Midterm.R

Fumonchu

2021-11-02

Company : Stevens Project : Midterm Purpose : Midterm First Name : David Last Name : Fu CWID : 10471854 Date : Novermber 2, 2021

Midterm

```
rm(list=ls())
set.seed(513)
```

1. The function d(x,y) = sum((xi-yi)^3) is not a proper distance function as Distance function have three main properties

- Always Non-negative
- Commutative distance between point A to B is the same as B to A
- Distance between A to C must be less than or equal to distance between A to B to C The distance function above violates both first and second property for the given example of (0,0,0) and (0,1,0) the distance is -1 which is negative Also the distance between (0,1,0) and (0,0,0) is 1 which is different from -1. (0-0)$^{3+(0-1)}$3+(0-0)^3 vs (0-0)$^{3+(1-0)}$3+(0-0)

2.
```
  Infection rate     Travel Chance   Chance of Infection Given Traveled
```

England .0012 .5 .0012*.5 = .00060 Italy .0015 .2 .0015*.2 = .00030 Spain .0016 .3 .0016*.3 = .00048

.0006+.0003+.00048=.00138 * 100 = .138% chance of being infected

Given that employee traveled and was infected what is the .0006/.00138=.43478 * 100 The employee has 43.478% chance he/she traveled to England

3.

```
covidData=read.csv("C:/Users/Fumonchu/Documents/GitHub/School/CS513/Midterm/COVID19_v4.csv", hea
der=TRUE, colClasses=c("ID"="character",

"MaritalStatus"="factor",

"Infected"="factor"))
```

I

```
summary(covidData)
```

```
##       ID                 Age           Exposure     MaritalStatus
##  Length:147         Min.   :20.00   Min.   :1.00   Divorced:33  
##  Class :character   1st Qu.:31.00   1st Qu.:1.50   Married :65  
##  Mode  :character   Median :36.00   Median :3.00   Single  :49  
##                     Mean   :37.91   Mean   :2.66                
##                     3rd Qu.:45.00   3rd Qu.:4.00                
##                     Max.   :59.00   Max.   :4.00                
##                     NA's   :8                                   
##      Cases       MonthAtHospital  Infected 
##  Min.   : 5434   Min.   : 0.000   No :117  
##  1st Qu.:16513   1st Qu.: 3.000   Yes: 30  
##  Median :20385   Median : 6.000            
##  Mean   :18808   Mean   : 6.702            
##  3rd Qu.:22329   3rd Qu.: 9.000            
##  Max.   :25000   Max.   :32.000            
##                  NA's   :6                 
```

II

```
covidData[rowSums(is.na(covidData)) > 0,]
```

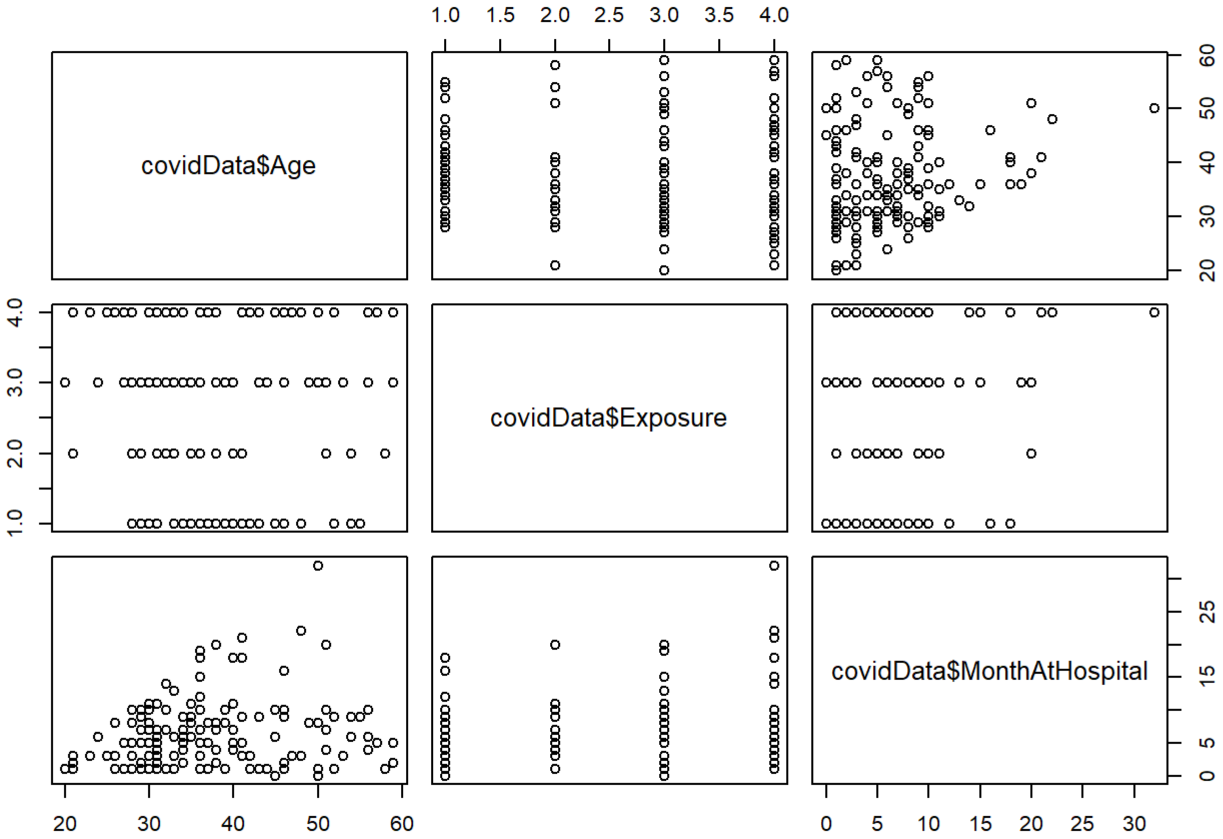
```
##       ID Age Exposure MaritalStatus Cases MonthAtHospital Infected
## 5   1001  NA        4      Divorced 10882               1       No
## 16  1001  NA        1       Married  5434              NA       No
## 39  1001  52        1       Married 25000              NA       No
## 45  1001  NA        3       Married 16177              19       No
## 51  1001  NA        3        Single  7556              10       No
## 52  1001  46        1        Single 25000              NA       No
## 55  1001  NA        4      Divorced 19837               1       No
## 75  1001  39        1       Married  7932              NA       No
## 79  1001  NA        1      Divorced 22927               3      Yes
## 93  1002  43        4       Married  8041              NA       No
## 118 1027  NA        3        Single 16211              10       No
## 127 1036  NA        3      Divorced 20999              15       No
## 131 1040  58        2        Single  5754              NA      Yes
```

