Homework 7: Unsupervised Learning **CSCE 633**
Due: 11:59pm on Jul 23, 2024

Instructions for homework submission
a) Please write a brief pdf report including your experiment details, results and discussion; **submit the pdf file on Canvas**. Also **submit a jupyter notebook** including your code.
b) Make sure TAs can run your code on Google Colab. For each question, please explain your thought process, results, and observations in a markdown cell after the code cells. Please do not just include your code without justification.
c) **You can use any available libraries for this homework.**
d) Please start early :)

Question: Unsupervised Learning

We use Wine dataset, which contains chemical analysis results of wines grown in the same region in Italy but derived from three different types.

The analysis determined the quantities of 13 constituents found in the three types of wines: (1) Alcohol, (2) Malic acid, (3) Ash, (4) Alcalinity of ash, (5) Magnesium, (6) Total phenols, (7) Flavanoids, (8) Nonflavanoid phenols, (9) Proanthocyanins, (10) Color intensity, (11) Hue, (12) OD280/OD315 of diluted wines, (13) Proline.

(a) Data Exploration

(1) Load data from 'wine.data' file and delete the label column, since we only need the 13 features for unsupervised learning.

(2) Plot the distribution histograms of the 13 variables. Provide a brief discussion on your intuition regarding the variables and the resulting histograms.

(3) Visualize the correlation matrix of all the variables with a *seaborn* heatmap and discuss potential associations between the variables.

(4) Consider data normalization to avoid artificially assigning higher importance to features of larger range.

(b) K-Means Clustering Use the k-means clustering algorithm to cluster the wine data based on all the variables.

(1) Choosing Number of Cluster Experiment with different numbers of clusters and use the Elbow method to identify the optimal number of clusters.

(2) K-Means Results Report the number of wines that were assigned to each cluster, the centroid of each cluster.

(c) Gaussian Mixture Model Use the GMMs to cluster the wine data based on all the variables.

(1) Choosing Number of Cluster The number of Gaussian mixtures can be close to

the optimal number of clusters found above. To make sure, you can use Akaike information criterion (AIC) and Bayesian information criterion (BIC) to select the number of clusters.

(2) GMM Results

- Report the mean and covariance for each Gaussian mixture component and discuss your findings. Note: You can use a heatmap to visualize the covariance matrices of the GMM, instead of printing their actual values.
- Compute the log-likelihood of each data sample belonging to the GMM. Plot and discuss the histogram of the resulting log-likelihood values.

(d) Principal Component Analysis Use PCA to reduce the dimensions of original wine dataset.

(1) Choosing Number of Principal Component First set number of components to 6, then check the explained variance ratios of the selected components. For the convenience of visualization, you may want to reduce the number of principal components base on your findings.

(2) PCA before Cluster Choose the top two components and make 2D scatter plots with colored labels predicted by both K-Means and GMM methods. The axes of plots are the first two principal c omponents. Use *m atplotlib.patches.Ellipse* to show how K-Means and GMM cluster the data.

(e) Evaluation Finally, use Silhouette score to evaluate the k-means clustering, Gaussian mixture model, PCA + k-means and PCA + GMM methods.