STAT7305 Assignment 3 - Classification

Due: Friday 25/11/2024 by 5pm (extended from date listed in course profile)

Weighting: 30%

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). Here we will deal with the task of classification - predicting which (known) species a given specimen belongs to.

The second dataset to be used to train classifiers is the Fashion MNIST database. The original MNIST (Modified National Institute of Standards and Technology) database is very well known in machine learning and contains 70,000 images of handwritten digits, of which 10,000 were reserved for testing. The Fashion MNIST dataset was introduced in 2017 by Xiao *et al*. who were employed by an online retailer.

The stated aim was to provide a more challenging classification task than MNIST for testing classifier performance, while maintaining MNIST's desirable aspects including computational feasibility and visual interpretability. The Fashion MNIST database contains images of clothes and shoes and has nothing to do with NIST. However, it has exactly the same sample size and contains images of the same resolution (28*28 pixels) and same 8-bit grey scale (256 levels).

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. Classifiers discussed in the course which are based on a probability model include linear, quadratic, mixture and kernel density discriminant analysis and logistic regression. Classifiers discussed which are not based on a probability model include *k* nearest neighbours, classification trees, support vector machines, neural networks, random forests and boosted trees. All of these are implemented via various packages in R and Python. If you wish to use a different method, please check with the lecturer. You cannot use a (classifier, dataset) combination that you have used or are using for an assignment in another course.

Tasks:

1. Apply one probability-based and one non-probability-based classifier to the Iris dataset (with all 4 predictor dimensions) using R, report the results and interpret them.

Results for each classifier should include the following:

(a) Characterisation of each class as modeled by the classifier. Note that, where possible, this should include parameter estimates for each class. If this does not seem to characterise the classes, another attempt should be made to do so, such as via numerical or visual summaries of the observations which the classifier puts into each class. If the set of parameter estimates is too large for e.g. a page in the appendix, you can put it in a file and refer to that. [1 mark]

(b) 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]

(c) Find, list and discuss any Iris observations which were misclassified in the CV checks. [1 mark]

(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. Note: you do not need to derive these boundaries – they can emerge from a plot. [1 mark]

(e) 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]

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

2. Choose two methods of classification that you have not used on the Iris dataset and apply them to the Fashion MNIST dataset. 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 good predictive performance with a particular classifier type, i.e. make reasonable choices with respect to e.g. hyper-parameters or any basic pre-processing. However, view this as primarily a learning exercise where you get a chance to see what a particular type of classifier can do on a problem of substantial size and difficulty without much or any tuning. You should not pre-process the data in a way which makes use of any knowledge you have of the image 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 (e.g. numerical summaries). [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. Also include a standard confusion matrix showing counts and a scaled confusion matrix showing estimated rates specific to each class. [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 one example of from each class (i.e. 20 in total across the two classifiers) of images 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. Note: defining and implementing "certainty" may take some thought, creativity or research, maths and coding. [3 marks]

(f) For each classifier, show one example from each class (i.e. 20 in total across the two classifiers) 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 photos of pullovers and those of coats according to each trained classifier? Try to explain what each classifier is doing in this case, i.e. what are the main things that each classifier considers to make this decision and how are they used? [3 marks]

(h) Consider the scenario where a classifier will be trained on the Fashion MNIST dataset and then used by an online retailer to automatically classify new images of clothes and shoes with the camera setup, resolution and grey scale. Also consider the perspective of a potential customer who has some purchases in mind. Assume that each time they want to see products, they are presented with the ten classes as options, which will then limit what they see to just items from the ten predicted classes. Propose a misclassification cost matrix for this scenario and justify this. [2 marks]

(i) Choose any kind of classifier model and train this classifier using the proposed classification cost matrix on the Fashion MNIST training data. Report the resulting class specific error rates, average cost, standard and scaled (per observation from a class) confusion matrices, all as estimated on the test set. Compare these results with those reported for 2(b) and comment. [3 marks]

