Statistical Methods in AI

Instructor : Prof Ravi Kiran Sarvadevabhatla

Deadline : 25 September 2023 11:55 P.M

Assignment - 2

General Instructions

- Your assignment must be implemented in Python.
- While you're allowed to use ChatGPT for assistance, you must explicitly declare in comments the prompts you used and indicate which parts of the code were generated with the help of ChatGPT.
- Plagiarism will only be taken into consideration for code that is not generated by ChatGPT. Any code generated with the assistance of ChatGPT should be considered as a resource, similar to using a textbook or online tutorial.
- The difficulty of your viva or assessment will be determined by the percentage of code in your assignment that is not attributed to ChatGPT. If during the viva if you are unable to explain any part of the code, that code will be considered as plagiarized.
- Clearly label and organize your code, including comments that explain the purpose of each section and key steps in your implementation.
- Properly document your code and include explanations for any non-trivial algorithms or techniques you employ.
- Ensure that your Jupyter Notebook is well-structured, with headings, subheadings, and explanations as necessary.
- Your assignment will be evaluated not only based on correctness but also on the quality of code, the clarity of explanations, and the extent to which you've understood and applied the concepts covered in the course.
- Make sure to test your code thoroughly before submission to avoid any runtime errors or unexpected behavior.
- The Deadline will not be extended.

- Moss will be run on all submissions along with checking against online resources.
- We are aware how easy it is to write code now in the presence of ChatGPT and Github Co-Pilot, but we strongly encourage you to write the code yourself.
- We are aware of the possibility of submitting the assignment late in github classrooms using various hacks. Note that we will have measures in place for that and anyone caught attempting to do the same would be give zero in the assignment.
- **SUBMISSION FORMAT : Submit seperate files with all the worked out codes and necessary observations in the MARKDOWN for each problem.**

1 Problem 1

This task involves exploring methods of dimensionality reduction. We will be looking into **PCA** (principal component analysis), for this task. Principal Component Analysis (PCA) is the general name for a technique which uses sophisticated underlying mathematical principles to transforms a number of possibly correlated variables into a smaller number of variables called principal components. IEEE Signal Processing Magazine (Accessible through college internet)

Use only NumPy, Pandas, Matplotlib, and Plotly libraries for the tasks. The use of any other libraries shall be accepted only upon the approval of the TAs.

1.1 PCA [25]

This task requires you to implement Principal Component Analysis and perform dimensionality reduction on a given dataset(s). The list of subtasks is given below.

- Perform dimensionality reduction on the IIIT-CFW dataset, varying the number of principle components. We have given the script to pre-process the data and to get the necessary information from the image Script.
- Plot the the relationship between the cumulative explained variance and the number of principal components. The x-axis of the plot typically represents the number of principal components, and the y-axis represents the cumulative explained variance.
- Perform the dimensionality reduction on features that you have used for assignment 1 (pictionary dataset) and show the metrics you have shown for the assignment 1. Compare the results and write down the observations in the MARKDOWN.

- Observe the impact of dimensionality reduction on the dataset. Use a classifier on the dataset pre and post-dimensionality reduction (if the number of features of the dataset is n, perform dimensionality reduction varying the principal components from 1 to n) and note the accuracies of the classifier. You are free to use external libraries for the classifier.

1.2 Pictionary Dataset [10]

This task is to perform the PCA on the Pictionary Dataset (Dataset).The attachment also contains the description for the Dataset. Perform PCA for both drawer and guesser.

- Plot the above features with respect to the obtained PCA axes.
- What does each of the new axes that are obtained from PCA represent ?

2 Problem 2

The EM algorithm is used for obtaining maximum likelihood estimates of parameters when some of the data is missing. More generally, however, the EM algorithm can also be applied when there is latent, i.e. unobserved, data which was never intended to be observed in the first place. In that case, we simply assume that the latent data is missing and proceed to apply the EM algorithm. The EM algorithm has many applications throughout statistics. It is often used for example, in machine learning and data mining applications, and in Bayesian statistics where it is often used to obtain the mode of the posterior marginal distributions of parameters. [Columbia University]

Membership value r_{ic} of a sample x_i is the probability that the sample belongs to cluster c, in a given GMM (Gaussian Mixture Model). Likelihood values for a set of samples, measures the likelihood of the given data under a fixed model. In other words, likelihoods are about how likely the data is given the model, while membership values are about how likely the model is given the data. [reference]

Use only NumPy, Pandas, Matplotlib, and Plotly libraries for the tasks. The use of any other libraries shall be accepted only upon the approval of the TAs.

2.1 GMM: Gaussian Mixture Models[25]

This task requires you to implement the EM algorithm for GMM and perform clustering operations on a given dataset(s). The list of subtasks is given below.

- Find the parameters of GMM associated with the customer-dataset, using the EM method. Vary the number of components, and observe the results. Implement GMM in a class which has the routines to fit data (e.g. *gmm.fit(data, number_of_clusters)*), a routine to obtain the parameters, a
- routine to calculate the likelihoods for a given set of samples and a routine to obtain the membership values of data samples.

