PCA and LDA

Jered Hightower, Haniyyah Hamid, & Sai Gonuguntla

Run PCA on the iris data

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
df <- read.csv("income_evaluation.csv", header=TRUE)
i <- sample(1:1500, 1000, replace=FALSE)

income_evaluation <- df[,c(1,5,13,15)]
income_evaluation$income <- factor(income_evaluation$income)
income_evaluation$age <- as.numeric(income_evaluation$age)
income_evaluation$education.num <- as.numeric(income_evaluation$education.num)
income_evaluation$hours.per.week <- as.numeric(income_evaluation$hours.per.week)
str(income_evaluation)
```

```
## 'data.frame':    32561 obs. of  4 variables:
##  $ age           : num  39 50 38 53 28 37 49 52 31 42 ...
##  $ education.num : num  13 13 9 7 13 14 5 9 14 13 ...
##  $ hours.per.week: num  40 13 40 40 40 40 16 45 50 40 ...
##  $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

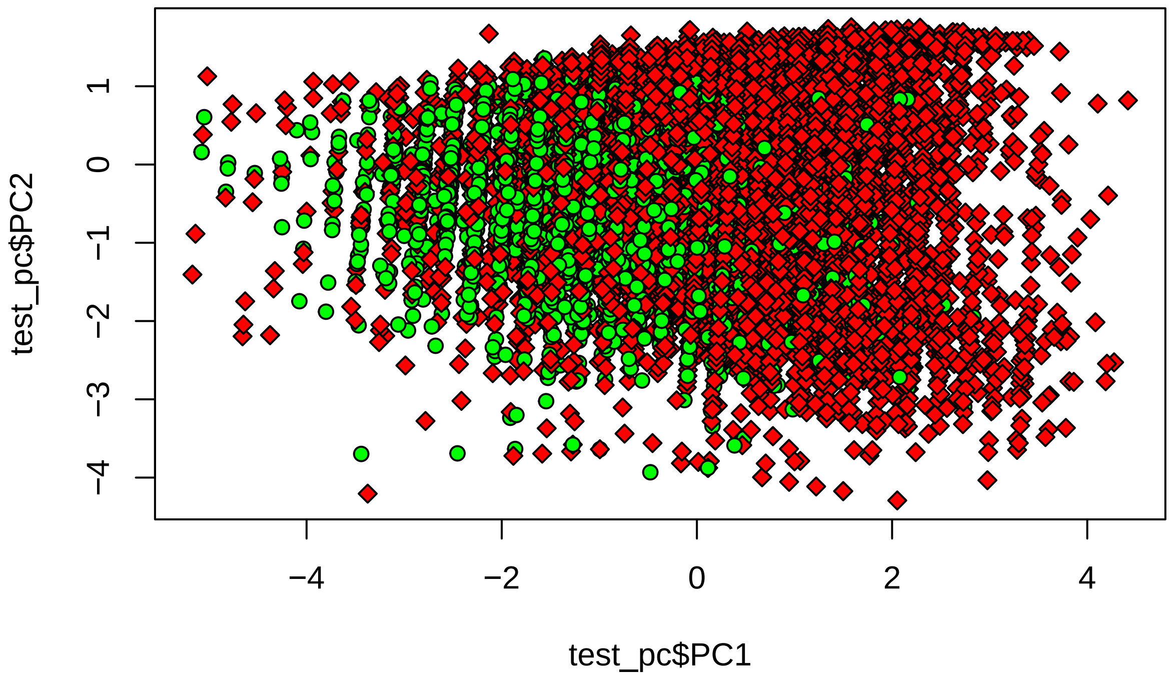
```
train <- income_evaluation[i,]
test <- income_evaluation[-i,]
set.seed(1234)
pca_out <- preProcess(train[,1:3], method=c("center", "scale", "pca"))
pca_out
```

```
## Created from 1000 samples and 3 variables
##
## Pre-processing:
##   - centered (3)
##   - ignored (0)
##   - principal component signal extraction (3)
##   - scaled (3)
##
## PCA needed 3 components to capture 95 percent of the variance
```

PCA plot

```
train_pc <- predict(pca_out, train[, 1:3])
test_pc <- predict(pca_out, test[,])

plot(test_pc$PC1, test_pc$PC2, pch=c(23,21)[unclass(test_pc$income)], bg=c("red","green")[unclass(test$i
```



PCA data in knn

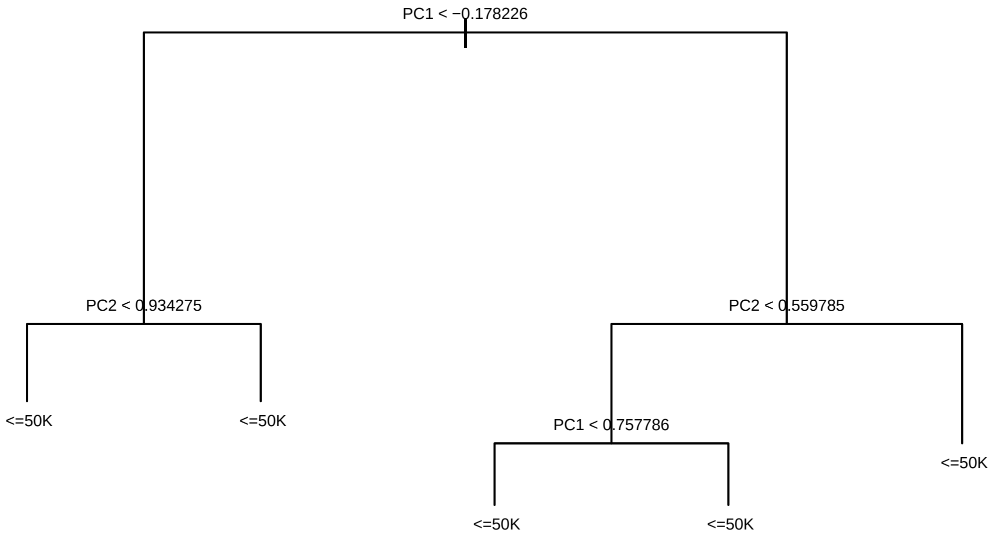
Now let's see if two principal components can predict class.

```
train_df <- data.frame(train_pc$PC1, train_pc$PC2, train$income)
test_df <- data.frame(test_pc$PC1, test_pc$PC2, test$income)
library(class)
set.seed(1234)
pred <- knn(train=train_df[,1:2], test=test_df[,1:2], cl=train_df[,3], k=2)
mean(pred==test$income)
```

```
## [1] 0.7264028
```

The accuracy is a lower than if we used all 3 predictors.

```
library(tree)
colnames(train_df) <- c("PC1", "PC2", "Income")
colnames(test_df) <- c("PC1", "PC2", "Income")
set.seed(1234)
tree1 <- tree(Income~., data=train_df)
plot(tree1)
text(tree1, cex=0.5, pretty=0)
```




```
pred <- predict(tree1, newdata=test_df, type="class")
mean(pred==test$income)
```

```
## [1] 0.759545
```

With the decision tree we got a little higher accuracy.

LDA

```
library(MASS)
lda1 <- lda(income~., data=train)
lda1$means
```

```
##              age education.num hours.per.week
##  <=50K 37.10027      9.673797       38.78075
##  >50K  44.00794     11.547619       45.30952
```

predict on test

```
lda_pred <- predict(lda1, newdata=test, type="class")
# lda_pred$class
mean(lda_pred$class==test$income)
```

```
## [1] 0.7890118
```

```
# nothing to plot, income is binary
```