MATH3021 Assignment 2, due on Sept 20

(1) Generate 50 samples from each of the two distributions: $N(\mu_1, \mathbf{I}_2)$ and $N(\mu_2, \mathbf{I}_2)$, where $\mu_1 = [1, 1]^\top$, $\mu_1 = [2, 2]^\top$ and $\mathbf{I}_2$ is the 2-by-2 identity matrix. Let the labels for the first sample to be 1 and labels for the second sample to be -1. Combine two samples to get a new sample:

$$\{(X_{1i}, X_{2i}, Y_i), i = 1, \cdots, 100\}.$$

- Derive the true decision boundary;
- Use KNN to learning the decision boundary. You may consider many grid points in the range of samples, and then draw the decision boundary. Consider three different choices of K.

(2) Generate $X_1, \cdots, X_5$ from the standard normal distribution $N(0, 1)$ and ϵ from $N(0, 0.5)$, and then let

$$Y = 2 + 3X_1 + X_2 + 2X_3 + \epsilon.$$

Considering a dataset $\{(Y_i, X_{1i}, \cdots, X_{5i}), i = 1, \cdots, n\}$ with sample size $n = 100$, find the best model by using best subset selection, forward selection, and backward selection approaches. You may consider using 70 samples as the training set and 30 samples as the test set, and when considering the p-value, you may set a threshold that you think appropriate.

Submit your results and R code. Remember to use the command *set.seed*() **so that your results can be reproduced.**