III

```
mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
modeAge <- mode(covidData$Age)
modeExpo <- mode(covidData$Exposure)
modeCases <- mode(covidData$Cases)
modeHos <- mode(covidData$MonthAtHospital)
covidData$Age[is.na(covidData$Age)] <- modeAge
covidData$Exposure[is.na(covidData$Exposure)] <- modeExpo
covidData$Cases[is.na(covidData$Cases)] <- modeCases
covidData$MonthAtHospital[is.na(covidData$MonthAtHospital)] <- modeHos
```

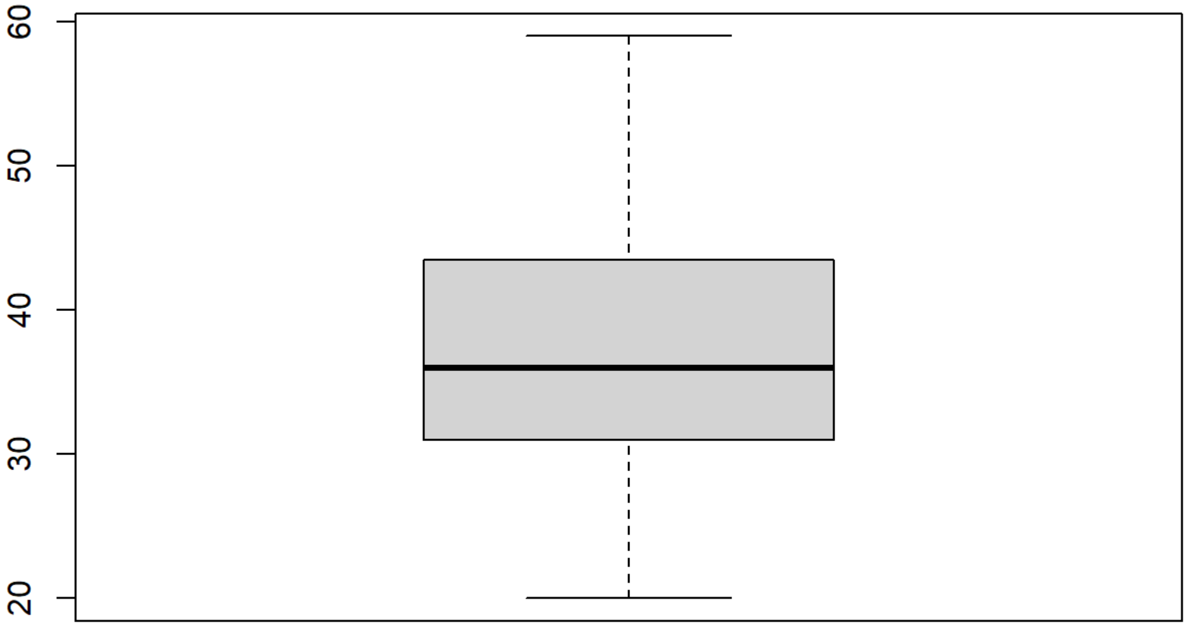
IV

```
pairs(~covidData$Age+covidData$Exposure+covidData$MonthAtHospital)
```