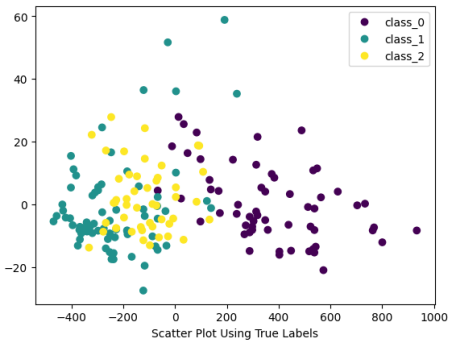

Figure 1: Scatter Plot of the Wine Dataset, after PCA with 2 principal components. (One of the axes, represents Principle Component - 1 and Other one, Principal Component - 2)

- Perform clustering on the wine-dataset using Gaussian Mixture Model (GMM) and K-Means algorithms. Find the optimal number of clusters for GMM using BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion). Reduce the dataset dimension to 2 using Principal Component Analysis (PCA), plot scatter plots for each of the clustering mentioned above, analyze your observations and report them. Also, compute the silhouette scores for each clustering and compare the results. You are free to use sklearn for the dataset, PCA, and Silhouette Score computation.

3 Problem 3

Hierarchical clustering is a popular method for grouping objects. It creates groups so that objects within a group are similar to each other and different from objects in other groups. Clusters are visually represented in a hierarchical tree called a dendrogram. [Reference for hierarchical clustering and linkages]

Use only NumPy, Pandas, Matplotlib, and Plotly libraries for the tasks. The use of any other libraries shall be accepted only upon the approval of the TAs.

3.1 Hierarchical Clustering: Linkages and Features [25]

This task requires you to implement Hierarchical clustering, and perform clustering on a given dataset(s). The list of subtasks is given below. You are expected to implement the required, using classes and methods. We expect to see routines like *hc.linkages(X, linkage_type)* (takes the data and provides linkage matrix), *hc.dendogram(Z)* (takes the linkage matrix and plots a dendogram).

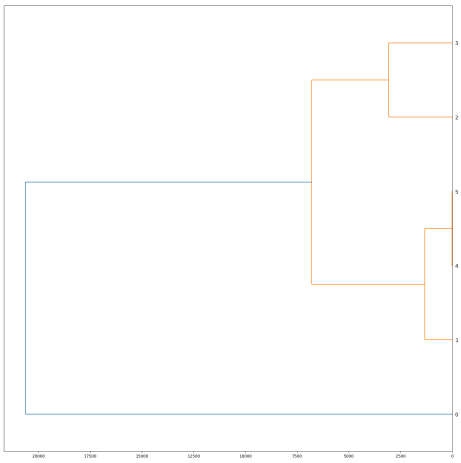

Figure 2: Example of a dendogram, obtained from a different dataset. The Vertical axis represents individual points. The horizontal one represents the distance between cluster.

- Perform hierarchical clustering on the dataset and obtain the linkage matrix. Vary the linkages and features used and state your observations. Plot the dendogram using the linkage matrix.
- Perform hierarchical clustering on the gene expression dataset and obtain the linkage matrix. Vary the linkages and features used and state your observations. (In the dataset, you are given 58 genes, their respective expression levels for 12 proteins and their IDs, resulting in a 58×13 matrix). Plot the dendogram, using the linkages obtained.

4 Problem 4

For this section, you are free to use external libraries to solve the question(s). Submit separate notebooks for each of the following sub-questions in this section.

4.1 Problem 4.1 [20]

You have been provided an dataset of 99 different shapes KIMIA-99. The task is to find the align the remaining shapes based on the orientation of the given template shape. Along with the code, write the flowchart of the algorithm that you will be using to implement the following task in the Jupyter Notebook itself as a MARKDOWN.
Attached is an example of the task Example .

4.2 Problem 4.2 [20]

The task is to determine the optimal horizontal and vertical euclidean distance thresholds between bounding boxes containing words on a document page. The objective of this task is to establish connections between boxes within a paragraph while ensuring that boxes across paragraphs and columns remain unconnected. Attached are illustrative examples showcasing the desired box connections and a sample visualization of the expected output ATTACHMENT.
You have also been given the following scripts -

- To visualize the enclosing boxes.
- Script to visualize the connecting boxes is there in the above attachment. The input for this script is a dataframe object with the following attributes.
  - ID : The ID number of the word which is in int datatype.
  - Top-Left, Bottom-Right, Top edge center, Bottom edge center, Right edge center, Left edge center : A list containing the x and y coordinates of the respective coordinates as understood by their attribute names.
  - Top box, Bottom box, Right box, Left Box - A list containing the distance and id of the nearest neighbour in the Top, Bottom, Right and Left directions respectively. **HINT - To remove the connection make sure that the list is [-1, 0].**

  NOTE : Make all necessary modifications to the script and then run the script.

4.3 Problem 4.3 [20]

For this problem, you shall be provided with a dataset of 2-dimensional points, of various colors. The data you will be given is in the form of an array, where each element, X, represents a point in the 2D color space. The data has been generated from 7 distinct Gaussian color components. The list of subtasks is given below.

- Find the likely color components which generate the dataset.
- Create a function which would take in an input of *(number_of_components (an integer, n), means (an numpy array of shape (n, 2)), covariances (a numpy array of shape n, 2, 2))*, and generates a sample dataset with the n likely components described by the above components. State your observations.

5 Relevant Readings

This section contains some reading material regarding the assignment, which may assist you in solving or understanding the question, couple with some resources to gain deeper knowledge regarding the topics. **This section is intended as just some help, and it is not graded or evaluated**.

- Reference for Gaussian Mixture Model
- Reference for Bayesian Information Criterion
- Reference for hierarchical clustering