Stat 427/627, Statistical Machine Learning

Homework 5

Due: Friday, June 21, 2024

Contents

- This assignment covers tree-based methods and support vector machines
- Finish Q.1 - Q.3 after Tuesday's class, and the rest after Thursday's class.
- 51 or 56 Points

Question	1	2	3	4	5	6	7	Total
427	6	4	22	3	6	10		51
627	6	4	22	3	6	10	5	56

1 Trees (Ex.8.4.4, p.362, 6 pts)

Refer to Figure 1.

(a) Draw a tree diagram corresponding to the partition of the predictor space shown the left-hand panel of Figure 1. The numbers inside the boxes indicate the mean of Y within each region.

(b) Look at the tree in the right-hand panel of Figure 1. Based on this tree, create a diagram similar to the left-hand panel of this figure where you divide the predictor space into the correct regions, and indicate the mean for each region.

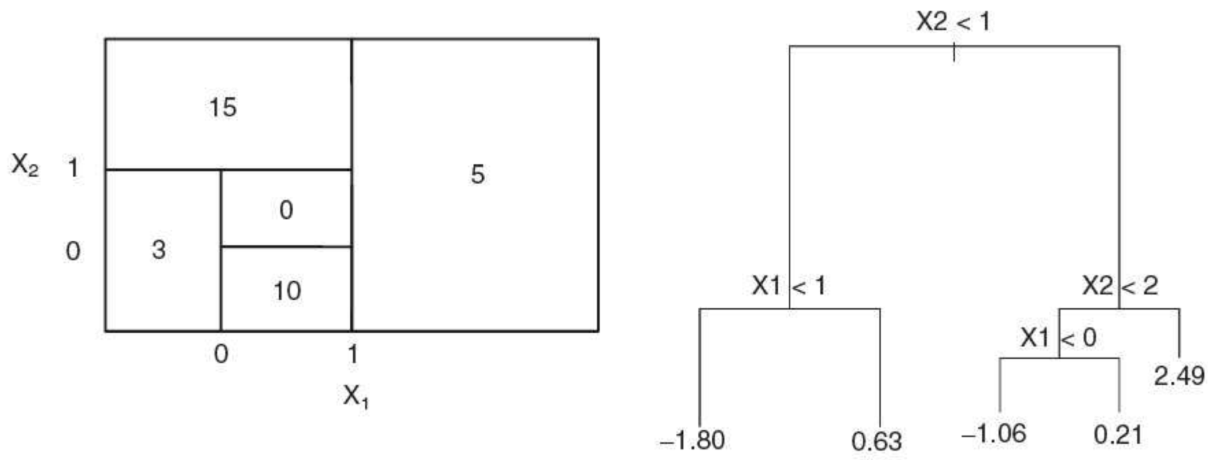


Figure 1: Tree Diagrams

2 Bagging (Ex.8.4.5, p.362, 4 pts)

Suppose we produce ten bootstrapped samples from a data set containing equal numbers of two classes red and green. We then apply a classification tree to each bootstrapped sample and, for a specific value of `X`, produce 10 estimates of $P(\text{Class is Red}|X)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, \text{ and } 0.75.$$

There are two common ways to combine these results together into a single class prediction.

- One is the majority vote approach.
- The second approach is to classify based on the average probability.

In this example, what is the final classification for each approach?

3 Orange Juice Purchases (Ex.8.4.9, p.363, 22 pts)

This problem involves the `OJ` (orange juice) data set which is part of the `ISLR2` package. To find its description, run the following codes.

```
library(ISLR2)
head(OJ)
? OJ
```

(a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations. Use `set.seed(2023)`.

(b) Use the `tree` package to fit a tree to the training data, with `Purchase` as the response and the other variables as predictors. Use the `summary()` to produce summary statistics about the tree, and describe the results obtained. Which variables are used in the tree? What is the training error rate? Do the variables make sense to you?

Remark: Several variables, such as `StoreID`, `Store7`, `STORE`, should be treated as factors. Furthermore, they are correlated. But since the tree-methods can select variables, we will use the predictors as is for now.

(c) Type in the name of your tree object to get a detailed text output. Pick one of the terminal nodes, and interpret the information displayed. How many decisions were made for that terminal node?

(d) Create a plot of the tree.

(e) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?

(f) Apply the `cv.tree()` function to the training set in order to determine the optimal tree size based on the misclassification rate. Use `set.seed(2023)`.

(g) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.

(h) Which tree size corresponds to the lowest cross-validated classification error rate?

(i) Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. ~~If cross-validation does not lead to selection of a pruned tree, then~~ **In addition**, create a pruned tree with **four** terminal nodes.

(j) Compare the training error rates between the pruned and unpruned trees. Which is higher?

