DS 303 Homework 12
Due: Dec. 04, 2023 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word file, or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Concept Review

(a) Any time clustering is performed on a dataset, we will obtain clusters - even if there are truly no clusters in the dataset. What we really want to know if whether the clusters we have found represent *real* subgroups in the data or whether they are simply the result of *clustering the noise*. Use the statistical knowledge you've accumulated this semester to propose an approach to validate the clusters we obtain from a dataset.

(b) Suppose we perform hierarchical clustering using single linkage and using complete linkage for a given dataset. We obtain two dendrograms;

i. At a certain point on the single linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ fuse. On the complete linkage dendrogram, the clusters $\{1, 2, 3\}$ and $\{4, 5\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

ii. At a certain point on a single linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ fuse. On the complete linkage dendrogram, the clusters $\{5\}$ and $\{6\}$ also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

(c) The variance explained by the mth principal component is the defined as:

$$\frac{1}{n}\sum_{i=1}^{n} z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2.$$

Therefore the proportion of variance explained (PVE) is

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2} = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}x_{ij}\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2}$$

On the `USArrests` data, calculate PVE (for all 4 principal components) in two ways:

i. Using the `sdev` output from the `prcomp()` function as we did in class.

ii. By applying the above formula directly. That is, use the `prcomp()` function to compute the principal component loadings (ϕ) and plug them into the formula to obtain the PVE.

These two approaches should give the same results.

(d) Let's revisit least squares using SVD. Note that any matrix X can be decomposed into U, V and D such that $X = UDV^T$. This decomposition is known as SVD (singular value decomposition). Prove that the following identities are true:

i. $(X^TX)^{-1} = VD^{-2}V^T$

ii. $X(X^TX)^{-1}X^T = UU^T$ (this is known as the Hat matrix and is a projection matrix that projects Y into the column space of X)

iii. Rewrite $\hat{\beta}$ in terms of U, V, D, and Y.

iv. Explain why rewriting $\hat{\beta}$ in terms of U, V, D, and Y has computational advantages. In fact, `lm()` uses matrix decomposition to speed up computing this step - which makes the function incredibly scalable for large datasets.

Problem 2: Simulations for Unsupervised Learning

(a) Generate a simulated dataset with 20 observations in each of 3 classes (for 60 observations total) and 50 features using the following code

```
set.seed(1)
set.seed(1)
c1 = c2 = c3 = matrix(NA,nrow=20,ncol=50)
for(i in 1:20){
  c1[i,]= rnorm(50,2,1.5)
  c2[i,] = rnorm(50,3,1.5)
  c3[i,] = rnorm(50,4,1.5)
}

data = rbind(c1,c2,c3)
```

(b) Perform PCA on the 60 observations. What is the PVE for the first two principal components?

(c) Your colleague, who misunderstands PCA, is surprised by the PVE in part (b). They point out that we've generated data such that each observation's 50 features come from a normal distribution (with only mean shift). We **did not** create a dataset where only a few dimensions are informative. Therefore, all of the dimensions should be equally interesting. Given this setup, your colleague does not understand why the first two principal components can explain so much of the variance in the dataset. Explain in plain language why your colleague's understanding of PCA is wrong.

(d) Plot the first two principal component score vectors. Use a different color to indicate observations in each of the three classes. Looking at the plot, are the classes well-separated?

(e) 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 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.

(f) Perform K-means clustering with $K = 2$. Describe your results.

(g) Perform K-means clustering with $K = 4$. Describe your results.

(h) 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. How do your results compare to using the raw data? Discuss.

Problem 3: Matrix Completion

Write an `R` function to perform matrix completion using the algorithm and example code outlined in lecture. In each iteration, the function should keep track of the relative error, as well as the iteration count. Iterations should continue until the relative error is small enough or until some maximum number of iterations is reached (set a default value for this maximum number). Furthermore, there should be an option to print out the progress in each iteration.

Test your function on the `Boston` data (part of `ISLR2 library`). First, standardize the features to have mean zero and standard deviation one using the `scale()` function. Run an experiment where you randomly leave out an increasing (and nested) number of observations from 5% to 30%, in steps of 5%. Apply the matrix completion algorithm with $M = 1, 2, \ldots, 8$. Plot the approximation error as a function of the fraction of observations that are missing, and the value of M, averaged over 10 repetitions of the experiment.

Problem 4: Hierarchical clustering and Classification

The dataset `gene.csv` consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group. Load in the data using `read.csv()`. You will need to select `header = F`.

(a) Apply hierarchical clustering to the samples using correlation-based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used? Discuss and report your results.
To implement the correlation-based distance, use the function `cor()` and then pass it through the `dist()` function: `distM = dist(cor(data)`.

(b) Technically, we can covert this to a classification problem where the label $Y = 0$ if the tissue is healthy and $Y = 1$ if the tissue is diseased. The measurements on 1,000 genes will be used as our predictors. Can logistic regression, LDA, or QDA be applied here? Explain why logistic/LDA/QDA will fail here.

(c) Convert the setting to a classification problem and implement a solution that allows us to fit logistic regression to the dataset. Make sure `Y` is stored as a `factor()`. Justify your approach. Report the confusion matrix and misclassification error rate. Since we have a limited sample size, you do not need to split the data into a training and test set.
Hint: PCA.

End of assignment