Stat 432 HW 06

Name: Your Name, netID: yournetID

Summer 2024

Include the R code for this HW.

```
knitr::opts_chunk$set(echo = TRUE)
library(ISLR2)
library(GGally)
library(tibble)
library(dplyr)
library(knitr)
library(kableExtra)
library(caret)
library(e1071)
library(gam)
library(splines)
#add more libraries as needed.
```

Question 1 (SVM)

You're given 9 observations in $p = 2$ dimensions. For each observations, there is an associated class label (y).

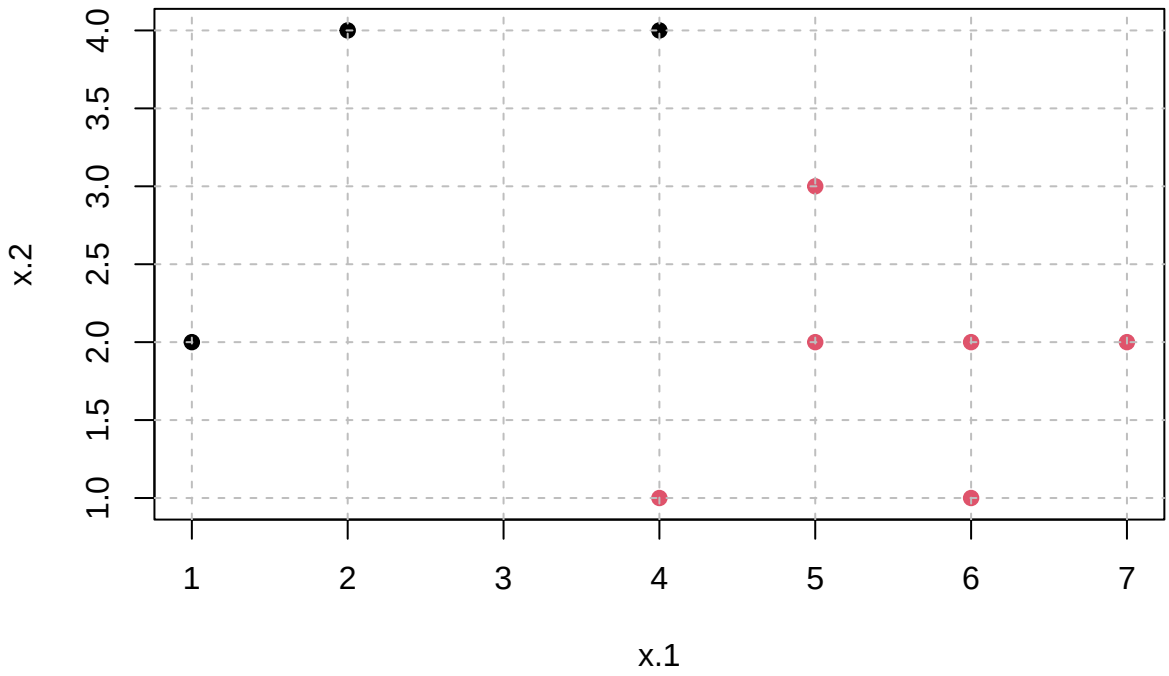
```
x.1=c(1,2,4,4,5,5,6,6,7)
x.2=c(2,4,4,1,3,2,1,2,2)
y=factor(c(rep(1,3),rep(2,6)))
my.data=data.frame(x.1,x.2,y)
print(my.data)
```

```
##   x.1 x.2 y
## 1   1   2 1
## 2   2   4 1
## 3   4   4 1
## 4   4   1 2
## 5   5   3 2
## 6   5   2 2
## 7   6   1 2
## 8   6   2 2
## 9   7   2 2
```

```
attach(my.data)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     x.1, x.2, y
```

```r
plot(x.2~x.1,col=y,pch=19,asp=1)
grid(nx = NULL, ny = NULL,
     lty = 2,      # Grid line type
     col = "gray", # Grid line color
     lwd = 1)
```



Answer following questions without using svm algorithm function in R.

(a) Find the optimal separating hyperplane define by the equation $-1 + X_1 + \beta_2 X_2 = 0$. Find β_2.

(b) Find all support vectors.

(c) If we add a new observation (x.1=1, x.2=4, y=1), would this affect the maximal margin classifier?

Now, use svm algorithm in R to answer the following question. *Use the option* `scale=FALSE` *for this question.*

(d) If we add a new observation (x.1=1, x.2=4, y=1), would this affect the maximal margin classifier? Add the observation to your dataset and see if it makes meaningful change.

Question 2 (SVM)

(a) Sketch the hyperplane $1 + 3X_1 - X_2 = 0$.

(b) For the given observations

```
set.seed(4)
x1=sample(1:10,5)
x2=sample(5:15,5)
print(data.frame(x1,x2))
```

```
##   x1 x2
## 1  8  7
## 2  3 10
## 3  9  9
## 4  7  6
## 5  4 15
```

indicate the set of points for which $1+3X_1-X_2>0$, as well as the set of points for which $1+3X_1-X_2<0$.

Question 3 (SVM)

We will use data found in (wisc-trn.csv) and (wisc-tst.csv) [check box folder for files] which contain train and test data respectively. 'This is a modification of the Breast Cancer Wisconsin (Diagnostic) dataset from the UCI Machine Learning Repository. Only the first 10 feature variables have been provided. (And these are all you should use.)

You should consider coercing the response (`class` variable) to be a factor variable.

(a) Fit a support vector classifier to the training data using cost = 0.01, with `class` as the response and the other variables as predictors. Use the `summary()` function to produce summary statistics, and describe the results obtained. Report training and test error rates.

(b) Use the `tune()` function to select an optimal `cost`. Consider values in range 0.01 to 10. If necessary, you can change the cost grid. Compute training and test error rates using this new value for `cost`.

(c) Repeat (b) using SVM with a radial kernel. Use default value for `gamma`.

(d) Repeat (b) using SVM with a polynomial kernel. Set `degree=2`. Hint: you can use `tune` function with `kernel='polynomial'` option. Students can also use other functions/options.

(e) Based on your analysis on (a) through (d) which approach seems to give the best results?

Question 4 (GAM: Poly)

In this question, we will conduct regression using `Boston` data from the `ISLR2` package.

```
set.seed(432)
trn.idx=sample(1:nrow(ISLR2::Boston),450)
tst.boston=ISLR2::Boston[-trn.idx,]
trn.boston=ISLR2::Boston[trn.idx,]
```

`nox` variable is your response variable. `dis` variable is your (only) predictor variable for this question.

(a) Use the `poly()` function to fit polynomials of degree 1 to 10. Plot polynomial fits.

(b) Using anova function to select the optimal degree for the polynomial, and explain your results. You do not have to pick one model, you can suggest multiple models or just explain why certain models are not ideal.

(c) Using 10-fold Cross-validation, select the optimal degree for the polynomial and explain your reason.

Question 5 (GAM)

In this question, we will conduct regression using `Boston` data from the `ISLR2` package.

```
set.seed(432)
trn.idx=sample(1:nrow(ISLR2::Boston),450)
tst.boston=ISLR2::Boston[-trn.idx,]
trn.boston=ISLR2::Boston[trn.idx,]
```

- `nox` variable is your response variable.
- You may choose your own predictor variables.

Come up with at least 3 generalized additive models, using variables of your own choice. Compare your models (on training data) and analyze which model works best.