Dim_Reduction

Chris Talley Group 12

2023-03-24

Dimensionality Reduction

Data Setup

In this notebook I perform PCA and LDA on the same airline dataset used in classification. This is done to better compare the accuracy of the final models. The following code sets the data up with the same seed as the classification notebook. Printing head(data) shows an output equivalent to classification's starting set.

```
set.seed(1111)
curr_path = rstudioapi::getActiveDocumentContext()$path
setwd(dirname(curr_path))
data1 <- read.csv("train.csv")
data2 <- read.csv("test.csv")
data <- rbind(data1, data2)
head(data)
```

```
##   X     id Gender     Customer.Type Age  Type.of.Travel    Class
## 1 0  70172   Male    Loyal Customer  13 Personal Travel Eco Plus
## 2 1   5047   Male disloyal Customer  25 Business travel Business
## 3 2 110028 Female    Loyal Customer  26 Business travel Business
## 4 3  24026 Female    Loyal Customer  25 Business travel Business
## 5 4 119299   Male    Loyal Customer  61 Business travel Business
## 6 5 111157 Female    Loyal Customer  26 Personal Travel      Eco
##   Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient
## 1             460                     3                                 4
## 2             235                     3                                 2
## 3            1142                     2                                 2
## 4             562                     2                                 5
## 5             214                     3                                 3
## 6            1180                     3                                 4
##   Ease.of.Online.booking Gate.location Food.and.drink Online.boarding
## 1                      3             1              5               3
## 2                      3             3              1               3
## 3                      2             2              5               5
## 4                      5             5              2               2
## 5                      3             3              4               5
## 6                      2             1              1               2
##   Seat.comfort Inflight.entertainment On.board.service Leg.room.service
## 1            5                      5                4                3
## 2            1                      1                1                5
## 3            5                      5                4                3
## 4            2                      2                2                5
## 5            5                      3                3                4
```

```
## 6            1                     1                3               4
##   Baggage.handling Checkin.service Inflight.service Cleanliness
## 1                4               4                5           5
## 2                3               1                4           1
## 3                4               4                4           5
## 4                3               1                4           2
## 5                4               3                3           3
## 6                4               4                4           1
##   Departure.Delay.in.Minutes Arrival.Delay.in.Minutes            satisfaction
## 1                         25                       18 neutral or dissatisfied
## 2                          1                        6 neutral or dissatisfied
## 3                          0                        0               satisfied
## 4                         11                        9 neutral or dissatisfied
## 5                          0                        0               satisfied
## 6                          0                        0 neutral or dissatisfied
```

Preprocessing and Data Cleaning

Drop non-essential predictor columns

```
data <- subset(data, select = -c(X, id, Customer.Type))
head(data)
```

```
##   Gender Age  Type.of.Travel    Class Flight.Distance Inflight.wifi.service
## 1   Male  13 Personal Travel Eco Plus             460                     3
## 2   Male  25 Business travel Business             235                     3
## 3 Female  26 Business travel Business            1142                     2
## 4 Female  25 Business travel Business             562                     2
## 5   Male  61 Business travel Business             214                     3
## 6 Female  26 Personal Travel      Eco            1180                     3
##   Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location
## 1                                 4                      3             1
## 2                                 2                      3             3
## 3                                 2                      2             2
## 4                                 5                      5             5
## 5                                 3                      3             3
## 6                                 4                      2             1
##   Food.and.drink Online.boarding Seat.comfort Inflight.entertainment
## 1              5               3            5                      5
## 2              1               3            1                      1
## 3              5               5            5                      5
## 4              2               2            2                      2
## 5              4               5            5                      3
## 6              1               2            1                      1
##   On.board.service Leg.room.service Baggage.handling Checkin.service
## 1                4                3                4               4
## 2                1                5                3               1
## 3                4                3                4               4
## 4                2                5                3               1
## 5                3                4                4               3
## 6                3                4                4               4
##   Inflight.service Cleanliness Departure.Delay.in.Minutes
## 1                5           5                        25
## 2                4           1                         1
## 3                4           5                         0
```

```
## 4                4           2                        11
## 5                3           3                         0
## 6                4           1                         0
##   Arrival.Delay.in.Minutes            satisfaction
## 1                       18 neutral or dissatisfied
## 2                        6 neutral or dissatisfied
## 3                        0               satisfied
## 4                        9 neutral or dissatisfied
## 5                        0               satisfied
## 6                        0 neutral or dissatisfied
```

