MAST90138 – Semester 2, 2023. Assignment #3

Instructions:

- This assignment contains 3 problems with a total of 19 +1 marks which count towards 15% of the final mark for the subject. It is due by 7pm Sunday 22 October, 2023.
- **[1 mark]** Your assignment should clearly show your name and student ID number, your tutor's name and the time and day of your tutorial class. Your answers must be clearly numbered and in the same order as the assignment questions (this includes R outputs, R code, figures and/or tables). Your answers must be easy to read (marks may be deducted for illegible handwriting). Include all of your working out in your answers. All R outputs, including graphs and tables, must be accompanied by your concise and clearly written R code used to produce it. Any graph, table or R code must be accompanied by clear and concise comments. Uncommented R code is not acceptable.
- Use concise text explanations to support your answers. Comments should be brief and concise: marks will be awarded for clarity.
- No late submission is allowed.
- Your lecturer may not help you directly with assignment questions, but may provide some appropriate guidance.

Data: The data come from Australian weather stations located in either the North or the South of the country. These weather stations have been separated in a training set available in the file `XGtrainRain.txt` which contains data on $ntrain = 150$ weather stations, and a test set available in the file `XGtestRain.txt` which contains data on $ntest = 41$ weather stations. In both files, each row corresponds to a different weather station and contains 366 values: the first 365 values are $p = 365$ explanatory variables $X_1, \ldots, X_p$; the last value is a class label ($G = 0$ if the weather station is located in the North, 1 if it is located in the South). For $j = 1, \ldots, p$, X_j is the amount of rainfall at day j of a year.

Throughout we use $\mathbf{X} = (X_1, \ldots, X_p)$ for a generic p-vector (with $p = 365$) whose components are the p explanatory variables. We also use `Xtest` to denote the 41 $\mathbf{X}_i$'s from the test data, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ is the vector of p explanatory variables for the ith individual from the test data.

The datasets are available on the LMS web page within the *Assessment* menu. You may need to manipulate the data format (data frames or matrices) depending on the task.

Questions [19 marks]

1. **[6 marks]**
   (a) If we do not apply any regularisation:
      (a1) how many parameters does the quadratic discriminant (QD) classifier need to estimate in order to predict the class label ($G = 0$ or 1) from $\mathbf{X}$? Justify your answer.
      (a2) how many parameters does the logistic classifier need to estimate in order to predict the class label ($G = 0$ or 1) from $\mathbf{X}$? Justify your answer.
   (b) Using the training data in `XGtrainRain.txt` and the R functions `qda` from the `MASS` package and `glm`, without applying any regularisation, construct:
      (b1) the QD classifier that predicts the class label (0 or 1) from $\mathbf{X}$ – what happens and why?
      (b2) the logistic classifier that predicts the class label (0 or 1) from $\mathbf{X}$, setting to 100 the maximum number of iterations of the function `glm`. Examine the output (but exceptionally, DO NOT show the output in your solution file as it is too long), and explain what is wrong with the estimated coefficients and why.
2. **[3 marks]** Using the training data in `XGtrainRain.txt` and the `glm` R function:

(a) construct the logistic classifier that predicts the class label (0 or 1) from the first q principal components of $\mathbf{X}$ as predictor, with $q \in [1, 30]$ selected by leave-one-out CV for classification (do not use any built-in function to compute CV; instead write your own code). Show the R outputs.

(b) Apply the classifier to the test data `Xtest`, and compute the classification error of the test data.

3. **[10 marks]** Using the training data in `XGtrainRain.text`:

(a) With the `randomForest` package in R, construct a random forest classifier using $\mathbf{X}$ as predictor with $B = 5000$ trees; for m (the number of random candidate variables for each split), use the value m_{OOB} that minimises the out-of-bag (OOB) classification error wrt m, on the following discrete grid of 10 candidate m values: $\{[c/4], [c/2], [c], [2c], [3c], [4c], [5c], [6c], [7c], [8c]\}$, where $c = \sqrt{365}$ and for any x, $[x]$ denotes the integer the closest to x. If you find several values of m on that grid that give the same OOB classification error, take the smallest of those values of m.

(b) For the random forest classifier you fitted in (a) with $m = m_{OOB}$ and $B = 5000$:

- plot the Gini importance of the X_j's versus j, for $j = 1, \ldots, 365$.
- Using that plot, identify the X_j's that are the most important for classification by the forest and explain (give an interpretation) why it makes sense that those particular X_j's play a big role in classification of these particular rainfall data.
- Apply the classifier to the test data `Xtest`, and compute the classification error of the test data.

(c) Using the R packages `pls` and `rpart`, compute the partial least squares (PLS) components of the training data and construct a single tree classifier based on the first 50 PLS components as predictor variables (you can keep the default values of the function that builds the tree). Then

- Plot the fitted tree using the function `rpart.plot` from the package `rpart.plot`. Which PLS components play a role in the tree? How are they related to the variables that are the most important for classification by the random forest in question (b)? Hint: for each PLS component used by the tree, plot its correlation with X_j versus j, for $j = 1, \ldots, 365$ (you must show this plot).
- Apply the fitted classifier to the test data `Xtest`, and compute the classification error of the test data.