STATS 216 Introduction to Statistical Learning
Stanford University, Winter 2018

Problem Set 4

Due: Friday, March 16, 2018.
No late submissions will be accepted for this problem set, as announced in the syllabus.
Remember the university honor code. All work and answers must be your own.

---

1. This question relates to the plots in Figure 8.12 of your textbook *An Introduction to Statistical Learning*.

   (a) Sketch the tree corresponding to the partition of the predictor space illustrated in the left-hand panel of Figure 8.12. The numbers inside the boxes indicate the mean of Y within each region.

   (b) Create a diagram similar to the left-hand panel of Figure 8.12, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

2. **You may work in groups up to size 4 on this problem. If you do work in groups, write the names of all your group members on your problem set.**

   In this problem, you will generate simulated data, and then perform PCA and K-means clustering on the data.

   (a) Generate a simulated data set with 25 observations in each of three classes (i.e. 75 observations total), and 45 variables.

   Hint: There are a number of functions in `R` *that you can use to generate data. One example is the* `rnorm()` *function;* `runif()` *is another option. Be sure to add a mean shift to the observations in each class so that there are three distinct classes.*

   (b) Perform PCA on the 75 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.

   (c) Perform K-means clustering of the observations with $K = 3$. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

   Hint: You can use the `table()` *function in* `R` *to compare the true class labels to the class labels obtained by clustering. Be careful how you interpret the results: K-means clustering will arbitrarily number the clusters, so you cannot simply check whether the true class labels and clustering labels are the same.*

(d) Perform K-means clustering with $K = 2$. Describe your results.

(e) Now perform K-means clustering with $K = 4$, and describe your results.

(f) Now perform K-means clustering with $K = 3$ on the first two principal component score vectors, rather than on the raw data. That is, perform K-means clustering on the 75×2 matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.

(g) Using the `scale()` function, perform K-means clustering with $K = 3$ on the data *after scaling each variable to have standard deviation one.* How do these results compare to those obtained in (c)? Explain.

3. **You may work in groups up to size 4 on this problem. If you do work in groups, write the names of all your group members on your problem set.**

Recall the `body` dataset from problem 4 of Homework 3. In that problem we used PCR and PLSR to predict someone's weight. Here we will re-visit this objective, using bagging and random forests. Start by setting aside 200 observations from your dataset to act as a test set, using the remaining 307 as a training set. Ideally, you would be able to use your code from Homework 3 to select the same test set as you did on that problem.

Using the `randomForest` package in `R` (hint: see section 8.3.3 in the textbook for guidance), use Bagging and Random Forests to predict the weights in the test set, so that you have two sets of predictions. Then answer the following questions:

(a) Produce a plot of test MSE (as in Figure 8.8 in the text) as a function of number of trees for Bagging and Random Forests. You should produce one plot with two curves, one corresponding to Bagging and the other to Random Forests. *Hint: If you read the documentation for the* `randomForest()` *function, you can find a way to obtain the data for both curves with only one call each to the* `randomForest()` *function.*

(b) Which variables does your random forest identify as most important? How do they compare with the most important variables as identified by Bagging?

(c) Compare the test error of your random forest (with 500 trees) against the test errors of the three methods you evaluated in problem 4(f) on Homework 3. Does your random forest make better predictions than your predictions from Homework 3?

If you did not successfully solve problem 4(f) on Homework 3, you may compare the test error of your random forest against the test errors in the Homework 3 solutions.

(d) The `randomForest()` function uses 500 as the default number of trees. For this problem, would it be valuable to include more trees? How can you tell?

4. In this problem, we will explore the maximal margin classifier on a toy data set. You should feel free to make use of R in carrying out any of the tasks below.

   (a) We are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label. Plot or sketch the observations.

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

   (b) Plot or sketch the separating hyperplane with maximum margin, and provide the equation for this hyperplane (of the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$).

   (c) Describe the classification rule for the maximal margin classifier. It should be something along the lines of "Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify to Blue otherwise." Provide the values for β_0, β_1, and β_2.

   (d) On your plot or sketch, indicate the margin for the maximal margin hyperplane. How wide is the margin?

   (e) Indicate the support vectors for the maximal margin classifier.

   (f) Would a slight movement of the seventh observation affect the maximal margin hyperplane? Why or why not?

   (g) Sketch a hyperplane that is *not* the optimal separating hyperplane, and provide the equation for this hyperplane.

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