Mapping categorical non-numerical predictors to different ranges

```
#data$Customer.Type <- ifelse(data$Customer.Type=="Local Customer", 1, 0)
data$Gender <- ifelse(data$Gender=="Female", 1, 0)
data$Type.of.Travel <- ifelse(data$Type.of.Travel=="Business travel", 1, 0)
data$Class[data$Class == "Eco"] <- 0
data$Class[data$Class == "Eco Plus"] <- 1
data$Class[data$Class == "Business"] <- 2
```

Check columns for NA's

```
print(sapply(data, function(y) sum(length(which(is.na(y))))))
```

```
##                            Gender                               Age
##                                 0                                 0
##                      Type.of.Travel                             Class
##                                 0                                 0
##                     Flight.Distance              Inflight.wifi.service
##                                 0                                 0
## Departure.Arrival.time.convenient             Ease.of.Online.booking
##                                 0                                 0
##                       Gate.location                    Food.and.drink
##                                 0                                 0
##                     Online.boarding                      Seat.comfort
##                                 0                                 0
##             Inflight.entertainment                  On.board.service
##                                 0                                 0
##                    Leg.room.service                  Baggage.handling
##                                 0                                 0
##                     Checkin.service                  Inflight.service
##                                 0                                 0
##                         Cleanliness         Departure.Delay.in.Minutes
##                                 0                                 0
##           Arrival.Delay.in.Minutes                      satisfaction
##                               393                                 0
```

Check columns for scores of 0

```
print(sapply(data, function(y) sum(length(which(y==0)))))
```

```
##                            Gender                               Age
##                             63981                                 0
##                      Type.of.Travel                             Class
##                             40187                             58309
##                     Flight.Distance              Inflight.wifi.service
##                                 0                              3916
```

```
## Departure.Arrival.time.convenient            Ease.of.Online.booking
##                                6681                              5682
##                      Gate.location                      Food.and.drink
##                                  1                               132
##                   Online.boarding                       Seat.comfort
##                               3080                                 1
##            Inflight.entertainment                   On.board.service
##                                 18                                 5
##                  Leg.room.service                   Baggage.handling
##                               598                                 0
##                   Checkin.service                   Inflight.service
##                                  1                                 5
##                       Cleanliness         Departure.Delay.in.Minutes
##                                 14                             73356
##          Arrival.Delay.in.Minutes                      satisfaction
##                             72753                                 0
```

Drop these observations then print number of observations

```
data <- data[!(is.na(data$Arrival.Delay.in.Minutes)),]
data <- data[!(data$Gate.location==0),]
data <- data[!(data$Food.and.drink==0),]
data <- data[!(data$Inflight.wifi.service==0),]
data <- data[!(data$Departure.Arrival.time.convenient==0),]
data <- data[!(data$Ease.of.Online.booking==0),]
data <- data[!(data$Online.boarding==0),]
data <- data[!(data$Seat.comfort==0),]
data <- data[!(data$Inflight.entertainment==0),]
data <- data[!(data$On.board.service==0),]
data <- data[!(data$Leg.room.service==0),]
data <- data[!(data$Checkin.service==0),]
data <- data[!(data$Inflight.service==0),]
data <- data[!(data$Cleanliness==0),]
print(nrow(data))
```

```
## [1] 119204
```

