Assignment 4

CS780/880: Introduction to Machine Learning

Due: By 12:40PM Thu Apr 6th, 2017
Submission: Turn in as a PDF on myCourses, or printed and turned in at class
Discussion forum: https://piazza.com/unh/spring2017/cs780cs880

Problem 1 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 20 observations in each of three classes (i.e. 60 observations total), and 50 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 60 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 on 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 60×2 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 (b)? Explain.

Problem 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 (*by hand!*) 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) Sketch (*by hand!*) the optimal separating hyperplane, and provide the equation for this hyperplane (of the form (9.1)).

(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 sketch, indicate the margin for the maximal margin hyperplane.

(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 hyper-plane, 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.

Problem 3 Now we try to solve a linear regression problem directly using linear algebra using the Auto dataset.

(a) Compute linear regression with mpg as the response and all the other variables except name as predictors. Do not forget about the intercept. Do not invert the design matrix.

(b) Compute the RSS of your solution and try to use linear algebra.

(c) Now use the lm command and compare the results: both the coefficients β and RSS.

Problem 4 In this problem, we explore the impact of *priors* on decisions. The question is as follows: "I put two randomly chosen sums of money x and $2 \cdot x$ in two sealed indistinguishable envelopes. I choose *one* of the envelopes randomly and give it to you. You may decide to keep the money in your envelope, or you can switch the envelopes. You cannot inspect the content inside before deciding to switch. Should you switch?"

(a) Compute the expected gain from switching the two envelopes. Let z be the amount in your envelope. What is your expectation for the amount in my envelope?

(b) If you had a chance to switch the envelopes again, would you do it? How much would you expect to get now compared with z?

(c) Discuss your results.

(d) Repeat part (a) assuming that you know that I only used \$150 dollars total. (That is x is chosen randomly from the uniform distribution between \$0 and \$50.)

(e) Are the results of (a) and (d) different. Why?