SVM Classification

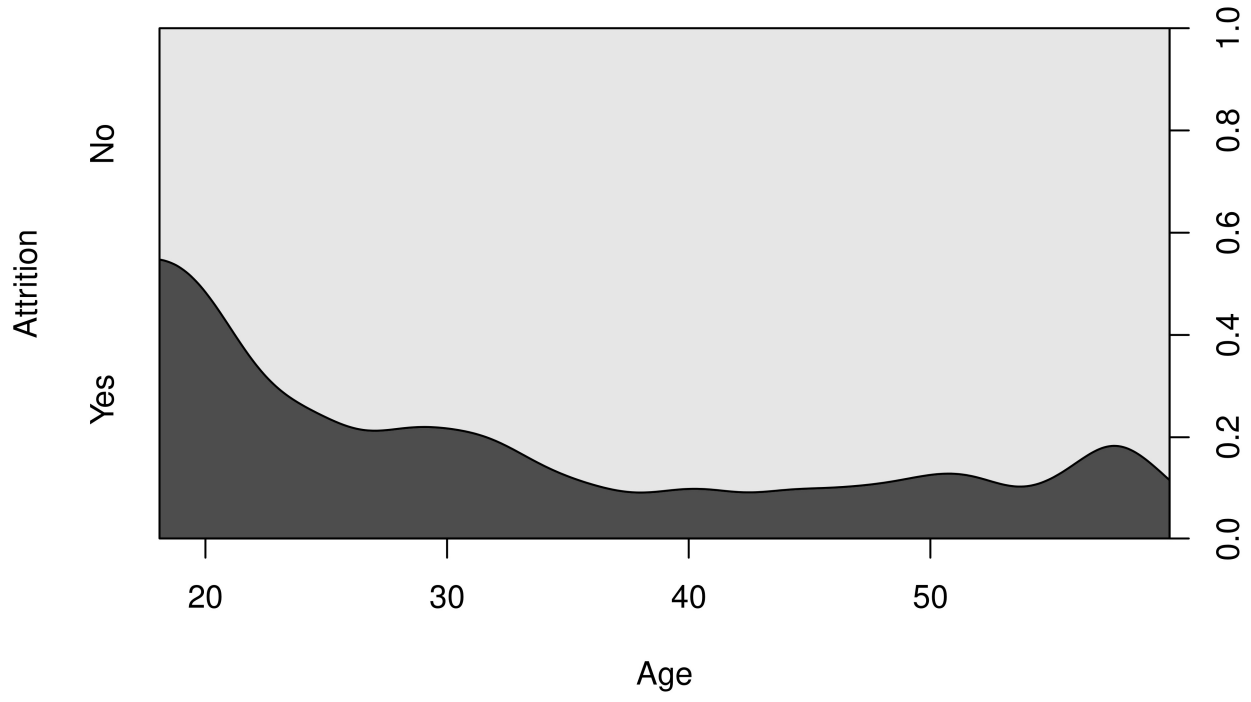
Team: Abed Ahmed (ASA190005) & Dylan Kapustka (DLK190000)

