Homework 10

STAT 430, Spring 2017

Due: Friday, April 28 by 11:59 PM

Please see the homework instructions document for detailed instructions and some grading notes. Failure to follow instructions will result in point reductions.

This is the last homework. :(

Exercise 1

[**10 points**] For this question we will return to the `OJ` data from the `ISLR` package. We will again attempt to predict the `Purchase` variable. After changing `uin` to your `UIN`, use the following code to test-train split the data.

```r
library(ISLR)
library(caret)
uin = 123456789
set.seed(uin)
oj_idx = createDataPartition(OJ$Purchase, p = 0.5, list = FALSE)
oj_trn = OJ[oj_idx,]
oj_tst = OJ[-oj_idx,]
```

(a) Tune a SVM with a linear kernel to the training data using 5-fold cross-validation. Use the following grid of values for `C`. Report the chosen values of any tuning parameters. Report test accuracy.

```r
lin_grid = expand.grid(C = c(2 ^ (-5:5)))
```

(b) Tune a SVM with a polynomial kernel to the training data using 5-fold cross-validation. Do not specify a tuning gird. (`caret` will create one for you.) Report the chosen values of any tuning parameters. Report a confusion matrix for the test data as well as the test accuracy.

(c) Tune a SVM with a radial kernel to the training data using 5-fold cross-validation. Use the following grid of values for `C` and `sigma`. Report the chosen values of any tuning parameters. Report test accuracy.

```r
rad_grid = expand.grid(C = c(2 ^ (-2:3)), sigma  = c(2 ^ (-3:1)))
```

(d) Tune a random forest using 5-fold cross-validation. Report the chosen values of any tuning parameters. Report test accuracy.

(e) Summarize the accuracies above. Which method performed the best?

Exercise 2

[**10 points**] For this question, use the data found in `clust_data.csv`. We will attempt to cluster this data using k-means. But, what k should we use?

(a) Apply k-means to this data 15 times, using number of centers from 1 to 15. Each time use `nstart = 10` and store the `tot.withinss` value from the resulting object. (Hint: write a for-loop.) The `tot.withinss` measures how variable the observations are within a cluster, which we would like to be low. So obviously this value will be lower with more centers, no matter how many clusters there truly are. Plot this value against the number of centers. Look for an "elbow", the number of centers where the improvement suddenly drops off. Based on this plot, how many cluster do you think should be used for this data?

(b) Re-apply k-means for your chosen number of centers. How many observations are placed in each cluster? What is the value of `tot.withinss`?

(c) Visualize this data. Plot the data using the first two variables and color the points according to the k-means clustering. Based on this plot, do you think you made a good choice for the number of centers? (Briefly explain.)

(d) Use PCA to visualize this data. Plot the data using the first two principal components and color the points according to the k-means clustering. Based on this plot, do you think you made a good choice for the number of centers? (Briefly explain.)

(e) Calculate the proportion of variation explained by the principal components. Make a plot of the cumulative proportion explained. How many principal components are need to explain 95% of the variation in the data?

Exercise 3

[**10 points**] For this question we will return to the `USArrests` data from the notes. (This is a default `R` dataset.)

(a) Perform hierarchical clustering six times. Consider all possible combinations of linkages (average, single, complete) and data scaling. (Scaled, Unscaled.)

Linkage	Scaling
Single	No
Average	No
Complete	No
Single	Yes
Average	Yes
Complete	Yes

Each time, cut the dendrogram at a height that results in four distinct clusters. Plot the results, with a color for each cluster.

(b) Based on the above plots, do any of the results seem more useful than the others? (There is no correct answer here.) Pick your favorite. (Again, no correct answer.)

(c) Use the documentation for `?hclust` to find other possible linkages. Pick one and try it. Compare the results to your favorite from **(b)**. Is it much different?

(d) Use the documentation for `?dist` to find other possible distance measures. (We have been using `euclidean`.) Pick one (not `binary`) and try it. Compare the results to your favorite from **(b)**. Is it much different?