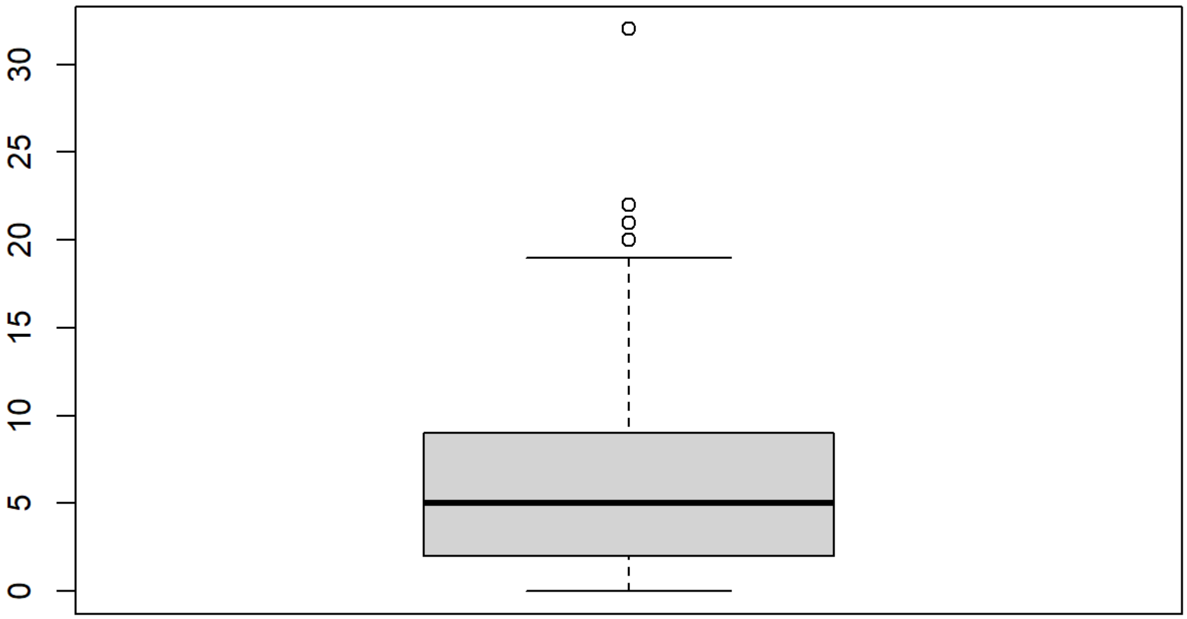


V

```
boxplot(covidData$Age)
```



```
boxplot(covidData$MonthAtHospital)
```



4 See Excel

5

```
covidData=read.csv("C:/Users/Fumonchu/Documents/GitHub/School/CS513/Midterm/COVID19_v4.csv", hea
der=TRUE, colClasses=c("ID"="character",

"MaritalStatus"="factor",

"Infected"="factor"))
covidData<-na.omit(covidData)
covidData$MonthAtHospital[covidData$MonthAtHospital < 6] <- 0
covidData$MonthAtHospital[covidData$MonthAtHospital >= 6] <- 1

covidData$Age[covidData$Age < 35] <- 0
covidData$Age[covidData$Age >= 35 & covidData$Age <= 50] <- 1
covidData$Age[covidData$Age >= 51] <- 2

idx<-sort(sample(nrow(covidData),as.integer(.70*nrow(covidData))))

training<-covidData[idx,]

test<-covidData[-idx,]

library(e1071)


nBayes <- naiveBayes(Infected~., data =training[,-1])

category_all<-predict(nBayes,test[,-1]  )


table(NBayes=category_all,Infected=test$Infected)
```

```
##       Infected
## NBayes No Yes
##    No  26   6
##    Yes  5   4
```

```
NB_wrong<-sum(category_all!=test$Infected )
NB_error_rate<-NB_wrong/length(category_all)
NB_error_rate
```

```
## [1] 0.2682927
```

6

```
library(rpart)

CART_infected<-rpart( Infected~.,data=training[,-1])
CART_predict2<-predict(CART_infected,test, type="class")
df<-as.data.frame(cbind(test,CART_predict2))
table(Actual=test[,"Infected"],CART=CART_predict2)
```

```
##       CART
## Actual No Yes
##    No  28   3
##    Yes  7   3
```

```
CART_wrong<-sum(test[,"Infected"]!=CART_predict2)

error_rate=CART_wrong/length(test$Infected)
error_rate
```

```
## [1] 0.2439024
```

7

```
library(kknn)
covidData=read.csv("C:/Users/Fumonchu/Documents/GitHub/School/CS513/Midterm/COVID19_v4.csv", header=TRUE, colClasses=c("ID"="character",

"MaritalStatus"="factor",

"Infected"="factor"))
covidData<-na.omit(covidData)
idx<-sort(sample(nrow(covidData),as.integer(.70*nrow(covidData))))

training<-covidData[idx,]

test<-covidData[-idx,]

predict_k1 <- kknn(formula= Infected~., training[,c(-1)] , test[,c(-1)], k=5,kernel ="rectangular"  )

fit <- fitted(predict_k1)
table(test$Infected,fit)
```

```
##      fit
##       No Yes
##   No  29   2
##   Yes 10   0
```

```
wrong<- ( test$Infected!=fit)
rate<-sum(wrong)/length(wrong)
rate
```

```
## [1] 0.2926829
```

8 See Excel

```
rm(list=ls())
```