MATH 189 Homework 5

Due Feb 12th, 2022

Q1. Places Rated Almanac rated 329 communities in the United States according to the following nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care & the Environment
4. Crime
5. Transportation
6. Education
7. The Arts
8. Recreation
9. Economics

The rating results can be found in Places_Rated.txt. In this dataset, the first 9 columns represent the above 9 variables. The 10th column is the index of communities, ranging from 1 to 329. Note that, except for housing and crime, the higher the score is the better condition the community has. Analyze this dataset according to the following steps.

(a). Calculate the eigenvalues and eigenvectors of the covariance matrix of **standardized** data (each column has mean 0 and variance 1). Calculate the proportion of total variance explained, by each eigenvector and the cumulative proportion of total variance explained by the first k (=1,…,9) eigenvectors. Report your results in scree plot and cumulative plot. Next,

repeat the above steps to raw data, and draw scree plot and cumulative plot. Compare the results obtained from raw and standardized data.

(b). Apply principal component analysis to the standardized data. Choose the number of principal components (k) according to the scree plot you obtained in Part 1. Report the corresponding principal component loading vectors. Visualize the dataset by projecting the observations onto the plane spanned by the first two principal components.

Q2. Consider the Weekly dataset, which is included in the ISLR package. This dataset consists of 1089 weekly percentage returns for the S&P 500 stock index over 21 years, from the beginning of 1990 to the end of 2010. It contains the following 9 variables.

Year: The year that the observation was recorded.

Lag1: Percentage return for previous week.

Lag2: Percentage return for 2 weeks previous.

Lag3: Percentage return for 3 weeks previous.

Lag4: Percentage return for 4 weeks previous.

Lag5: Percentage return for 5 weeks previous.

Volume: Volume of shares traded (average number of daily shares traded in billions).

Today: Percentage return for this week.

Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week.

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

(b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as covariates/predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).