Indian Institute of Technology Ropar

MA515 Foundations of Data Science

Assignment

Deadline: November 10, 2024 (11:59 pm)

Instructions:

- Each student is supposed to do the assignment individually.
- Students, please be advised that any form of cheating or plagiarism, including copying from peers or the internet, will result in 0 marks. The assignments will be screened through **Plagiarism Checker by Grammarly**.
- Your solutions should be neatly presented. Attach Python codes named in the format **"Assignment_EntryNumber_Codes.ipynb"** and a PDF file named **"Assignment_EntryNumber_Solutions.pdf"**. The Python codes should execute without any issues and should be well-commented.
- There will be a viva after the submission deadline is over. The dates will be announced later.
- How to submit: Submit a zip file named as 'Assignment_EntryNumber_Solutions.zip containing the python notebook and its pdf on the google classroom.

---

1. **LDA and QDA:** Generate synthetic data of 3000 samples having 2 features and 2 unique classes. Let n_1 and n_2 (randomly generated) be the number of samples in each class. Split this dataset into training and testing datasets. The test dataset must have s% of samples with $s \sim Unif(0, 0.3)$.
   - Write two functions that implements (fit and predict) linear discriminant analysis and quadratic discriminant analysis.
   - Fit the models on the training dataset and predict the classes for test dataset.
   - Find the confusion matrix and accuracy score for both models on the test datasets.
   - Plot obtained results for both training and test datasets.
   - Compare the results from both LDA and QDA.
2. **k-means Clustering Algorithm:**
   - Write Python code to implement the k-means clustering algorithm from scratch. The algorithm should use **Euclidean distance** as the distance measure between data points and cluster centroids.
   - Your implementation should:
     - Initialize k cluster centroids randomly from the data points.
     - Assign each data point to the nearest centroid based on Euclidean distance.
     - Update each centroid to be the mean of the points assigned to it.
     - Repeat the assignment and update steps until the centroids no longer change or a maximum number of iterations is reached.
   - For dataset 1 implement the algorithm for $k = 2$ and 3 clusters, Compare the results and plot the clusters. For dataset 2 implement the K-Means algorithm for $k = 2$, 3 and 4 clusters and print the centroids and the respective clusters.