STAT 318 Data Mining
Assignment 1
Due Date: 4pm, 25 March, 2024

Please submit your assignment as a single pdf on Learn. Try to give answers that are as short as possible and as long as necessary. Marks will be lost for unexplained, poorly presented and incomplete answers. Whenever you are asked to do computations with data, please provide your R code. All figures and plots must be clearly labelled.

1. **(6 marks)** We evaluate the accuracy of a regression.

   (a) Define the **testing mean-squared error** and **training mean-squared error**. Explain the relevant difference between the two definitions.

   (b) Assume that we use linear and polynomial regression. Describe the general behaviour of testing mean-squared error and training mean-squared error when we move from linear regression through and low polynomial degrees to higher polynomial degrees.

   (c) What is the process of finding the best polynomial degree with respect to the testing mean-squared error called?

2. (**8 marks**) In this question, you will fit kNN regression models to the `Auto` data set to predict $Y =$ `mpg` using $X =$ `horsepower`. 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 training data, testing data and the best kNN model in the same figure. (*The* `points()` *function is useful to plot the kNN model because it is discontinuous.*)

   (d) Describe the bias-variance trade-off for kNN regression.

3. **(6 marks)** Consider a binary classification problem $Y \in \{0, 1\}$ with one predictor X. The prior probability of being in class 0 is $\Pr(Y = 0) = \pi_0 = 0.5$ and the density function for X in class 0 is a standard normal

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

   The distribution for X in class 1 is the exponential distribution with parameter $\lambda = 1/2$. The density function is

   $$f_1(x) = \text{Exp}(x; 0.5) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2}\exp\left(-\frac{1}{2}x\right) & \text{if } x \geq 0. \end{cases}$$

   (a) Plot $\pi_0 f_0(x)$ and $\pi_1 f_1(x)$ in the same figure.

(b) Find the Bayes decision boundary (*Hint:* $\pi_0 f_0(x) \leq \pi_1 f_1(x)$ on one side of the boundary and $\pi_0 f_0(x) \geq \pi_1 f_1(x)$ on the other side of the boundary).

(c) Using Bayes classifier, classify the observation $X = 3$. **Justify your prediction.**

(d) What is the proportion of class 0 elements that will get a false classification, and what is the proportion of class 1 elements that will get a false classification? Explain your reasoning.