BMTRY 790, Spring 2023: Assignment 1

Exercises

1. Recall the k-NN approach for classifying an observation $\mathbf{x}$ based on the average class of the k data points nearest to $\mathbf{x}$.

    i. Write psuedo code for a function to fit a k-NN classifier for a case where the outcome has 2 classes.

    ii. How would you extend a k-NN classifier for the case where the outcome has more than 2 classes (i.e. what would change?).

    iii The k-NN classifier discussed in class took a simple average of the k closest points to make a prediction of class. Describe an alternative to a simple average and reasons why you feel it would or wouldn't improve classification?

2. We often need to simulate data under specific distributional conditions in order to compare the performance of different statistical methods. The adta for the two class problem discussed in class were simulated in R according to the following rules. First 10 values, m_1, were generated from a bivariate normal distribution, $N\left((1,-1),\mathbf{I}\right)$ and labeled as class 1 (i.e. $Y=1$). Similarly, 10 values, m_0, were generated from a bivariate normal distribution, $N\left((-1,1),\mathbf{I}\right)$ and labeled as class 0 (i.e. $Y=0$). The for each class, 100 observations were generated from each of the classes as follows: for each observation, a mean ($m_k = m_1$ or m_0) from among the 10 values was selected with probability $\frac{1}{10}$, and then an observation was generated as $N\left(m_k,\mathbf{I}\right)$ resulting in a mixture of normals for each class.

    i. Write a function to generate data as described above.

    ii. Fit a linear and logistic classifier to data with 100 observations per class. Construct a plot with each class labeled by a different color and the decision boundary for each classifier.