Tutorial 5

```
# create positive class sample with 2 descriptive features
set.seed(3)
f1 <- rnorm(100, mean=6, sd = 1.2)
set.seed(4)
f2 <- rnorm(100, mean=6, sd = 1.2)
P.data <- cbind(f1, f2)

# create positive class sample with 2 descriptive features
set.seed(7)
f1 <- rnorm(300, mean=4, sd = 1.2)
set.seed(8)
f2 <- rnorm(300, mean=4, sd = 1.2)
N.data <- cbind(f1, f2)

# combine all samples
data.mat <- data.frame(rbind(P.data, N.data), Class=rep(c(1, 0), time=c(nrow(P.data), nrow(N.data))))
```

The above code will create a dataset with two class and two features, each follows a normal distribution.

(1) Partition the data into 80% for model training (training set) and 20% for model testing (test set).
(2) Train a Logistic Regression, a LDA and a *k*NN (try different *k* value) classifier using training dataset. Compare their performance using test dataset.
(3) For *k*NN, identify optimal *k* value by minimising classification error on test set.
(4) Now we used test set to select optimal *k*, is it still valid to use this test set to evaluate the performance of our optimised *k*NN classifier? Why or why not?