Convert satisfaction to a factor

```
data$satisfaction<-as.numeric(as.factor(data$satisfaction))
head(data)
```

```
##   Gender Age Type.of.Travel Class Flight.Distance Inflight.wifi.service
## 1      0  13              0     1             460                     3
## 2      0  25              1     2             235                     3
## 3      1  26              1     2            1142                     2
## 4      1  25              1     2             562                     2
## 5      0  61              1     2             214                     3
## 6      1  26              0     0            1180                     3
##   Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location
## 1                                 4                      3             1
## 2                                 2                      3             3
## 3                                 2                      2             2
## 4                                 5                      5             5
## 5                                 3                      3             3
## 6                                 4                      2             1
```

```
##   Food.and.drink Online.boarding Seat.comfort Inflight.entertainment
## 1              5               3            5                      5
## 2              1               3            1                      1
## 3              5               5            5                      5
## 4              2               2            2                      2
## 5              4               5            5                      3
## 6              1               2            1                      1
##   On.board.service Leg.room.service Baggage.handling Checkin.service
## 1                4                3                4                4
## 2                1                5                3                1
## 3                4                3                4                4
## 4                2                5                3                1
## 5                3                4                4                3
## 6                3                4                4                4
##   Inflight.service Cleanliness Departure.Delay.in.Minutes
## 1                5           5                        25
## 2                4           1                         1
## 3                4           5                         0
## 4                4           2                        11
## 5                3           3                         0
## 6                4           1                         0
##   Arrival.Delay.in.Minutes satisfaction
## 1                       18            1
## 2                        6            1
## 3                        0            2
## 4                        9            1
## 5                        0            2
## 6                        0            1
```

80:20 Train:Test split then print number of training observations

```
i <- sample(1:nrow(data), nrow(data)*0.80, replace=FALSE)
train <- data[i,]
test <- data[-i,]
print(nrow(train))
```

```
## [1] 95363
```

PCA and Predictions

I will now perform PCA on the dataset to reduce the dimensions

```
pca_out <- preProcess(data[,1:22],method=c("center","scale","pca"))
train_pc <- predict(pca_out, train[, 1:22])
test_pc <- predict(pca_out, test[,])
```

Using KNN with PCA Data

```
set.seed(1111)
pred <- knn(train=train_pc[,2:17], test=test_pc[,2:17], cl=train_pc[,1], k=3)
acc_pca <- mean(pred==test$satisfaction)
print(paste("PCA KNN Accuracy: ", acc_pca))
```

```
## [1] "PCA KNN Accuracy:  0.371586762300239"
```

LDA and Predictions

```
lda_out <- lda(satisfaction~., data=train)
lda_out$means
```

```
##      Gender      Age Type.of.Travel     Class1    Class2 Flight.Distance
## 1 0.5116598 37.85834      0.5069227 0.09821674 0.2684225        961.3404
## 2 0.4978864 42.54586      0.9398103 0.03954483 0.7801317       1580.7479
##   Inflight.wifi.service Departure.Arrival.time.convenient
## 1              2.411321                         3.277421
## 2              3.360352                         3.112048
##   Ease.of.Online.booking Gate.location Food.and.drink Online.boarding
## 1               2.621509      2.982314       2.955866        2.707526
## 2               3.221466      2.987097       3.566727        4.164004
##   Seat.comfort Inflight.entertainment On.board.service Leg.room.service
## 1     3.036013               2.881792         3.001189         2.997952
## 2     4.025610               4.055102         3.906631         3.897046
##   Baggage.handling Checkin.service Inflight.service Cleanliness
## 1         3.361152        3.030416         3.380247    2.923951
## 2         4.002359        3.654198         4.006931    3.791658
##   Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
## 1                   16.46738                 17.20627
## 2                   12.62129                 12.71230
```

```
lda_pred <- predict(lda_out, newdata=test, type="class")
acc_lda <- mean(lda_pred$class==test$satisfaction)
print(paste("LDA Accuracy: ", acc_lda))
```

```
## [1] "LDA Accuracy:  0.88255526194371"
```

Result Analysis

For this analysis I ran KNN with the PCA output, and predicted with LDA output. The PCA accuracy is significantly decreased. Originally, KNN on the full data set had an accuracy of 0.714483452875299. My PCA KNN is an abysmal 0.371586762300239. The 22 variables were broken down into 17 principal components, capturing 95 percent of variance. As for LDA, much of the accuracy is retained. Logistic regression on the original data had an accuracy of 0.887378885113879. My LDA model output resulted in 0.88255526194371. This is a minimal difference, showing that LDA may be a better method for dimensionality reduction on this specific data set.