Date: 09/25/2022

HomeWork 5 - Kernel and Ensemble Methods

```
data <- read.csv(file="C:/Users/Abed/Desktop/employeeData.csv",header=TRUE)
data <- na.omit(data)
data$Attrition <- as.factor(data$Attrition)
set.seed(2)
library(caTools)


cdplot(Attrition ~ Age, data=data,xlab="Age",ylab="Attrition")
```



```r
data$Attrition <- as.numeric(data$Attrition)
for(i in 1:length(data$Attrition)){
  if(data$Attrition[i] == 1){
    data$Attrition[i] = 0
  }
  else {
    data$Attrition[i] = 1
  }
}

spec <- c(train=.6, test=.4)
i <- sample(cut(1:nrow(data),
                nrow(data)*cumsum(c(0,spec)), labels=names(spec)))
train <- data[i=="train",]
test <- data[i=="test",]
summary(train)
```

```
##       Age          Attrition      BusinessTravel      Department
##  Min.   :18.00   Min.   :0.0000   Length:2629        Length:2629
##  1st Qu.:30.00   1st Qu.:0.0000   Class :character   Class :character
##  Median :36.00   Median :0.0000   Mode  :character   Mode  :character
##  Mean   :36.95   Mean   :0.1643
##  3rd Qu.:43.00   3rd Qu.:0.0000
##  Max.   :60.00   Max.   :1.0000
##  DistanceFromHome   Education     EducationField     EmployeeCount
##  Min.   : 1.000   Min.   :1.000   Length:2629        Min.   :1
##  1st Qu.: 2.000   1st Qu.:2.000   Class :character   1st Qu.:1
##  Median : 7.000   Median :3.000   Mode  :character   Median :1
##  Mean   : 9.231   Mean   :2.932                      Mean   :1
##  3rd Qu.:14.000   3rd Qu.:4.000                      3rd Qu.:1
##  Max.   :29.000   Max.   :5.000                      Max.   :1
##    EmployeeID      Gender            JobLevel       JobRole
##  Min.   :   3   Length:2629        Min.   :1.000   Length:2629
##  1st Qu.:1102   Class :character   1st Qu.:1.000   Class :character
##  Median :2223   Mode  :character   Median :2.000   Mode  :character
##  Mean   :2219                      Mean   :2.081
##  3rd Qu.:3320                      3rd Qu.:3.000
##  Max.   :4409                      Max.   :5.000
##  MaritalStatus      MonthlyIncome    NumCompaniesWorked    Over18
##  Length:2629        Min.   : 10090   Min.   :0.000      Length:2629
##  Class :character   1st Qu.: 29040   1st Qu.:1.000      Class :character
##  Mode  :character   Median : 49300   Median :2.000      Mode  :character
##                     Mean   : 64524   Mean   :2.716
##                     3rd Qu.: 81890   3rd Qu.:4.000
##                     Max.   :199990   Max.   :9.000
##  PercentSalaryHike StandardHours StockOptionLevel TotalWorkingYears
##  Min.   :11.00     Min.   :8     Min.   :0.0000   Min.   : 0.00
##  1st Qu.:12.00     1st Qu.:8     1st Qu.:0.0000   1st Qu.: 6.00
##  Median :14.00     Median :8     Median :1.0000   Median :10.00
##  Mean   :15.23     Mean   :8     Mean   :0.7923   Mean   :11.45
##  3rd Qu.:18.00     3rd Qu.:8     3rd Qu.:1.0000   3rd Qu.:16.00
##  Max.   :25.00     Max.   :8     Max.   :3.0000   Max.   :40.00
##  TrainingTimesLastYear YearsAtCompany   YearsSinceLastPromotion
```

```
##  Min.   :0.000         Min.   : 0.000   Min.   : 0.000
##  1st Qu.:2.000         1st Qu.: 3.000   1st Qu.: 0.000
##  Median :3.000         Median : 5.000   Median : 1.000
##  Mean   :2.826         Mean   : 7.051   Mean   : 2.177
##  3rd Qu.:3.000         3rd Qu.: 9.000   3rd Qu.: 3.000
##  Max.   :6.000         Max.   :40.000   Max.   :15.000
##  YearsWithCurrManager
##  Min.   : 0.000
##  1st Qu.: 2.000
##  Median : 3.000
##  Mean   : 4.121
##  3rd Qu.: 7.000
##  Max.   :17.000
```

```
mean(train$Age)
```

```
## [1] 36.94827
```

```
median(train$Age)
```

```
## [1] 36
```

Linear Kernel

```
library(e1071)
svm1 <- svm(Attrition~Age, data=train, kernel="linear", cost=10, scale=TRUE)
summary(svm1)
```

```
##
## Call:
## svm(formula = Attrition ~ Age, data = train, kernel = "linear", cost = 10,
##     scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  linear
##        cost:  10
##       gamma:  1
##     epsilon:  0.1
##
##
## Number of Support Vectors:  867
```

```
pred <- predict(svm1,newdata=test)
prob <- ifelse(pred>0.5,1,0)
acc <- mean(prob==test$Attrition)

print(paste("Accuracy of Linear Kernel is: ", acc))
```

```
## [1] "Accuracy of Linear Kernel is:  0.844266970907017"
```

Polynomial Kernel

```
library(e1071)
svm2 <- svm(Attrition~Age, data=train, kernel="polynomial", cost=10, scale=TRUE)
summary(svm2)
```

```
##
## Call:
## svm(formula = Attrition ~ Age, data = train, kernel = "polynomial",
##     cost = 10, scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  polynomial
##        cost:  10
##      degree:  3
##       gamma:  1
##      coef.0:  0
##     epsilon:  0.1
##
##
## Number of Support Vectors:  868
```

```
pred2 <- predict(svm2,newdata=test)
prob2 <- ifelse(pred2>0.5,1,0)
acc2 <- mean(prob2==test$Attrition)
print(paste("Accuracy of Polynomial Kernel is: ", acc2))
```

```
## [1] "Accuracy of Polynomial Kernel is:  0.844266970907017"
```

```
library(e1071)
svm3 <- svm(Attrition~Age, data=train, kernel="radial", cost=10, scale=TRUE)
summary(svm3)
```

```
##
## Call:
## svm(formula = Attrition ~ Age, data = train, kernel = "radial", cost = 10,
##     scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  10
##       gamma:  1
##     epsilon:  0.1
##
##
## Number of Support Vectors:  950
```

```r
pred3 <- predict(svm3,newdata=test)
prob3 <- ifelse(pred3>0.5,1,0)
acc3 <- mean(prob3==test$Attrition)
print(paste("Accuracy of Radial Kernel is: ", acc3))
```

```
## [1] "Accuracy of Radial Kernel is:  0.844266970907017"
```

Summary and Analysis of Kernels

All the kernels performed about the same. This is most likely due to the fact that this is a simple binary classifer, there is already a very strong correlation and linear trend between age and employee Attrition, and as a result, the added dimensions with kernels are already overkill.