Assignment_2

Dengcheng Chen

2022/2/20

```
library(readr)
Ubank <- read_delim(file = 'UniversalBank.csv',delim=',')
```

```
## Rows: 5000 Columns: 14
## -- Column specification --------------------------------------------------
------
## Delimiter: ","
## dbl (14): ID, Age, Experience, Income, ZIP Code, Family, CCAvg, Education,
 M...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this m
essage.
```

```
library(reshape)
Ubank <- rename(Ubank,c(`Personal Loan` = 'PL','Securities Account'='SA','CD
Account'='CDA'))
names(Ubank)
```

```
##  [1] "ID"         "Age"         "Experience" "Income"      "ZIP Code"
##  [6] "Family"     "CCAvg"       "Education"  "Mortgage"    "PL"
## [11] "SA"         "CDA"         "Online"     "CreditCard"
```

```
summary(Ubank)
```

```
##        ID              Age          Experience       Income           ZIP Cod
e
##  Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9
307
##  1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91
911
##  Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93
437
##  Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93
153
##  3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94
608
##  Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96
651
##      Family          CCAvg          Education        Mortgage
##  Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0
##  1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0
##  Median :2.000   Median : 1.500   Median :2.000   Median :  0.0
##  Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5
##  3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0
##  Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0
##        PL              SA              CDA              Online
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000   Median :0.0000   Median :1.0000
##  Mean   :0.096   Mean   :0.1044   Mean   :0.0604   Mean   :0.5968
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##    CreditCard
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.294
##  3rd Qu.:1.000
##  Max.   :1.000
```

```
Ubank$ID <-  NULL
Ubank$`ZIP Code` <- NULL

Ubank$Education = as.factor(Ubank$Education)
Ubank$PL = as.factor(Ubank$PL)
summary(Ubank)
```

```
##       Age          Experience       Income           Family
##  Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   :1.000
##  1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:1.000
##  Median :45.00   Median :20.0   Median : 64.00   Median :2.000
##  Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :2.396
##  3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:3.000
##  Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :4.000
##      CCAvg        Education    Mortgage      PL             SA
##  Min.   : 0.000   1:2096    Min.   :  0.0   0:4520   Min.   :0.0000
##  1st Qu.: 0.700   2:1403    1st Qu.:  0.0   1: 480   1st Qu.:0.0000
##  Median : 1.500   3:1501    Median :  0.0            Median :0.0000
##  Mean   : 1.938             Mean   : 56.5            Mean   :0.1044
##  3rd Qu.: 2.500             3rd Qu.:101.0            3rd Qu.:0.0000
##  Max.   :10.000             Max.   :635.0            Max.   :1.0000
##       CDA             Online          CreditCard
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :0.0000   Median :1.0000   Median :0.000
##  Mean   :0.0604   Mean   :0.5968   Mean   :0.294
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.000
```

```
library(caret)
```

```
## 载入需要的程辑包：ggplot2
## 载入需要的程辑包：lattice
```

```
library(class)
```

```
##
## 载入程辑包：'class'
```

```
## The following object is masked from 'package:reshape':
##
##     condense
```

```
dummies <- dummyVars(PL ~ ., data = Ubank)
Ubank_dummy <- as.data.frame(predict(dummies, newdata= Ubank))
```

```
## Warning in model.frame.default(Terms, newdata, na.action = na.action, xlev
 =
## object$lvls): variable 'PL' is not a factor
```

```
head(Ubank_dummy)
```

```
##   Age Experience Income Family CCAvg Education.1 Education.2 Education.3
## 1  25          1     49      4   1.6           1           0           0
## 2  45         19     34      3   1.5           1           0           0
## 3  39         15     11      1   1.0           1           0           0
## 4  35          9    100      1   2.7           0           1           0
## 5  35          8     45      4   1.0           0           1           0
## 6  37         13     29      4   0.4           0           1           0
##   Mortgage SA CDA Online CreditCard
## 1        0  1   0      0          0
## 2        0  1   0      0          0
## 3        0  0   0      0          0
## 4        0  0   0      0          0
## 5        0  0   0      0          1
## 6      155  0   0      1          0
```

```
Norm_model <- preProcess(Ubank_dummy,method = c("center", "scale"))
Ubank_norm = predict(Norm_model,Ubank_dummy)
summary(Ubank_norm)
```

```
##       Age               Experience             Income             Family
##  Min.   :-1.94871   Min.   :-2.014710   Min.   :-1.4288   Min.   :-1.2167
##  1st Qu.:-0.90188   1st Qu.:-0.881116   1st Qu.:-0.7554   1st Qu.:-1.2167
##  Median :-0.02952   Median :-0.009121   Median :-0.2123   Median :-0.3454
##  Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.84284   3rd Qu.: 0.862874   3rd Qu.: 0.5263   3rd Qu.: 0.5259
##  Max.   : 1.88967   Max.   : 1.996468   Max.   : 3.2634   Max.   : 1.3973

##      CCAvg          Education.1       Education.2       Education.3    
##  Min.   :-1.1089   Min.   :-0.8495   Min.   :-0.6245   Min.   :-0.6549  
##  1st Qu.:-0.7083   1st Qu.:-0.8495   1st Qu.:-0.6245   1st Qu.:-0.6549  
##  Median :-0.2506   Median :-0.8495   Median :-0.6245   Median :-0.6549  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.3216   3rd Qu.: 1.1770   3rd Qu.: 1.6010   3rd Qu.: 1.5266  
##  Max.   : 4.6131   Max.   : 1.1770   Max.   : 1.6010   Max.   : 1.5266  
##     Mortgage             SA               CDA              Online       
##  Min.   :-0.5555   Min.   :-0.3414   Min.   :-0.2535   Min.   :-1.2165  
##  1st Qu.:-0.5555   1st Qu.:-0.3414   1st Qu.:-0.2535   1st Qu.:-1.2165  
##  Median :-0.5555   Median :-0.3414   Median :-0.2535   Median : 0.8219  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.4375   3rd Qu.:-0.3414   3rd Qu.:-0.2535   3rd Qu.: 0.8219  
##  Max.   : 5.6875   Max.   : 2.9286   Max.   : 3.9438   Max.   : 0.8219  
##    CreditCard     
##  Min.   :-0.6452  
##  1st Qu.:-0.6452  
##  Median :-0.6452  
##  Mean   : 0.0000  
##  3rd Qu.: 1.5495  
##  Max.   : 1.5495
```

```
Ubank_norm$PL=Ubank$PL

Train_Index = createDataPartition(Ubank$PL,p=0.6, list=FALSE)
Train.df = Ubank_norm[Train_Index,]
Validation.df = Ubank_norm[-Train_Index,]
```

