STAT 318/462: Data Mining
Assignment 1
Due Date: 3pm, 15th August, 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. **(3 marks)** Describe one advantage and one disadvantage of less flexible (verses a flexible) approach for supervised learning? Under what conditions might a flexible approach be preferred?

2. **(2 marks)** Consider a classification problem for which the Bayes decision boundary is approximately linear. If we perform k-nearest neighbour classification, would we expect a lower testing error rate for a relatively small or relatively large value of k? Why?

3. **(2 marks)** Show that the expected test MSE, for a given $\mathbf{x}_0$, can be decomposed into the sum of three fundamental quantities, the variance of $\hat{f}(\mathbf{x}_0)$, the squared bias of $\hat{f}(\mathbf{x}_0)$ and the variance of the error terms ϵ. That is

$$E[y_0 - \hat{f}(\mathbf{x}_0)]^2 = V(\hat{f}(\mathbf{x}_0)) + [\text{Bias}(\hat{f}(\mathbf{x}_0))]^2 + V(\epsilon).$$

   You may use the following properties without proof:

$$E(a) = a; \quad E(aX) = aE(X); \quad \text{and} \quad E(X+Y) = E(X) + E(Y),$$

   where a is a number and X and Y are random variables.

4. **(5 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

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

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

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

   (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, what class is assigned to an observation with $X = 3$?

   (d) What is the probability that an observation with $X = 3$ is in class 1?

5. (**8 marks**) In this question, you will fit two regression models to the `Wage` data set to predict $Y =$ `wage` using $X =$ `age`. This data has been divided into training and testing sets: `WageTrain.csv` and `WageTest.csv` (download these sets from Learn). The `kNN()` *R* function on Learn should be used to perform k-nearest neighbour (kNN) regression (you need to run the `kNN` code before calling it).

   (a) Perform kNN regression with $k = 1, 5, 10, 20, 50, 75, 100, 150, 200$ and 300, learning from the training data. Plot the **testing MSE** for each value of k against $1/k$. Which value of k performed best? **Explain.**

   (b) Loess is a local smoothing function in R (use ?`loess` in R for details on how to fit the model and how to make predictions). Fit the loess model (using default settings or user specified) to the training data and calculate the **testing MSE**. To make predictions outside the range of the training data (extrapolation) you need to include the following `control` setting when fitting the model:

   ```
   > loess( ..., control=loess.control(surface="direct"))
   ```

   (c) Plot the best kNN model, the loess model, and the full data set in the same figure. Comment on the two regression models. (*Hint: the* `points()` *function could be useful for plotting kNN because it is discontinuous.*)

   (d) These regression methods are local methods that require neighbourhood definitions. Comment on the bias-variance trade-off when defining a neighbourhood for kNN regression.