(k) Compare the test error rates between the pruned and unpruned trees. Which is higher?

4 Hyperplane (line for 2-dimension) (Ex.9.7.1.a, 3 pts)

This problem involves a hyperplane in two dimensions. Sketch the hyperplane $1 + 3X_1 - X_2 = 0$. Indicate the set of points for which $1 + 3X_1 - X_2 > 0$, as well as the set of points for which $1 + 3X_1 - X_2 < 0$.

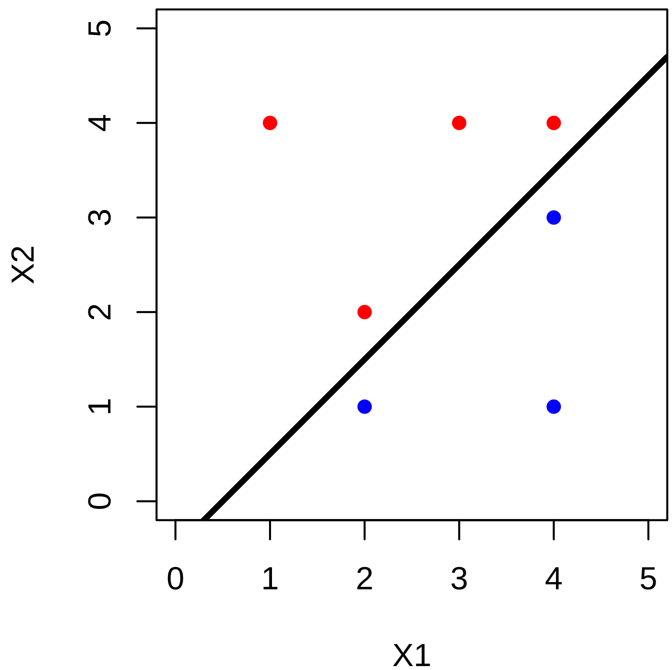
5 Maximal Margin Classifier (Ex.9.7.3, with revision, p.368, 6 pts)

Here we explore the maximal margin classifier on a toy data set.

Obs	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

The data and the optimal separating hyperplane, $X_1 - X_2 - 0.5 = 0$, are sketched in the scatter plot.

```r
X1 <- c(3,2,4,1,2,4,4)
X2 <- c(4,2,4,4,1,3,1)
Y <- c(rep("red",4), rep("blue",3))
par(pty="s")
plot(X1, X2, pch=16, col=Y, xlim=c(0,5), ylim=c(0,5), xlab="X1", ylab="X2")
abline(a=-0.5, b=1, lwd=3)
```



(a) Identify the support vectors (there are 4) for the maximal margin classifier. What is the margin of the classifier?

(b) Argue that a slight movement of the seventh observation $(X_1, X_2, Y) = (4, 1, Blue)$ would not affect the maximal margin hyperplane.

(c) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

6 SVM with application in Auto data (Ex.9.7.7, with revision, p.371, 10 pts)

In this problem, you will use support vector approaches in order to predict whether a given car gets high or low gas mileage based on the `Auto` data set in the `ISLR2` package.

(a) Create a binary variable that takes on a 1 for cars with gas mileage above the median, and a 0 for cars with gas mileage below the median.

(b) Fit a support vector classifier (i.e., SVM with linear kernel) to the data, in order to predict whether a car gets high or low gas mileage. Use the following variables in the `Auto` data set: `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, and `year`. Report the 10-fold cross-validation error. (You can use the default `cost` value.)

(c) Repeat (b), using SVMs with radial and polynomial basis kernels. Use the default values for `cost`, `gamma` and `degree` in the `svm()` function. Comment on your results.

(d) Make some plots to illustrate your results from (b) and (c). When $p > 2$, you can use the `plot()` function to create plots displaying one pair of variables at a time. For example, to plot `horsepower` and `year` the syntax is

```
plot(auto.svmlin, data=auto.data, horsepower ~ year)
```

7 Stat 627. Non-linear boundary in SVM (Ex.9.7.2, with revision, p.398, 5 pts)

We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$. We now investigate a non-linear decision boundary.

Suppose a classifier assigns an observation to the *BLUE* class if

$$(X_1 + 1)^2 + (X_2 - 2)^2 > 4$$

, and to the *RED* class otherwise.

(a) Sketch the boundary. Verify that the boundary is a circle. You do not need to attach the plot. Instead, report the following:

- The center of the circle. (Report in terms of $(X_1,\, X_2)$ coordinate.)
- The radius of the circle.
- Is the *BLUE* class inside or outside the circle?

(b) To what class the following observations are classified?

- (0, 0)
- (-1, 1)
- (2, 2)
- (3, 8)

—— **This is the end of HW 5.** ——