# STAT 318/462: Data Mining
# Assignment 3
# Due Date: 3pm, 15th October, 2018

**Your printed assignment must be submitted in the STAT318/462 assignment box on the fourth floor of the Erskine building (by MATH/STAT reception).**

You can work in pairs on this assignment and submit a single co-authored set of answers. 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. **(6 marks)** In this question, you will build CART trees for a two-class classification problem using the $n = 100$ training observations in Table 1. This table summarizes a data set with three binary-valued (0 or 1) features, $X_1, X_2$, and $X_3$, with two class labels *High* or *Low*. For example, there are 5 *low* observations of the form (0,0,1) in the data. In the questions that follow, use Gini index for measuring impurity and consider splits of the form $X_i < 0.5$ where $i = 1, 2, 3$.

| $X_1$ | $X_2$ | $X_3$ | # *High* observations | # *Low* observations |
|---|---|---|---|---|
| 1 | 1 | 1 | 5 | 0 |
| 0 | 1 | 1 | 0 | 20 |
| 1 | 0 | 1 | 20 | 0 |
| 0 | 0 | 1 | 0 | 5 |
| 0 | 1 | 0 | 25 | 0 |
| 0 | 0 | 0 | 0 | 25 |

Table 1: Training data for Question 4.

(a) Find the best split in the training data.

(b) Sketch the first split of your tree, showing the number of *High* and *Low* observations at each daughter node.

(c) Find the best split for each impure daughter node found in part (b). Sketch the tree you obtain.

(d) How many training observations are misclassified in your tree?

(e) Build a tree in the following way. First, split the root node using $X_3$. Then, split the daughter nodes using the best split. Sketch the tree you obtain.

(f) Use parts (d) and (e) to conclude about the greedy nature of CART.

2. **(8 marks)** In this question, you will fit regression trees to predict *sales* using the Carseats data. This data has been divided into training and testing sets: `carTrain.csv` and `carTest.csv` (download these sets from Learn). Use the `tree()` 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 see if pruning will improve your tree's performance. Does pruning improve the test MSE?

   (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 MSE for your best tree?

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

3. **(6 marks)** This question considers clustering the A3data1 and A3data2 data (both sets are on the Learn page). **When clustering these datasets, use $x_1$ and $x_2$.** The third variable 'Cluster' is the actual cluster label of each point and should not be used for clustering. This variable is included so that you plot the clusters and assess the performance of each clustering method.

   (a) Use the A3data1 data to answer the following questions.

      i. Perform $k$-means clustering using $k = 5$. Plot the clustering using a different colour for each cluster.

      ii. Using hierarchical clustering with complete linkage and Euclidean distance, cluster the data and plot the clustering with 5 clusters. Repeat using single linkage.

      iii. Compare the clustering methods from parts i. and ii. (using 5 clusters) with the actual cluster labels using the method described on page 412-413 of the course textbook. **Comment on your results.**

   (b) Repeat part (a) using the A3data2 data, but use 3 clusters (not 5).

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

4. **(4 marks)** In this question, you will fit support vector machines to the Banknote data from Assignment 2 (on the Learn page). **Only use the predictors $x_1$ and $x_3$ to fit your classifiers.** You will find Section 9.6 of the course textbook useful.

   (a) Is it possible to find a separating hyperplane for the training data? **Explain.**

   (b) Fit a support vector classifier to the training data using `tune()` to find the best `cost` value. Plot the best classifier and produce a confusion matrix for the testing data. **Comment on your results.**

   (c) Fit a support vector machine (SVM) to the training data using the radial kernel. Use `tune()` to find the best `cost` and `gamma` values. Plot the best SVM and produce a confusion matrix for the testing data. Compare your results with those obtained in part (b).