STAT3006 Assignment 3, 2023

Classification

30% - due 16/10/2023

This assignment inv[illegible]d assessing classifiers. We will do this using the Statlog vehicle silho[illegible]

The Statlog project ([illegible], 1994 and King *et al*., 1995) aimed to practically compare a wide range of classifiers, from statistics and machine learning, on a range of practical problems of relevance to industry. It was carried out over 1990-1993, funded by the European Community and coordinated by a researcher from Daimler-Benz (Nakhaeizadeh, 1995).

The vehicle silhouettes dataset was collected in 1987 by J.P. Siebert, to try to distinguish classes of 3D object based on 2D images and included in the collection analysed by the Statlog consortium. It consists of 18 quantitative variables based on the images of four "Corgi" model vehicles: a double-decker bus, a Chevrolet van, a Saab 9000 car and an Opel Manta 400 car. The task was to correctly classify observations based on these predictor variables into the four vehicle types. The bus and van were expected to be fairly easy to distinguish, but it was expected to be harder to separate the two types of car.

The dataset was first made available to the public at Strathclyde university and later via UCI's Machine Learning Repository https://archive.ics.uci.edu/. For reasons unknown, the data is no longer available from that site, but it is available from many other sources in many versions. The original dataset is available from the R package mlbench as Vehicle, and from blackboard as vehicle_orig.csv . Note that many other versions contain fewer observations, just 3 classes and missing data. They are not to be used for this assignment.

Your overall aim is to use the given labelled data to fit classifiers based on this data which can accurately classify unlabelled observations from the same population.

You need to use four different classifier types, one of which must be *k* nearest neighbours. You can select the other classifier types, given the conditions below. At least one of the remaining three classifiers must be based on a probability model. Each classifier has preferably been 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. In addition to *k* nearest neighbours, classifiers discussed which are not based on a probability model include classification trees, support vector machines and neural networks.

All of these classifier types along with fitting methods are implemented via various packages in R. If you wish to use Python instead, or a mix of both, that's ok. We are better able to support use of R. You may find that R has better packages for statistical methods and Python has an edge in machine learning methods. If you want to use a classification method not

listed above, please check with the lecturer (via Ed discussion board would be good). If you have used the vehicle dataset for classification in another course, you must use different classifier types here.

Aim for good predic t view this as primarily a learning exercise. You should not pre-proce which makes use of any knowledge you have of the problem. You can u on if you wish (e.g. principal component analysis) but most methods w from it.

You will need to sul ument and a file (e.g. .zip) containing all of your code to two separate links on Blackboard.

Tasks:

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

Apply all four classifiers to the Vehicle dataset, report the results and interpret them.

Results for each classifier should include the following:

(b) Characterisation of each class as modeled by the classifier. Note that, where possible, this should include parameter estimates for each class. Where not possible, another attempt should be made to characterise the class, 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 along with the code and refer to that. [2 marks]

(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 mathematical description and justification of how this was obtained.

For k nearest neighbours classification, use nested cross-validation (cv) to both select an optimal value of k and evaluate the classifier with this choice. Also perform standard cv to choose a single optimal value of k that could be used in a "final classifier" (see lecture notes including p34 and p38). Detail and explain what has been observed re: optimal values for k across both types of cv.

Consider whether or not you might like to transform and/or normalise any of the predictor variables to try to aid a particular classifier type. Explain why you either did or did not do this.

[5 marks]

(d) Find, list (by index in an appendix) and discuss observations which were misclassified by each classifier, with particular consideration of observations misclassified by multiple classifiers. [1 mark]

(e) For each classifi  rank all the predictor variables. You can do this either by researchin ing one or developing your own method. In either case, mathematicall l and explain why it may be reasonable or useful. Then use it to rank a bles for each classifier and choose the top two from each. [2 marks]

(f) Use the top two predictor variables found in (d) to plot 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.

Re: the top two predictor variables. Your classifier will require values for all 18 predictor variables, so you need to provide values for the other 16 variables. I suggest you fix these equal to their means for the dataset, hence producing a slice through the higher-dimensional space. Note: you do not need to derive the decision boundaries – they can emerge from the plot.

