STAT 318/462: Data Mining
Assignment 1
Due Date: 3pm, 15th August, 2019

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 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. **(4 marks)** Describe **one** advantage and **one** disadvantage of flexible (verses a less flexible) approaches for regression. Under what conditions might a less flexible approach be preferred?

2. **(6 marks)** Consider a binary classification problem $Y \in \{0, 1\}$ with one predictor X. The prior probability of being in class 0 is $\Pr(Y = 0) = \pi_0 = 0.69$ and the density function for X in class 0 is a standard normal

$$f_0(x) = \text{Normal}(0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

   The density function for X in class 1 is also normal, but with $\mu = 1$ and $\sigma^2 = 0.5$

$$f_1(x) = \text{Normal}(1, 0.5) = \frac{1}{\sqrt{\pi}} \exp\left(-(x-1)^2\right).$$

   (a) Plot $\pi_0 f_0(x)$ and $\pi_1 f_1(x)$ in the same figure.
   (b) Find the Bayes decision boundary (*Hint:* $\pi_0 f_0(x) = \pi_1 f_1(x)$ on the boundary).
   (c) Using Bayes classifier, classify the observation $X = 3$. **Justify your prediction.**
   (d) What is the probability that an observation with $X = 2$ is in class 1?

3. (**8 marks**) In this question, you will fit kNN regression models to the `Auto` data set to predict $Y =$ `mpg` using $X =$ `weight`. This data has been divided into training and testing sets: `AutoTrain.csv` and `AutoTest.csv` (download these sets from Learn). The `kNN()` *R* function on Learn should be used to answer this question (*you need to run the* `kNN` *code before calling the function*).

   (a) Perform kNN regression with $k = 2, 5, 10, 20, 30, 50$ and 100, (learning from the **training data**) and compute the **training** and **testing MSE** for each value of k.
   (b) Which value of k performed best? **Explain.**
   (c) Plot the best kNN model and all the data (use different colours for the training and testing sets) in the same figure. **Comment about your results.** (*the* `points()` *function could be useful for plotting kNN because it is discontinuous.*)
   (d) Comment on the bias-variance trade-off when defining a neighbourhood for kNN regression.