Introduction to Statistical learning

DS413 - Assignment 3

Introduction

This assignment involves building Bayes classifiers using Gaussian Mixture Models (GMM) on synthetic and real-world datasets. It includes linearly and nonlinearly separable data, as well as vowel formant frequency data. Each class is split into 70% training and 30% testing data. Classifier performance is evaluated under different covariance assumptions. GMM parameters are initialized using K-means, and experiments are conducted with varying numbers of mixtures to study their impact on classification accuracy and model performance.

Introduction to Data set

Dataset 1: Linearly separable classes: 3 classes, 2-dimensional linearly separable data is given. Each class has 500 data points.

Dataset 2: Nonlinearly separable classes: 2-dimensional data of 2 or 3 classes that are nonlinearly separable. The number of examples in each class and their order is given at the beginning of each file.

Dataset 3: Real-world data (Vowel data): 2-dimensional data of 3. The real-world data set correspond to the formant frequencies F1 and F2 for vowel utterances.

Divide the data from each class into training, and test data. From each class, train, and test split should be 70% and 30% respectively.

Assumption: Class-conditional densities are Gaussian

Problem I: Bayes Classifier

Classifiers to be built for each dataset: Bayes classifier with,

(a) Covariance matrix for all the classes is the same and is $\sigma^2\mathbf{I}$.

- You can obtain the same covariance matrix for all the classes by taking the average of covariance matrices of all the classes. You can obtain the same variance by averaging all the variances.

(b) Full Covariance matrix for all the classes is the same and is $\mathbf{\Sigma}$.

- You can obtain the same covariance matrix for all the classes by taking the average of covariance matrices of all the classes.

(c) Covariance matric is diagonal and is different for each class.

(d) Full covariance matrix for each class is different.

Presentation of results

Report should include the results of studies presented in the following forms for each classifier and for each dataset.

(a) Confusion matrix, classification accuracy, precision for every class, mean precision, recall for every class, mean recall, F-measure for every class and mean F-measure on test data.

(b) Inferences on the plots and inferences on the results observed (such as performance, nature of decision surface etc.) for each dataset.

Problem 2: Bayes classifier using GMM

Build the Bayes classifier using GMM on given datasets, Parameters of GMM are to be initialized using k-means clustering.

(a) Perform the experiments on different number of mixtures of GMM $(1, 2, 4, 8, 16, 32, 64)$.

Presentation of results

Report should include the results of studies presented in the following forms for each classifier and for each dataset.

(a) Classification accuracy, precision for every class, mean precision, recall for every class, mean recall, F-measure for every class and mean F-measure on test data (for each of the different parameters).

(b) Confusion matrix based on the performance for test data (for the best GMM model).

(c) Constant density contour plot for all the classes with the training data superposed.

(d) Decision regions plot with the training data superposed. Comparison with the results from previous problems.

(e) Report should also include the graph of iterations vs log likelihood for all the datasets with different number of components.

Conclusion

All the results need to be included in a **report**, and the report should contain the results of all the models mentioned above, along with their explanations.

Bonus Question

Please provide a summary of the methodology adopted to complete the assignment, along with the key challenges encountered during the process. Additionally, include a table at the end of the report outlining the distribution of tasks among group members.

References

List all references and sources used for the assignment(sources can be some generative mode or anything across the internet).