Select some of the data to add to your plot (via e.g. nearness to the slice or some form of projection), explain how and why you chose these points and comment on the resulting plot.

[3 marks]

(g) Compare and contrast the decision boundaries between classes produced by the different classifiers and try to explain their shapes. Which method do you think was best for this dataset? Explain. Describe some aspects of each method that you think are well-suited or not so well-suited for this classification problem. [2 marks]

(h) Fisher's linear discriminant analysis can be extended to multiple classes. This typically results in G-1 projections of the data, where G is the number of classes. This can be viewed as a form of supervised dimensionality reduction.

Mathematically explain how this set of projections is determined and why there are at most G-1 of them. Also determine the first two linear discriminant projection vectors for this dataset. [2 marks]

(i) The linear discriminant transformation is generally far from reversible due to the level of dimension reduction. Hence it would be difficult to plot a meaningful representation of any of your classifiers in this space. Instead, choose the classifier type that performed best and complete the following. Use cross-validation to train and assess this classifier in the space of the top two linear discriminant projections only. However, the discriminant projections are determined using the labels, so they will need to be re-calculated in each training step before training and testing. Note that if you choose a k nearest neighbour classifier, use the best k value from (b). Report the estimated overall and class-specific error rates. [2 marks]

(j) Determine the first two linear discriminant projections of the whole dataset and retrain this classifier on just this projected data. Then produce a plot of the decision boundaries, similar to that in (g), but also including all the data (projected to the first two LDs). Comment on this plot, including comparison to the plot for the same classifier type in (e). [2 marks]

(k) You are now ask[illegible]s of new observations collected under different conditions where th[illegible]ave changed to 0.4, 0.3, 0.2 and 0.1 for the classes: saab, opel, van and b[illegible]wever, you only have the existing Vehicle dataset.

Describe (with math[illegible]v you would change or refit each classifier to give it the lowest possible expected error rate under these circumstances. There is no need to carry this out. Explain the nature of the changes. [2 marks]

(l) For each classifier type, determine a measure of "certainty" about the predicted classification that can be applied to any new observation. This may take some thought, creativity, mathematics and coding. [2 marks]

(m) For each classifier, find 1 example per class (i.e. 4 in total) of observations which were classified into the correct class with the most certainty. Explain why you think the classifiers were particularly successful at classifying these correctly and with certainty. [1 mark]

(n) For each classifier, show 1 example per class (i.e. 4 in total) of the worst errors made by your classifier and quantify what you mean by worst, making use of your certainty function. Explain why you think some of these errors may have been made by your classifier. [1 mark]

(o) What is the difference between a Saab and an Opel (of the kinds seen in the dataset) 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? [2 marks]

Notes:

(i) Some R commands you might find useful:

objects() – gives the [illegible]s in memory.

attributes(x) – gives [illegible]of an object x.

(ii) Make it a habit t[illegible]tifications for decisions or statements.

(iii) Please put all th[illegible]eparate file or files and submit these separately via a single text file or a zip [illegible]. Do 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 summarise in the text.

(iv) Please name your files something like student_number_STAT3006_A3.pdf and student_number_STAT3006_A3.zip to assist with marking.

(v) 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. 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, a* ition, Springer, 2009.

McLachlan, G.J. an *ture Models*, Wiley, 2000.

McLachlan, G.J. *Di* *nd Statistical Pattern Recognition*, Wiley, 1992.

Scholkopf, B. and S *with Kernels*, MIT Press, 2001.

Statlog and the Vehicle Silhouette Data:

King, R. D., Feng, C. and Sutherland, A. Statlog: comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence*, 9(3), 289-333, 1995.

Michie, D., Spiegelhalter, D. J., and Taylor, C. C. *Machine learning, neural and statistical classification*, 1994.

Nakhaeizadeh, G. What Daimler-Benz has learned as an industrial partner from the machine learning project Statlog. *Proc. of the Workshop on Applying Machine Learning in Practice*. 1995.