Objective:

In this session, you will learn to implement kNN for classification and regression problems.

Key takeaways:

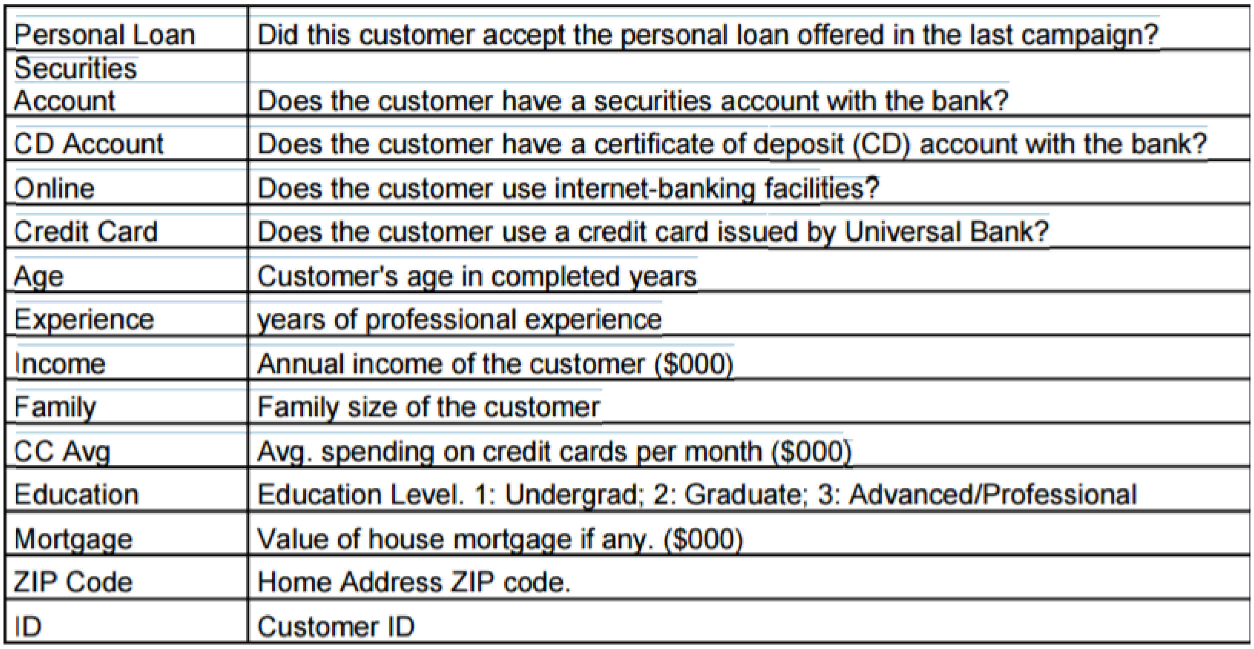
- kNN concepts – impact of variable standardization, arriving at number of neighbors
- Applying kNN to solve classification and regression problems respectively
- Condensing point/Border points

Case Study:

Banking Data:

The objective is to study the characteristics of loan customers and classify the loan takers and non-loan takers. The study is to provide the marketing division of a young bank with information to set up a new campaign to gain more loan customers. The specific questions of interest are what combination of parameters makes a customer more likely to accept a personal loan and are there any association among special offers as online services, security accounts, credit cards that support cross-selling opportunities.

The data set includes 5000 observations with 14 variables divided into four different measurement categories. The binary category has five variables, including the target variable personal loan, also securities account, CD account, online banking and credit card. The interval category contains five variables: age, experience, income, CC avg and mortgage. The ordinal category includes then variables family and education. The last category is nominal with ID and Zip code.



Personal Loan	Did this customer accept the personal loan offered in the last campaign?
Securities Account	Does the customer have a securities account with the bank?
CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Does the customer use internet-banking facilities?
Credit Card	Does the customer use a credit card issued by Universal Bank?
Age	Customer's age in completed years
Experience	years of professional experience
Income	Annual income of the customer ($000)
Family	Family size of the customer
CC Avg	Avg. spending on credit cards per month ($000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. ($000)
ZIP Code	Home Address ZIP code.
ID	Customer ID

KNN-Classification:

Steps to follow to execute the problem:

1. Load data ‘UniversalBank.csv’ into R
2. Consider “Personal.Loan “ as target attribute
3. Understand the summary of data
4. Check for the missing values, if so, impute them
5. Remove the columns that may not be used for analysis (ID and Zip Code)

```
bankdata2 = subset(bankdata, select=-c(ID,ZIP.Code))
```

