STATS 216v – Introduction to Statistical Learning
Stanford University, Summer 2017

Problem Set 4

Due: August 16, 2017

Remember the university honor code. All work and answers must be your own.

1. In this problem we investigate a simple example of hierarchical clustering. Problem 4 will use this method on a real data set.

   Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

   For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observation is 0.8.

   (a) Based on this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using **complete linkage**. Indicate on the plot the height at which each fusion occurs and the observations corresponding to each leaf of the dendrogram.

   (b) Repeat part (a) using **single linkage** clustering

   (c) Suppose that we cut the dendrogram from (a) in such a way that results in two clusters. Which observations are in each cluster?

   (d) Suppose that we cut the dendrogram from (b) in such a way that results in two clusters. Which observations are in each cluster?

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

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

Obs.	X_1	X_2	Y
1	13.58	17.50	blue
2	9.60	9.60	blue
3	17.60	17.60	blue
4	5.56	17.60	blue
5	9.60	5.60	red
6	17.60	13.60	red
7	17.58	5.60	red

(b) Sketch the optimal separating hyperplane, 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 Green otherwise." Provide the values for β_0, β_1, and β_2.

(d) On your 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) Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.

(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.

3. This problem involves the `OJ` data set which is part of the `ISLR` package.

(a) Create a training set containing a random sample of 535 observations, and a test set containing the remaining observations. Use the commands `set.seed(2017);` `train = sample(1:nrow(OJ), 535); test = setdiff(1:nrow(OJ), train)`.

(b) Fit a (linear) support vector classifier to the training data using `cost=1`, with `Purchase` as the response and the other variables as predictors. Use the `summary()` function to produce summary statistics about the SVM, and describe the results obtained.

(c) What are the training and test error rates?

(d) Use the `tune()` function to select an optimal `cost`. Consider values in the range 0.01 to 10.

(e) Compute the training and test error rates using this new value for `cost`.

(f) Repeat parts (b) through (e) using a support vector machine with a radial kernel. Use the default value for `gamma`.

(g) Repeat parts (b) through (e) using a support vector machine with a polynomial kernel of degree 2.

(h) Repeat parts (b) through (e) using a linear support vector machine, applied to an expanded feature set consisting of linear and all possible quadratic terms for the predictors. How does this compare to the polynomial kernel both conceptually and in terms of the results for this problem?

(i) Overall, which approach seems to give the best results on this data?

4. In this problem we will explore K-means and hierarchical clustering on a wheat seed data set. The data file `SeedData.csv` is available on the class website. There are $n = 210$ seeds with $p = 7$ real valued features for each observation. The data set is from Lichman, M. (2013) from the UCI Machine Learning Repository [1]. You can find a full description of the data at `http://archive.ics.uci.edu/ml/datasets/seeds#`.

   (a) Set `set.seed(2017)`. Using the `kmeans()` function, perform K-means clustering (using all 7 features) with $K = 3$ and 20 random starting positions. Plot the data in terms of the variables in the 2nd and 6th columns, with the points colored according to the clusters you just obtained.

   (b) Using the `hclust()` function, perform single linkage hierarchical clustering on the data. Plot the corresponding dendrogram.

   (c) Using the `cutree()` function, make a cut on the tree that splits the data into 3 clusters. Plot the data again in terms of the variables in the 2nd and 6th columns, coloring the points by the clusters you just obtained. How does it compare to the result you got using k-means clustering?

   (d) Using the `hclust()` function, perform complete linkage hierarchical clustering on the data. Plot the corresponding dendrogram. What do you see? If we wanted to cluster our data into 3 clusters, do you prefer single linkage clustering or complete linkage clustering in this example? Why?

   (e) It turns out that there are three different types of wheat seeds: Kama, Rosa, and Canadian. If you had access to the class labels, briefly describe one way to build a model for classifying wheat seeds. Limit your answer to 1 sentence.

5. This problem uses the `ALS` data set. Recall that in previous problem sets we tried to predict the rate of progression of ALS in patients using linear regression, the Lasso, decision trees, boosting and random forests, and saw varying values of the test RMSE. Now we will see if we can improve on these results by first preprocessing the data in an unsupervised way.

   (a) Use the `prcomp()` function to run a principal components analysis on the training set.

   (b) Use the standard deviations to calculate the proportion of variance explained (PVE) by each principal component. Plot the PVE and the cumulative PVE.

[1] `http://archive.ics.uci.edu/ml`

(c) Use the `pcr()` function in the `pls` library to run PCR on the training set using cross-validation; pass along the parameter `ncomp = 100` and call your object `pcr.fit`. Plot the cross-validation RMSE via the function `validationplot()`.

(d) What is the number of components that minimize the RMSE? You can extract the errors for each component via `RMSEP(pcr.fit)$val[2,1,1:101]`.

(e) Compute the test set RMSE using the number of components you found in the previous item. How does this compare to our previous results? Recall that the test RMSE for linear regression was 0.7527, for the Lasso it was 0.5209, for Boosted Trees it was 0.5115 and for Random Forests it was 0.5123.