STATS 415 Homework 8

Due by 2:30pm on Dec 3, 2019

1. Suppose you have a dataset with two predictors and perform PCA by carrying out an eigen-decomposition on the covariance matrix of the data (the two predictors are on the same scale). You find that its two eigenvalues are $\lambda_1 = 4$ and $\lambda_2 = 1$, and their corresponding eigenvectors are

$$u_1 = \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix} \quad u_2 = \begin{bmatrix} -0.8 \\ 0.6 \end{bmatrix}$$

   (a) Reconstruct the covariance matrix of the data. (15 points)

   (b) What percentage of variance is explained by the first principal component? (10 points)

   (c) For a new data point $X = (1, 2)$, find its scores on the first and second principal components. (15 points)

2. This exercise continues Q2 of HW 7 and Q3 of HW 6. Use the same training and test datasets. The goal is to predict the acceptance rate from the other variables in the `College` data set. (20 points for each question)

   (a) Perform Principal Component Analysis on the predictors. Explain why you chose to standardize or not standardize the predictors first. Make a scree plot of the eigenvalues. How many eigenvalues does one need to explain 95% of the variance in the data? Report loadings of the first two PCs. Interpret them if you can.

   (b) Fit a PLS model on the training set, with the number of principal components K chosen by cross-validation. Report the training and test error obtained, along with the value of K selected.

   (c) Comment on the results obtained, including also the methods from HW 6 and 7. Which approach would you recommend for this dataset and why?

Please limit your answer to Q2 to 6 pages, organized into a coherent typed data analysis report. Answers to Q1 may be either typed or handwritten. Please staple everything together and clearly write your name, your UMID, and your GSI/lab number on the homework.