**DSCI 343: Introduction to Data Analysis** Fall 2019

ASSIGNMENT 3: CLUSTERING AND PREDICTIVE MODELING

Instructor: Mehmet Koyutürk, TA: Serhan Yılmaz Due: December 6 before 11:59 PM

# 1 Problem 1

In this exercise, our aim is to predict a quantitative outcome using a set of feature variables. For this purpose, we will build a linear model using the training sample and validate the prediction performance on the test sample. We will use the following datasets that are provided with the assignment:

- The file "p1_features_training.csv" contains a matrix $X \in \mathbb{R}^{60 \times 10}$ which contains 60 observations and 10 variables. This is our training sample. We will use these 10 variables as features to build a linear model to predict an outcome variable.
- The file "p1_outcome_training.csv" contains a vector $Y \in \mathbb{R}^{60 \times 1}$ having 60 observations. This variable is the outcome variable that we are going to predict. Again, this is our training sample. We will use the observations here to learn how to predict the outcome variable using the provided feature variables.
- The file "p1_features_test.csv" contains a matrix $X \in \mathbb{R}^{40 \times 10}$ which contains 40 observations and 10 features. The observations belong to the same 10 features given in "p1_features_training.csv", but this is our test sample. We will use the observations in this file validate our model.
- The file "p1_outcome_test.csv" contains a vector $Y \in \mathbb{R}^{40 \times 1}$ having 60 observations. These observations belong to the same outcome variable given in "p1_outcome_test.csv", but this is our test sample. We will use the observations in this file to validate our model.

## 1.1 Part a

Use the training observations of the first feature (denoted $X_1$) given in "p1_features_training.csv" and outcome variable (denoted $Y$) given in "p1_outcome_training.csv" to build a linear model:

$$Y = \beta_0 + \beta_1 X_1 \tag{1}$$

Estimate the parameters $\beta_0$ and $\beta_1$ using ordinary least squares (OLS) loss function:

$$\min \sum_i (Y^{(i)} - \hat{Y}^{(i)})^2 \tag{2}$$

where $\hat{Y}^{(i)}$ is the predicted outcome for the $i$th observation according to Equation 1.

After finding the $\beta$ parameters, predict the outcome on the training sample and assess the model predictivity using the $R^2$ measure (also known as coefficient of determination). Next, predict the outcome on the test sample (using the observations in "p1_features_test.csv") and assess the model predictivity using $R^2$ on the test sample by comparing with the true outcome values (given in "p1_outcome_test.csv"). Which one is higher (training $R^2$ vs. test $R^2$)? Discuss why.

### 1.2 Part b

Repeat (a) for 10 different models where model $k$ uses the first $k$ features (instead of only the first feature) to predict the outcome. Assess the model predictivity using $R^2$ for each of the 10 models on the training sample as well as on the test sample. Plot the prediction performance $R^2$ as a function of number of features in the model. Which model (thus, number of variables) seem to the best according to $R^2$ on the training sample? Which model seem to the best according to $R^2$ on the test sample? Do they agree? Comment on why they agree (or not). If they do not agree, which one do you think is the most reliable?

Note that, the linear model for model $k$ is:

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j \tag{3}$$

## 2 Problem 2

In this exercise, our aim is to analyze a high dimensional dataset using dimensionaly reduction techniques and unsupervised learning. For this purpose, we will use the following congressional votes dataset that is provided with the assignment:

- The file “p2_congress_1984_votes.csv“ contains a matrix $X \in \{-1, 0, 1\}^{435 \times 16}$ indicating the votes of 435 U.S. House of Representatives congress members on 16 key issues in the congress of 1984. Here, -1, 0 and 1 denote *reject*, *neutral* and *accept* votes respectively.
- The file “p2_congress_1984_party_affiliations.csv“ contains a vector $Y$ of size 435 x 1 indicating the party affiliations (Republican or Democrat) of 435 congress members in the congress of 1984. See data source for more information.

### 2.1 Part a

On the congressional votes dataset given in “p2_congress_1984_votes.csv“, apply principal component analysis (PCA). Plot the cumulative variance explained by top $k$ principal components (with highest eigenvalues) as a function of $k$ (see ”example_pca_figure.png”). How many principal components do you think are enough to sufficiently summarize the data (according to the explained variance)? Next, project the data onto the first 3 principal components with highest eigenvalues. For each of the three principal component pairs (PC1-PC2, PC1-PC3, PC2-PC3), draw a scatter plot of congress members colored according to their party affiliations (as given in “p2_congress_1984_party_affiliations.csv“). Which of the principal component pair separates the congress members best according to their party affiliations? Are the congress members with the same party affiliation seem to be clustered according to their votes on the congress?

### 2.2 Part b

Use your favorite clustering algorithm (e.g., k-means, k-medoids) to cluster the congress members into two groups based on their congress votes on 16 issues. Visualize these groups with scatter plots on the first two top principal components you identified in part (a). Are these groups seem visually separated? How much do they agree with the party affiliations? Make sure to quantify the agreement of the clusters with the party affiliations (e.g., with accuracy or $f_1$ score). Now, repeat the clustering analysis using the first two principal components (instead of all 16 votes). Again, quantify the agreement of the clusters with the party affiliations. Which clustering (using principal components vs. using all 16 votes) agrees with the party affiliations more? Comment on why this might be the case.

*Note*: Make sure to briefly explain the clustering algorithm and the distance function that you use to cluster the congress members.