STAT 598 Homework 1 Due: 1/23/2019

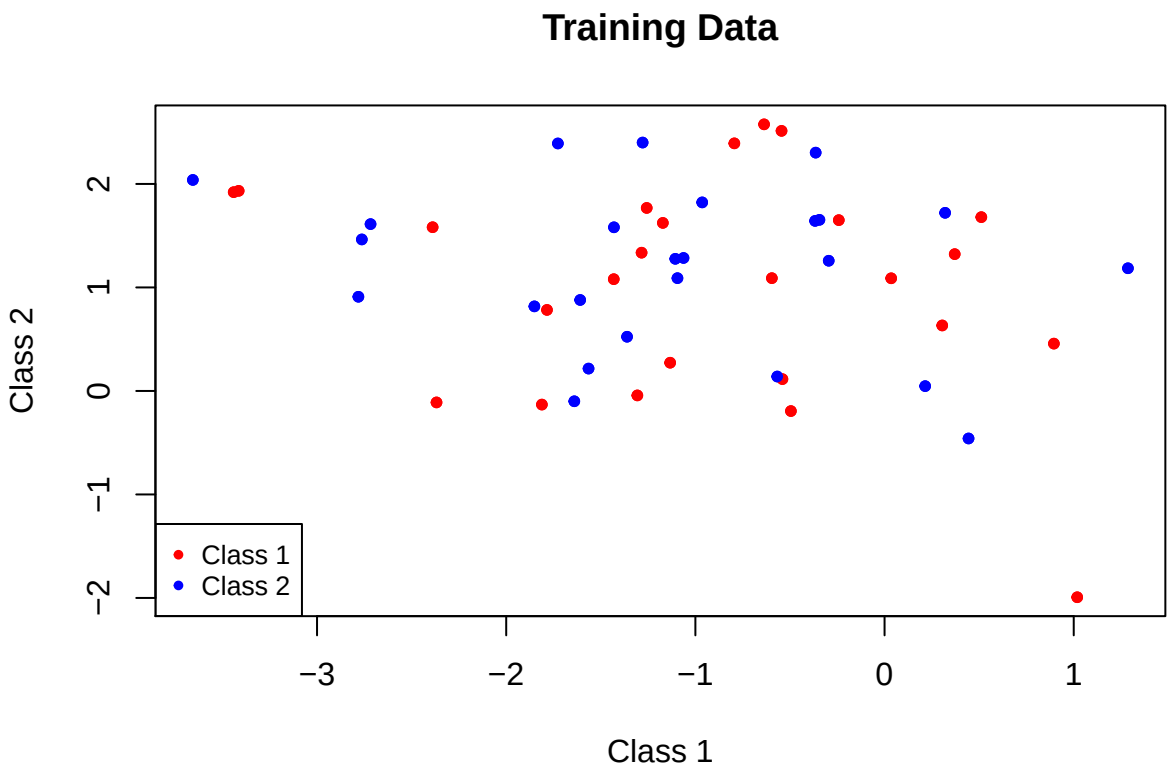
Bryan Melchert

1. (10 points) Write an R program to generate data in two classes, with two features. These features are all independent Gaussian variates with standard deviation 1. Their mean vectors are $(-1, -1)$ in class 1 and $(1, 1)$ in class 2. Generate 50 observations from each class to form the training data set, and plot the observations.