(j) Additionally consider a scenario where the class proportions are also different to the equal proportion case considered with Fashion MNIST. Specifically, assume that the new proportions are 0.01*(30,15,4,6,3,5,15,11,6,5) across classes 0-9. Research or invent two ways in which you could alter or refit the classifier considered in (i) to be optimal with respect to these changed class proportions. Explain these (with mathematical details) and their strengths and weaknesses. Compare the effect of these changes to changes in the cost matrix. [3 marks]

Notes:

- Your main response to these questions should consist of a single pdf file, submitted by the relevant link for this assignment on Blackboard. If you wish to use other software to help prepare this, that's fine, but the submitted file must be a pdf and contain all your answers. You should not include any raw output from software except figures and these should have a title, axis labels, a legend where appropriate, a caption, a figure number and be referenced by the figure number at least once in your report text. Any other output should be manually processed/selected before being included in e.g. text or tables.

- All your code and any supplementary files should be submitted via a separate .zip file to a second link for this assignment, also on Blackboard. Note that no code should be included in the .pdf file. All code should be written in R or Python and be readable via a text editor.

- Please name your files something like studentnumber_STAT7305_A3_report.pdf and studentnumber_STAT7305_A3_supp.zip to assist with marking.

- As per https://my.uq.edu.au/information-and-services/manage-my-program/student-integrity-and-conduct/academic-integrity-and-student-conduct ,
what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. Equations are either correct or not, but you should use consistent notation throughout your assignment and define all of it. The use of AI technologies to develop responses is strictly prohibited and may constitute misconduct under the Student Code of Conduct.

- 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.)

- Make it a habit to give reasons or justifications for decisions or statements.

References:

Anderson, E., "The problem of species in the northern blue flags, Iris Versicolor L. and Iris Virginica L.", *Annals of the Missouri Botanical Garden*, 15 (3): 241–332, 1928.

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

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

Devroye, L., Gyorfi, L. and Lugosi, G., *A Probabilistic Theory of Pattern Recognition*, Springer, 1996. https://www.szit.bme.hu/~gyorfi/pbook.pdf

Duda, R.O., Hart, P.E. and Stork, D.G., *Pattern Classification*, 2nd ed., Wiley, 2001. https://search.library.uq.edu.au/permalink/f/18av8c1/61UQ_ALMA2180308820003131

Fisher, R.A., "The Use of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics*, 1936.

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

Kuncheva, L. *Combining pattern classifiers: methods and algorithms* 2nd ed., Wiley, 2014. https://lucykuncheva.co.uk/Combining_Pattern_Classifiers_Methods_and_Algorithms_2nd_ed_Kuncheva%202014-09-09.pdf

Maindonald, J. and Braun, J. *Data Analysis and Graphics Using R - An Example-Based Approach*, 3rd ed., Cambridge University Press, 2010. https://search.library.uq.edu.au/permalink/f/12kerkf/61UQ_ALMA51301441400003131

McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, 1992. https://search.library.uq.edu.au/permalink/f/18av8c1/61UQ_ALMA51302650780003131

Schapire, R.E. and Freund, Y. *Boosting: Foundations and Algorithms*, MIT Press, 2012. https://direct.mit.edu/books/oa-monograph/5342/BoostingFoundations-and-Algorithms

Scholkopf, B. and Smola, A. J. *Learning with Kernels*, MIT Press, 2001. https://search.library.uq.edu.au/permalink/f/tbms52/TN_cdi_proquest_ebookcentral_EBC3338886

Venables, W.N. and Ripley, B.D., *Modern Applied Statistics with S*, 4th ed., Springer, 2002. https://search.library.uq.edu.au/permalink/f/tbms52/TN_cdi_askewsholts_vlebooks_9780387217062

Wickham, H. and Grolemund, G. *R for Data Science*, 2nd ed., O'Reilly, 2023. https://search.library.uq.edu.au/permalink/f/12kerkf/61UQ_ALMA51314916030003131

Xiao, H., Rasul, K., & Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017. *arXiv:1708.07747*. https://arxiv.org/abs/1708.07747