DSM Lab 2: KNN, Logistic Regression and LDA

Exercise 1

In this question you will produce a picture like Figure 2.2 in the book Elements of Statistical Learning on pp.15, but for a different dataset. Try to understand first the code for this Figure by running the posted file **mixture.R**, where the generated dataset for Figure 2.2 and the code is given.

a) Generate a dataset consisting of n = 100 observations from the logit model below:

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-\beta_1 x_1 - \beta_2 x_2)} \tag{1}$$

with $x_1 \sim iid\, \mathcal{N}(0, 1)$, $x_2 \sim iid\, \mathcal{N}(0, 4)$, $\beta_1 = 2$ and $\beta_2 = 3$. Here $x_1 \perp x_2$ but you don't need to impose this in the analysis.

Hint: use **rnorm()** for generating normally distributed variables, calculate $P(y = 1 \mid x)$. To generate y for each x, draw from the binomial distribution with probability $P(y = 1 \mid x)$. Use the command **rbinom()** for binomial draws. Use **set.seed()** to obtain the same results when repeating the code.

b) Plot the data in the two dimensions x_1 and x_2, using orange and blue circles for the two classes in y.

Hint: set the color palette with:
col.list = c("blue","orange")
palette(col.list)

c) The Bayes decision boundary are the points x_1 and x_2 such that $P(y = 1 \mid x) = 0.5$. For each x_1 in the simulated (or training) data, calculate x_2 such that $P(y = 1 \mid x) = 0.5$ and add the Bayes decision boundary on the plot in b) using a dashed purple line. Is this boundary linear and can you find the exact formula for it? Explain.

Hint: the Bayes decision boundary is linear in this case.

d) Construct a test set on a grid of $g = 50$ values for x_1 and x_2, ranging from their minimum to their maximum. Generate a test set from each combination of x_1 and x_2, and call it test. Gather the training set for x into a data frame called train.

e) Run a KNN analysis with $k = 3$ nearest neighbors on the test data using the training data and the realizations of y. Use the command **knn()**.

f) Following the **mixture.R** code, plot the training data with circles, the test data with dots, each with the color blue or orange according to which class they either belong to (in the training data) or to which class they were assigned to (in the test data). Add the KNN decision boundary to the plot using contour() and the Bayes decision boundary.

g) Repeat e) and f) with $k = 10$ and $k = 15$. Calculate the test error rate for each of $k = 3; 10; 15$, and the Bayes decision boundary. Which k gets closest to the Bayes decision boundary? Explain why this makes sense or not.

Exercise 2

Solve Question 10 in Chapter 4 of the book Intro to Statistical Learning in Edition 1 (Question 13 in Edition 2).