Make sure that the observations are distinguishable by different colors, and use `set.seed()` *function to make your work repeatable.*

```
# Set seed for reproducibility
set.seed(42)
# Generate feature variabes on Gaussian(normal) distribution
class1 <- matrix(data = rnorm(50, mean=c(-1,-1), sd=1), nrow=50, ncol = 2)
class2 <- matrix(data = rnorm(50, mean=c(1,1), sd=1), nrow=50, ncol = 2)
# Encode target variables, Merge dataframes
a = data.frame(rep(0,50))
d1 = cbind.data.frame(class1, a)
colnames(d1) = c("DataPoint1", "DataPoint2", "Y")
b = data.frame(rep(1,50))
d2 = cbind.data.frame(class2, b)
colnames(d2) = c("DataPoint1", "DataPoint2", "Y")
train = merge(d1,d2, all=TRUE)
# Convert NA values to 0
train[is.na(train)] <- 0

# Plot features versus target according to class color
plot1 = plot(class1, class2, xlab ="Class 1", ylab ="Class 2", main = "Training Data", pch = 20, col=c(
legend(x ="bottomleft", legend=c("Class 1", "Class 2"), col=c("red", "blue"), pch=20, cex=0.8)
```



2. (20 points) Based on the training set in #1, which class the following new data points will belong to, respectively?

$$(0,0), \quad (-0.5,0), \quad (0,0.5)$$

a. If linear regression is used? Comment on your finding.
b. If KNN classification is used ($k = 1, 3, 5$)? Comment on your finding.

```
# Linear Regression
reg2 = lm(Y~DataPoint1+DataPoint2, data=train)
# Input new data into trained model and predict Y
x.new1 = data.frame(DataPoint1=0, DataPoint2=0)
pred1 = predict(reg2, newdata=x.new1)
```

```
## Warning in predict.lm(reg2, newdata = x.new1): prediction from a rank-
## deficient fit may be misleading
```

```
# Print associated Class given the prediction on new data
final.pred1 = ifelse(pred1<.5, "Class 1", "Class 2")
x.new2 = data.frame(DataPoint1=-0.5, DataPoint2=0)
pred2 = predict(reg2, newdata=x.new2)
```

```
## Warning in predict.lm(reg2, newdata = x.new2): prediction from a rank-
## deficient fit may be misleading
```

```r
final.pred2 = ifelse(pred2<.5, "Class 1", "Class 2")
x.new3 = data.frame(DataPoint1=0, DataPoint2=0.5)
pred3 = predict(reg2, newdata=x.new3)
```

```
## Warning in predict.lm(reg2, newdata = x.new3): prediction from a rank-
## deficient fit may be misleading
```

```r
final.pred3 = ifelse(pred3<.5, "Class 1", "Class 2")

# K-Nearest Neighbors classification
library(class)
knn_pred11 = knn(train=train[,1:2], test=x.new1, cl=train[,3])  # Levels: 0 1
table(knn_pred11)
```

```
## knn_pred11
## 0 1
## 1 0
```

```r
knn_pred13 = knn(train=train[,1:2], test=x.new1, cl=train[,3], k=3)
knn_pred15 = knn(train=train[,1:2], test=x.new1, cl=train[,3], k=5)
knn_pred21 = knn(train=train[,1:2], test=x.new2, cl=train[,3])
knn_pred23 = knn(train=train[,1:2], test=x.new2, cl=train[,3], k=3)
knn_pred25 = knn(train=train[,1:2], test=x.new2, cl=train[,3], k=5)
knn_pred31 = knn(train=train[,1:2], test=x.new3, cl=train[,3])
knn_pred33 = knn(train=train[,1:2], test=x.new3, cl=train[,3], k=3)
knn_pred35 = knn(train=train[,1:2], test=x.new3, cl=train[,3], k=5)
```

3. (10 points) Residential real estate prices depend, in part, on property size and number of bedrooms. The house size X_1 (in hundreds of square feet), number of bedrooms X_2, and house price Y (in thousands of dollars) of a random sample of houses in a certain county were observed.

House	1	2	3	4	5	6	7
House Size (X_1)	18	20	25	22	33	19	17
Number of Bedrooms (X_2)	3	3	4	4	5	4	3
House Price (Y)	160	190	208	220	350	170	178

If a new property in the same county has a size of 2100 square feet and 3 bedrooms. Use nearest neighbor method ($k = 3$) to predict the house price.

No R program is allowed, you must do the computation by hand and type up the result.