CS 534: Machine Learning

Homework #5

Due: April 18th at 11:59 PM on Canvas

Submission Instructions: The homework should be submitted electronically on Canvas. The code can be done in Python, R, or Matlab with the hints mostly focusing on those who will be programming in Python. Also be sure to attach the code to ensure you'll get partial / full credit if the answer is not what was expected.

1. **(5 + 3 + 2 = 10 pts)** (**Illustrating the "curse of dimensionality"**)

   For a hypersphere of radius a in d dimensions, the volume is related to the surface area of a unit hypersphere (S) as

   $$V = \frac{S \times a^d}{d}.$$

   (a) Use this result to show that the fraction of the volume which lies at values of the radius between $a - \epsilon$ and a, where $0 < \epsilon < a$, is given by $f = 1 - (1 - \epsilon/a)^d$. Hence, show that for any fixed ϵ, no matter how small, this fraction tends to 1 as $d \to \infty$.

   (b) Evaluate the ratio f numerically by plotting the results for different values of $\epsilon/a = 0.01$ and $d = 1$, 10, 100, and 1000.

   (c) What conclusions can you draw from the plot.

2. **(5+5+5+15 = 30 points) k-Means**

   We will be using the famous Iris dataset and trying to cluster the dataset to find any similar groups of flowers. However, we are not certain that the scales of the various features are well-suited for clustering. Since we have the labels, we will use them to determine the quality of our clusters:

   $$\text{purity}(C_i) = \frac{1}{|C_i|} \max_j (|C_i|_{\text{class}=j}),$$

   $$\text{total purity} = \sum_{i=1}^{k} \frac{|C_i|}{|D|} \text{purity}(C_i),$$

   where $|C_i|$ is the total number of data points assigned to cluster C_i, $|C_i|_{\text{class}=j}$ is the number of data points from class j assigned to cluster C_i, and $|D|$ is is the total number of data points in the dataset.

   - Cluster the data into 3 clusters using K-means and calculate the cluster purity for your solution. Report the cluster purity and also mention how you seeded the algorithm.

- Now linearly scale each feature so that the values range from 0 to 1. Cluster the data using the k-means algorithm as before and calculate the cluster purity for the clustering. Report the calculations you used to scale the features as well as the cluster purity.
- Linearly scale the original dataset features so that the distribution of values for each feature has a mean of 0 and a standard deviation of 1. Cluster the data as before and report the cluster purity obtained.
- Now repeat the above 3 steps for K = {4, 5, 6}. Based on the results, what preprocessing would you use and what is the "optimal" value of K based on purity?

3. **(20 pts) Agglomerative Clustering**

   Run agglomerative clustering using single, average, complete, and ward linkage on the iris dataset (You might want to use the hierarchical clustering package in `scipy`.) What distance metric did you use? Plot the dendrograms and for each dendrogram, visually inspect it to suggest what value(s) of k (between 1 and 6) that seem reasonable to choose for this dataset.

4. **(15 pts) Gaussian Mixture Models**

   Learn a Gaussian Mixture Model from the iris data with K = 3, 4, 5, 6. How does the cluster purity compare to k-Means?

5. **(3 + 5 + 5 + 2 + 10 = 25 pts) PCA & NMF**

   Load the college dataset `Colleges.txt` provided.

   (a) Preprocess the data by removing missing data and properly dealing with categorical data.

   (b) Run PCA on this processed data. Report how many components were needed to capture 95% of the variance in the normalized data. Discuss what characterizes the first 3 principal components (i.e., which original features are important).

   (c) Normalize the data (where applicable) and run PCA on the normalized data. Report how many components were needed to capture 95% of the variance in the normalized data. Discuss what characterizes the first 3 principal components (i.e., which original features are important).

   (d) Discuss why you should normalize the data before performing PCA.

   (e) Run NMF on the normalized data using $R = 3$. Discuss what characterizes the 3 components. How much variance does it capture?