Dimensionality Reduction

Jeffrey Li

10/9/2022

Source for the data: https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+#

Note: The data sets provided above do not meet the 10K rows requirement, so Excel was used to combine two data sets into one.

This notebook performs PCA and LDA dimensionality reduction on the data set.

Data Cleaning

The date column is removed, and "Occupancy" is turned into a factor.

```
data <- read.csv("data.csv")
data <- data[,c(2:7)]
data$Occupancy <- factor(data$Occupancy)
```

Run PCA on the data

```
library(caret)
```

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

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

```
set.seed(1234)
i <- sample(1:nrow(data), nrow(data)*0.8, replace=FALSE)
train <- data[i,]
test <- data[-i,]

pca_out <- preProcess(train[,1:5], method=c("center", "scale", "pca"))
pca_out
```

```
## Created from 9933 samples and 5 variables
##
## Pre-processing:
##   - centered (5)
##   - ignored (0)
##   - principal component signal extraction (5)
##   - scaled (5)
##
## PCA needed 4 components to capture 95 percent of the variance
```

PCA data in kNN regression

See if the 4 principle components can predict occupancy.

```
library(class)
train_pc <- predict(pca_out, train[, 1:5])
test_pc <- predict(pca_out, test[,])

train_df <- data.frame(train_pc$PC1, train_pc$PC2, train_pc$PC3, train_pc$PC4, train$Occupancy)
test_df <- data.frame(test_pc$PC1, test_pc$PC2, test_pc$PC3, test_pc$PC4, test$Occupancy)

set.seed(1234)
pred1 <- knn(train = train_df[,1:4], test=test_df[,1:4], cl=train_df[,5], k=13)
mean(pred1==test$Occupancy)
```

```
## [1] 0.9927536
```

The accuracy is 0.9927536 which is lower than the accuracy of 0.9951691 if we used all 5 variables. This means 0.0024155 accuracy was lost.

PCA data in decision tree regression

```
library(tree)
colnames(train_df) <- c("PC1", "PC2", "PC3", "PC4", "Occupancy")
colnames(test_df) <- c("PC1", "PC2", "PC3", "PC4", "Occupancy")

set.seed(1234)
tree1 <- tree(Occupancy~., data=train_df)
pred2 <- predict(tree1, newdata=test_df, type="class")
mean(pred2==test$Occupancy)
```

```
## [1] 0.9907407
```

The accuracy is 0.9907407 which is lower than the previous accuracy of 0.9943639. This means 0.0036232 accuracy was lost.

Run LDA on the data

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

```
##   Temperature Humidity     Light      CO2 HumidityRatio
## 0    20.74905 29.02691  23.62548 683.5429   0.004386398
## 1    22.14193 28.60673 493.45701 938.0636   0.004720380
```

Predict on test

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

```
## [1] 0.9879227
```

The accuracy is 0.9879227 which is lower than the previous accuracy of 0.9939614. This means 0.0060387 accuracy was lost.