

Data Science II: Homework 5

Name: Jasmin Martinez (JRM2319) Date: 05/03/25

Question 1: In this problem, we will apply support vector machines to predict whether a given car gets high or low gas mileage based on the dataset “auto.csv” (used in Homework 3; see Homework 3 for more details of the dataset). The response variable is mpg cat. The predictors are cylinders, displacement, horsepower, weight, acceleration, year, and origin. Split the dataset into two parts: training data (70%) and test data (30%).

```
auto = read.csv("auto.csv")
head(auto)
```

```
##   cylinders displacement horsepower weight acceleration year origin mpg_cat
## 1         8          307         130   3504          12.0   70      1     low
## 2         8          350         165   3693          11.5   70      1     low
## 3         8          318         150   3436          11.0   70      1     low
## 4         8          304         150   3433          12.0   70      1     low
## 5         8          302         140   3449          10.5   70      1     low
## 6         8          429         198   4341          10.0   70      1     low
```

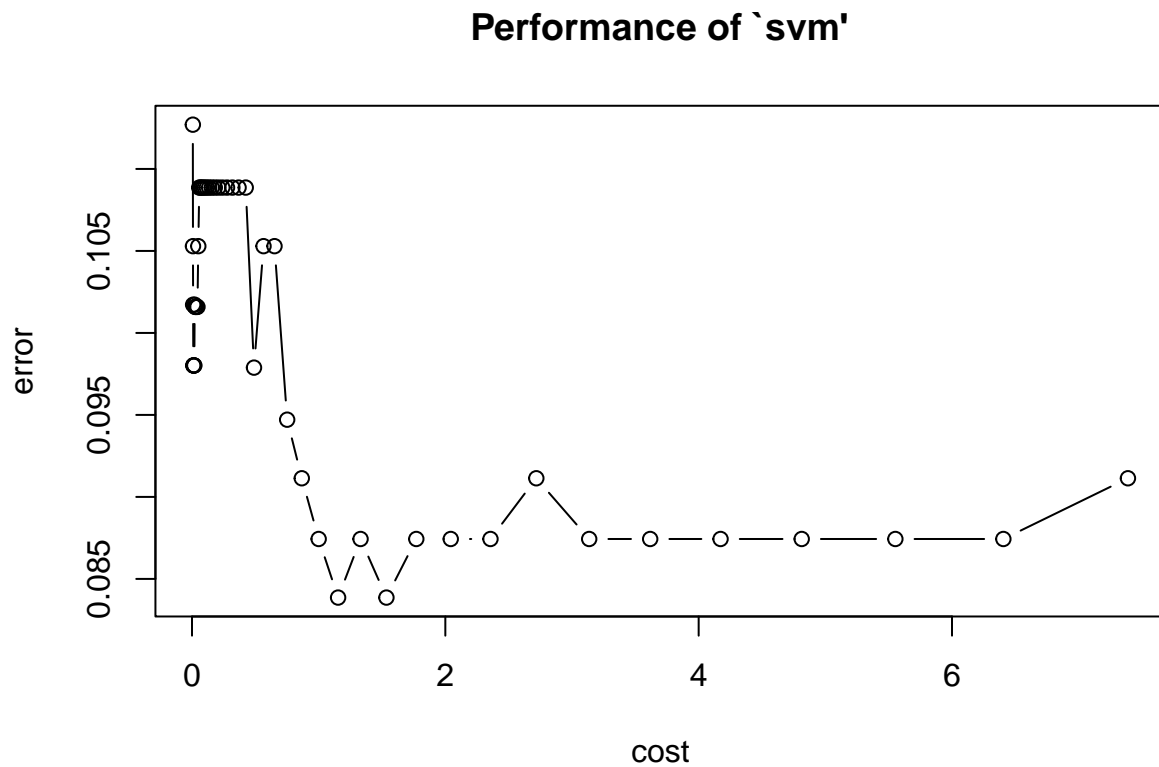
```
set.seed(111111)
datSplit = initial_split(data = auto, prop = 0.7)
trainData = training(datSplit)
testData = testing(datSplit)
head(trainData)
```

```
##   cylinders displacement horsepower weight acceleration year origin mpg_cat
## 1         4          134          95   2515          14.8   78      3     low
## 2         4          156          92   2585          14.5   82      1    high
## 3         6          168         120   3820          16.7   76      2     low
## 4         4          151          90   2670          16.0   79      1    high
## 5         6          258         110   3632          18.0   74      1     low
## 6         4           98          68   2135          16.6   78      3    high
```

```
trainData$mpg_cat = as.factor(trainData$mpg_cat)
testData$mpg_cat = as.factor(testData$mpg_cat)
```

(a): Fit a support vector classifier to the training data. What are the training and test error rates?

```
set.seed(1)
linear.tune <- tune.svm(mpg_cat ~ .,
                        data = trainData,
                        kernel = "linear",
                        cost = exp(seq(-5,2, len = 50)),
                        scale = TRUE)
plot(linear.tune)
```



(b): Fit a support vector machine with a radial kernel to the training data. What are the training and test error rates?

Question 2: In this problem, we perform hierarchical clustering on the states using the USArrests data in the ISLR package. For each of the 50 states in the United States, the dataset contains the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape. The dataset also contains the percent of the population in each state living in urban areas, UrbanPop. The four variables will be used as features for clustering.

(a): Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

(b): Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. Does scaling the variables change the clustering results? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed?