PCA and LDA

Karen Mazidi

Run PCA on the iris data

```
library(caret)
```

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

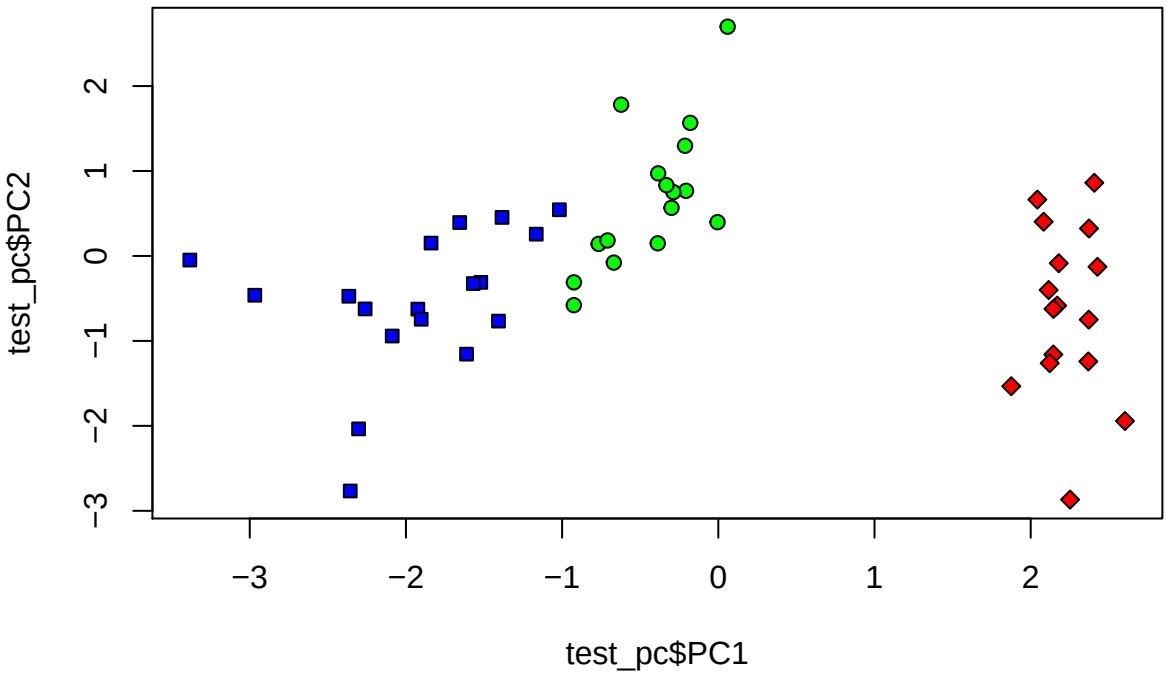
```
## Loading required package: ggplot2
```

```
data(iris)
i <- sample(1:150, 100, replace=FALSE)
train <- iris[i,]
test <- iris[-i,]
set.seed(1234)
pca_out <- preProcess(train[,1:4], method=c("center", "scale", "pca"))
pca_out
```

```
## Created from 100 samples and 4 variables
##
## Pre-processing:
##   - centered (4)
##   - ignored (0)
##   - principal component signal extraction (4)
##   - scaled (4)
##
## PCA needed 2 components to capture 95 percent of the variance
```

PCA plot

```
train_pc <- predict(pca_out, train[, 1:4])
test_pc <- predict(pca_out, test[,])
plot(test_pc$PC1, test_pc$PC2, pch=c(23,21,22)[unclass(test_pc$Species)], bg=c("red","green","blue")[un
```



###

PCA data in knn

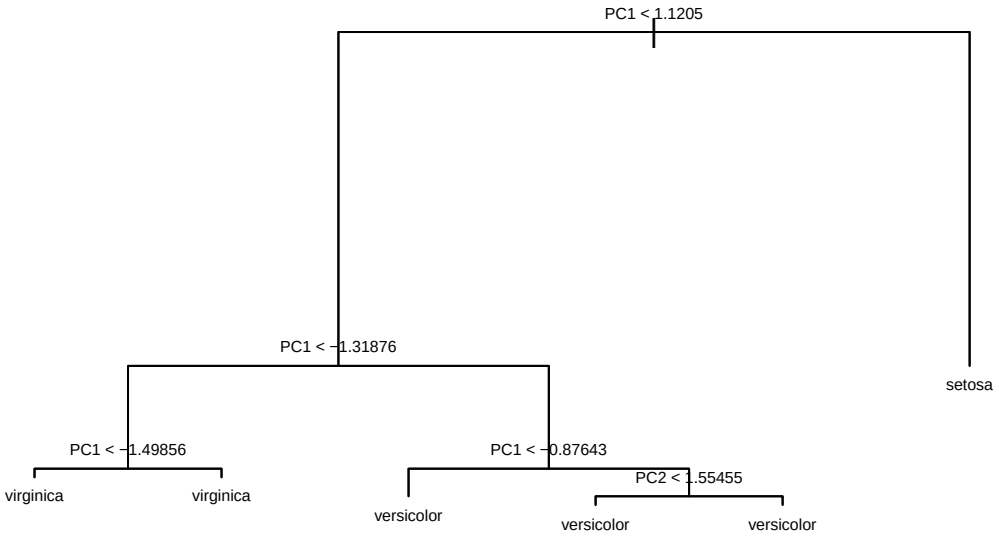
Now let's see if our two principal components can predict class.

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

```
## [1] 0.98
```

The accuracy is lower than if we used all 4 predictors.

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




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

```
## [1] 0.96
```

With the decison tree we got a little lower accuracy.

LDA

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

```
##            Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa         4.979412    3.376471     1.450000   0.2470588
## versicolor     5.979412    2.797059     4.317647   1.3588235
## virginica      6.462500    2.946875     5.468750   2.0343750
```

predict on test

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

```
##  [1] setosa     setosa     setosa     setosa     setosa     setosa
##  [7] setosa     setosa     setosa     setosa     setosa     setosa
## [13] setosa     setosa     setosa     setosa     versicolor versicolor
## [19] versicolor versicolor versicolor versicolor versicolor versicolor
## [25] versicolor versicolor versicolor versicolor versicolor versicolor
## [31] versicolor versicolor virginica  virginica  virginica  virginica
## [37] virginica  virginica  virginica  virginica  virginica  virginica
## [43] virginica  virginica  virginica  versicolor virginica  virginica
## [49] virginica  virginica
## Levels: setosa versicolor virginica
```

```
mean(lda_pred$class==test$Species)
```

```
## [1] 0.98
```

plot

```
plot(lda_pred$x[,1], lda_pred$x[,2], pch=c(23,21,22)[unclass(lda_pred$class)], bg=c("red","green","blue
```