Q1 Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education_1 = 0, Education_2 = 1, Education_3 = 0, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1, and CreditCard = 1. Perform a k-NN classification with all predictors except ID and ZIP code using k = 1. Remember to transform categorical predictors with more than two categories into dummy variables first. Specify the success class as 1 (loan acceptance), and use the default cutoff value of 0.5. How would this customer be classified?

```
To_Predict=data.frame(Age=40, Experience = 10, Income = 84, Family = 2, CCAvg
 = 2,
                      Education.1 = 0, Education.2 = 1, Education.3 = 0, Mort
gage = 0,
                     SA = 0, CDA = 0, Online = 1,CreditCard = 1)

print(To_Predict)

##   Age Experience Income Family CCAvg Education.1 Education.2 Education.3
## 1  40         10     84      2     2           0           1           0
##   Mortgage SA CDA Online CreditCard
## 1        0  0   0      1          1
```

```
To_Predict_norm <- predict(Norm_model,To_Predict)
print(To_Predict_norm)

##          Age Experience    Income     Family     CCAvg Education.1 Educati
on.2
## 1 -0.4657003 -0.8811162 0.2221371 -0.3453975 0.0355115  -0.8494814    1.60
1024
##   Education.3   Mortgage         SA        CDA    Online CreditCard
## 1  -0.6548999 -0.5554684 -0.3413892 -0.2535149 0.8218687   1.549477

Prediction <- knn(train = Train.df[1:13],
                  test = To_Predict_norm[1:13],
                  cl=Train.df$PL,
                  k=1)
print(Prediction)

## [1] 0
## Levels: 0 1
```

Addicting to the result, the customer will not be the targeted one.

Q2 What is a choice of k that balances between overfitting and ignoring the predictor information?

```
set.seed(123)
fitControl <- trainControl(method = "repeatedcv",
                           number = 3,
                           repeats = 2)
searchGrid=expand.grid(k = 1:15)
Knn.model=train(PL~.,
                data=Train.df,
                method='knn',
                tuneGrid=searchGrid,
                trControl = fitControl,)

Knn.model

## k-Nearest Neighbors
##
## 3000 samples
##   13 predictor
##    2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (3 fold, repeated 2 times)
## Summary of sample sizes: 2000, 2000, 2000, 2000, 2000, 2000, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.9561667  0.7225052
##    2  0.9531667  0.7043490
##    3  0.9573333  0.7078895
##    4  0.9553333  0.6929361
##    5  0.9558333  0.6916585
##    6  0.9548333  0.6848947
##    7  0.9521667  0.6592196
##    8  0.9496667  0.6340200
##    9  0.9488333  0.6274694
##   10  0.9480000  0.6163411
##   11  0.9480000  0.6139964
##   12  0.9473333  0.6082642
##   13  0.9470000  0.6048033
##   14  0.9438333  0.5748143
##   15  0.9441667  0.5757008
## 
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 3.
```

Q3 Show the confusion matrix for the validation data that results from using the best k.

```
predictions<-predict(Knn.model,Validation.df)
confusionMatrix(predictions,Validation.df$PL)
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0 1800   73
##          1    8  119
##                                           
##                Accuracy : 0.9595          
##                  95% CI : (0.9499, 0.9677)
##     No Information Rate : 0.904           
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.7251          
##                                           
##  Mcnemar's Test P-Value : 1.151e-12       
##                                           
##             Sensitivity : 0.9956          
##             Specificity : 0.6198          
##          Pos Pred Value : 0.9610          
##          Neg Pred Value : 0.9370          
##              Prevalence : 0.9040          
##          Detection Rate : 0.9000          
##    Detection Prevalence : 0.9365          
##       Balanced Accuracy : 0.8077          
##                                           
##        'Positive' Class : 0               
## 
```

Q4 Consider the following customer: Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education_1 = 0, Education_2 = 1, Education_3 = 0, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1 and Credit Card = 1. Classify the customer using the best k.

```
To_Predict=data.frame(Age=40, Experience = 10, Income = 84, Family = 2, CCAvg
 = 2,
                      Education.1 = 0, Education.2 = 1, Education.3 = 0, Mort
gage = 0,
                     SA = 0, CDA = 0, Online = 1,CreditCard = 1)
To_Predict_norm=predict(Norm_model,To_Predict)
predict(Knn.model,To_Predict_norm)

## [1] 0
## Levels: 0 1
```

Also, this cumstomer is not the targeted one.