STAT 318/462: Data Mining
Assignment 3
Due Date: 11-59pm, 25th October, 2021

Your assignment must be submitted electronically to the STAT318/462 Learn page, under Assignment 3.

You may do the assignment by yourself or with one other person from the same cohort (300-level students cannot work with 400-level students). If you hand in a joint assignment, you will each be given the same mark. Marks will be lost for unexplained, poorly presented and incomplete answers. Whenever you are asked to do computations with data, feel free to do them any way that is convenient. If you use *R* (recommended), please provide your code. All figures and plots must be clearly labelled.

1. **(10 marks)** In this question, you will work with one of the "classical" datasets in data mining - *Iris* and a package **Caret**.

   (a) Examine the *Iris* dataset from the *caret* package. Describe the size of the dataset and all the variables in it.

   (b) Visualise features in the *Iris* dataset via box and whisker plots of each feature.

   (c) Assume a classification setting where *Species* is the target variable and all the other variables are predictors. Examine the number of the data points per class. If we would like to build a classifier to predict species of iris, would we deal with imbalanced data? Explain your reasoning.

   (d) Using *caret* function `featurePlot()`, construct and describe overlaying probability density plots of class distribution for each feature - 4 plots.

   (e) Assume classification problem where *Species* is the target variable and all the other variables are predictors. Prepare dataset for cross-validation using function `trainControl()` in the **Caret** package. Using function `train()` with accuracy as metric perform 10-fold cross-validation for LDA and KNN models and compare and discuss results of these two models. Hint: The following code could be of help:

```
results <- resamples(list(lda=fit.lda, knn=fit.knn))
summary(results)
dotplot(results)
```

2. **(12 marks)** In this question, you will utilise knowledge about bootstrap in order to analyse data from *Credit* dataset.

   (a) Examine a Credit dataset from the ICLR2 package (please note that you would need to install this package in R first). Describe the size of the dataset and all the variables in it.

   (b) Based on this data set, provide an estimate for the population mean of the credit card balance. Let's denote it $\hat{\mu}$.

(c) Provide an estimate of the standard error of $\hat{\mu}$. Interpret your result. Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

(d) Estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (c)?

(e) Based on your bootstrap estimate from (d), provide a 95% confidence interval for the mean of the credit card balance. Compare it to the results obtained using t-test. Hint: You can approximate a 95% confidence interval using the formula $[\hat{\mu} - 2 * SE(\hat{\mu}), \hat{\mu} + 2 * SE(\hat{\mu})]$.

3. **(10 marks)** In this question, you will fit regression trees to predict *sales* using the Carseats data. This dataset has been divided into training and testing sets: `carseatsTrain.csv` and `carseatsTest.csv` (download these sets from Learn). Use the `tree(), randomForest()` and `gbm()` *R* functions to answer this question (see Section 8.3 of the course textbook).

(a) Fit a regression tree to the training set (do not prune the tree). Plot the tree and interpret the results. What are the test and training MSEs for your tree?

(b) Use the `cv.tree()` *R* function to prune your tree (use your judgement here). Does the pruned tree perform better?

(c) Fit a bagged regression tree and a random forest to the training set. What are the test and training MSEs for each model? Was decorrelating trees an effective strategy for this problem?

(d) Fit a boosted regression tree to the training set. Experiment with different tree depths, shrinkage parameters and the number of trees. What are the test and training MSEs for your best tree? Comment on your results.

(e) Which model performed best and which predictors were the most important in this model?

Question 4 is for students taking STAT462. STAT318 students will NOT receive additional credit if they choose to answer this question. This is an independent research question

4. **(5 marks)** In this question, you will investigate scientific literature using a database called Google Scholar. Identify 3 scientific papers published not earlier than 2019 which use one of the classification algorithms discussed in this course so far. Write a short summary overview of the three papers highlighting algorithms which were utilised and the major results of the study. The limit is half of the page.