6. Convert categorical attributes into numeric
   a. Convert “education” as factor variable
   b. Modify into dummy variable, add the dummy attributes to data and drop the original one

```
bankdata2$Education=as.factor(as.character(bankdata2$Education))
Education=dummy(bankdata2$Education)
bankdata3=subset(bankdata2,select=-c(Education)) bankdata4=cbind(bankdata3,Education)
```

7. Should we standardize the data?
   a. Let us check the results with and without standardizing the data

<u>Without standardizing the data</u>

- ➢ Split this data set into train and test and observe the distribution of personal loan in train and test data
- ➢ Separate out independent attributes and target variable into two data frames

```
bankdata_trainwithoutclass=subset(bankdata_train,select=-c(Personal.Loan))
bankdata_testwithoutclass=subset(bankdata_test,select=-c(Personal.Loan))
```

- ➢ Run the model on test data using different k = 2,3,4,5,6,7 and check the accuracy values and come up with optimal k value

```
pred=knn(bankdata_trainwithoutclass,bankdata_testwithoutclass,
bankdata_train$Personal.Loan,k=1)
a=table(pred,bankdata_test$Personal.Loan)
a = sum(diag(a))/nrow(bankdata_testwithoutclass)
accu
```

<u>Standardizing the data</u>

- ➢ Standardize the independent attributes data using ‘Range’ method and then merge the target variable with this standardize data

```
library(vegan)
bankdata5 = decostand(bankdata4,"range")
```

- Split this data set into train and test and observe the distribution of personal loan in train and test data

- Separate out independent attributes and target variable in two data frames

  ```
  bankdata_trainwithoutclass=subset(bankdata_train,select=-c(Personal.Loan))
  bankdata_testwithoutclass = subset(bankdata_test,select=-c(Personal.Loan))
  ```

- Run the model on test data using different k values and check the accuracy values and come up with optimal k value

  ```
  pred=knn(bankdata_trainwithoutclass,bankdata_testwithoutclass,
  bankdata_train$Personal.Loan,k=1)
  a=table(pred,bankdata_test$Personal.Loan)
  a
  accu= sum(diag(a))/nrow(bankdata_testwithoutclass)
  accu
  ```

  Write a for loop to compute the accuracy for K = 1 to 10.

8. Condensing to reduce the complexity of the model

   ```
   keep = condense(bankdata_trainwithoutclass, bankdata_train$Personal.Loan)
   keep
   ```

9. Take condensed data and run the model compare the accuracies with whole data and condensed data

   ```
   pred=knn(bankdata_trainwithoutclass[keep,],bankdata_testwithoutclass,
   bankdata_train$Personal.Loan[keep],k=5)
   a <- table(pred,bankdata_test$Personal.Loan)
   a
   accu=sum(diag(a))/nrow(bankdata_testwithoutclass)
   accu
   ```

10. Now we can find the indices of the records that are considered for prediction in the model for a specific record of test data using FNN library.

    - First install library FNN and do the following

      ```
      # run the model using FNN library
      library(FNN)
      pred=FNN::knn(bankdata_trainwithoutclass[keep,],bankdata_testwithoutclass,
                          bankdata_train$Personal.Loan[keep],k=5)
      a <- table(pred,bankdata_test$Personal.Loan)
      a
      accu = sum(diag(a))/nrow(bankdata_test)
      accu
      ```

```
indices = knnx.index(bankdata_trainwithoutclass[keep,], bankdata_testwithoutclass,
          k=5)
```

```
# If you want the indices of the 5 nearest neighbors for the row 20 of test dataset
print(indices[20, ])
```

KNN-Regression:

Objective is to predict the ccavg of customers using the Knn regression method. Here is the sample code for a dummy dataset.

Steps to follow to execute the problem:

1. Install the packages FNN, Metrics

```
install.packages("FNN")  #"Fast Nearest Neighbours" for knn regression
install.packages("Metrics") #to calculate error metrics for regression
```

2. Let us generate the data for regression

```
#set.seed()
set.seed(12345) #to get same random numbers generated every time
#Create a dataframe of 100 rows and 25 columns
data <- data.frame(matrix(data = runif(2500, 24,65), nrow = 100, ncol = 25))
```

3. Target attribute is “x25”

4. Split this data set into train and test

5. Separate out independent attributes and target variable in two data frame

```
## Excluding Target Variable
testData <- data[sample(81:100),1:24]
trainData <- data[1:80,1:24]
train.tgt <- data[1:80,25]
test.tgt <- data[sample(81:100),25]
```

6. Let us run KNN model now & compute rmse for different k values and come up with optimal k value

```
# Run the model
pred <- knn.reg(train = trainData, test = testData, y = train.tgt, k = 1 )
actual <- test.tgt
pred <- data.frame(pred$pred)
result2 <- rmse(actual = actual, predicted = pred)
```

Assignment:

- Download the data from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
- Apply KNN to predict if the customer will subscribe a term deposit