematically-oriented overview of how you propose to choose all parameters of the method. [3 marks]

2. Apply one probability-based and one non-probability-based classifier to the Iris dataset using R, report the results and interpret them.

Results for each classifier should include the following:

a) characterisation of each class (including parameter estimates). [1 mark]

b) apparent error rates for each class and overall: obtained by training the classifier on the whole dataset and then applying it to the same dataset and checking error rates. [1 mark]

c) cross-validation (CV)-based estimates of the overall and class-specific error rates: obtained by training the classifier on a large fraction of the whole dataset and then applying it to the remaining data and checking error rates. You may use 5-fold, 10-fold or leave-one-out cross-validation to estimate performance, but you should give a statistical reason for your choice.

Also include an approximate 95% confidence interval for each error rate, along with a description of how this was obtained. [2 marks]

One option for error rate estimation is via the R package **ipred**, which contains a function called **errorest** (error rate estimation). With a bit of manipulation, errorest should be able to produce CV estimates for most classifiers. There are many alternative packages however.

d) plots of the predicted classes as they apply to the data and the data space, including visual representation of the decision boundaries, covering all unique pairs of explanatory variables. Repeat this with a set of zoomed out plots (wider ranges per axis) to check how new outliers would be handled – comment if you see anything surprising. [2 marks]

e) find, list and discuss any Iris observations which were misclassified in the apparent error rate and CV checks. [1 mark]

f) compare and contrast the decision boundaries between classes produced by the two methods and try to explain their shapes. Which method do you think was best for this dataset? Explain. Describe some aspects of either method that you think are appropriate or inappropriate for this classification problem. [2 marks]

g) You are now asked to predict the class of new observations collected from an area where the class proportions have changed to 0.2, 0.2 and 0.6 for setosa, virginica, and versicolor, respectively. Describe (with mathematical details) how you would change or refit each classifier to give it the best possible predictive error rate under these circumstances. Change or refit the classifier as necessary to do this and report point estimates of the new classifier parameters. Explain the nature of the changes. [3 marks]

Note: we will not compare our classification results with those of Fisher's 1936 paper "The Use of Multiple Measurements in Taxonomic Problems". However, it is worth reading that paper for background on the dataset and some of the aims of its analysis.

3. Choose two methods of classification that you have not used on the Iris dataset and apply them to the MNIST dataset – see http://yann.lecun.com/exdb/mnist/ . You are welcome to combine techniques if you wish. Leave the train and test split as it is, but feel free to use some of the training data to help choose a model, if desired. Aim for best possible predictive performance, but view this as primarily a learning exercise. I.e. you do not need to choose methods with the best performance. However, you should aim to get reasonable performance out of any method chosen, e.g. with reasonable choice of any hyperparameters. You should not pre-process the data in a way which makes use of your knowledge of the digit recognition problem, i.e. don't try to produce new explanatory variables which represent image features, even though this would likely help performance. You can use dimension reduction if you wish (e.g. PCA).

(a) Give a brief introduction to the dataset, including quantitative aspects. [1 mark]

(b) Give a summary of the predictive performance on the test set for each classifier. Make sure you do not use the test set at all before doing this. Include at least estimated overall error rate and class-specific error rates, along with approximate 95% confidence intervals for these. [2 marks]

(c) For each classifier, also report error rates as estimated using the training set. Attempt to explain any differences between the error rates estimated from the training and test sets. Note that reference to training and test sets here are to the labelling of the original data, not to how you may have used them. [2 marks]

(d) Explain why you chose each classifier type and describe some of their apparent strengths and weaknesses for this problem. [2 marks]

(e) For each classifier, show 1 example per digit (i.e. 20 in total) of handwritten digits which were classified into the correct class with the most certainty, and quantify what you mean by certainty. Explain why you think the classifiers were particularly successful at classifying these correctly and with certainty. [3 marks]

(f) For each classifier, show 1 example per digit (i.e. 20 in total) of the worst errors made by your classifier and quantify what you mean by worst. Explain why you think some of these errors may have been made by your classifier and been among the worst seen. [2 marks]

(g) What is the difference between a handwritten 7 and a 1 according to each classifier? Try to explain what each classifier is doing in this case, i.e. what are the main things the classifier considers to make this decision and how are they used? [3 marks]

Notes:

(i) Some R commands you might find useful:

objects() – gives the current list of objects in memory.

attributes(x) – gives the set of attributes of an object x.

(ii) For the Iris dataset, we will assume that each species was collected from an environment where all three are equally likely to be selected in a random sample. We can view the sample as representative and the prevalence of each species is similar in some environments. (See section VI of Fisher, 1936 for some details on how the observations were collected.)

(iii) Make it a habit to give reasons or justifications for decisions or statements.

(iv) Please put all the R commands in a separate file or files and submit these separately via a single text file or a zip file. You should not give any R commands in your main report and should not include any raw output – i.e. just include figures from R (each with a title, axis labels and caption below) and put any relevant numerical output in a table or within the text.

(v) As per http://www.uq.edu.au/myadvisor/academic-integrity-and-plagiarism, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. Use consistent notation throughout your assignment and define all of it.

(vi) Some references

R:

Maindonald, J. and Braun, J. *Data Analysis and Graphics Using R - An Example-Based Approach*, 3rd edition, Cambridge University Press, 2010.

Venables, W.N. and Ripley, B.D., *Modern Applied Statistics with S,* Fourth Edition, Springer, 2002.

Wickham, H. and Grolemund, G. *R for Data Science,* O'Reilly, 2017.

Classification and Clustering:

Bishop, C. *Pattern Recognition & Machine Learning*, Springer, 2006.

Devroye, L., Gyorfi, L. and Lugosi, G., *A Probabilistic Theory of Pattern Recognition*, Springer, 1996.

Duda, R.O., Hart, P.E. and Stork, D.G., *Pattern Classification*, Wiley, 2001.

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep Learning*, MIT Press

Hardle, W.K. and Simar, L., *Applied Multivariate Statistical Analysis*, 4th ed., Springer, 2015.

Hastie, T. and Tibshirani, R. Discriminant analyses by Gaussian mixtures, *Journal of the Royal Statistical Society B*, 8, 155-176. (MDA paper), 1996.

Hastie, T. and Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition, Springer, 2009.

McLachlan, G.J. and Peel, D. *Finite Mixture Models,* Wiley, 2000.

McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, 1992.

Scholkopf, B. and Smola, A. J. *Learning with Kernels*, MIT Press, 2001.

Data:

Anderson, E. The species problem in Iris, *Annals of the Missouri Botanical Garden* 23 (3): 457–509, 1936.

Fisher, R.A. The use of multiple measurements in taxonomic problems, *Annals of Eugenics* 7 (2): 179–188, 1936.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. Gradient-Based Learning Applied to Document Recognition, *Proceedings of the IEEE*, 86, 2278-2324, 1998.