National Institute of Technology Calicut

Department of Computer Science and Engineering

CS4092D MACHINE LEARNING LABORATORY (S7 B Tech)

Problem Set 02

Submission deadline (on or before):

- **21/08/2025, 10:00 AM**

Policies for Submission and Evaluation:

- You must submit your programs in the (Eduserver) course page, on or before the submission deadline. Also, ensure that your programs compile and execute without errors in python. During evaluation, failure to execute programs without compilation errors may lead to zero marks for that program. Detection of any malpractice can lead to awarding an F grade in the course.
- You must execute the programs in offline compilers (such as Jupyter Notebook) on the day of evaluation.

Naming Conventions for Individual Program

- PS0X < PROBLEM SET NUMBER > < ROLL NO > <FIRST − NAME > < PROGRAM − NUMBER > . <extension > (For example: PS02 BxxyyyyCS LAXMAN 1.py). Please make sure that you follow the naming conventions correctly.

Naming Conventions for Submission

- Submit a single ZIP (.zip) file (do not submit in any other archived formats like .rar, .tar, .gz) containing the source code (.py file) for the three programs. The name of this file must be P S <PROBLEM SET NUMBER > < ROLLNO > < FIRST − NAME > .zip (For example: PS02 BxxyyyyCS LAXMAN.zip). DO NOT add any other files (like temporary files, input files, etc.) except your source code, into the zip archive.

PS-02

Dataset Description:

The Breast Cancer Wisconsin (Diagnostic) dataset contains 569 observations and 32 variables, including an ID column, a diagnosis label, and 30 numeric features computed from digitized images of fine needle aspirate (FNA) of breast masses. The target variable `diagnosis` is binary, where 'M' indicates malignant and 'B' indicates benign tumors. The 30 input features represent properties of the cell nuclei present in the image, categorized into three groups of 10 attributes each: mean, standard error, and worst (i.e., the largest value) of measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. All these features are continuous and suitable for statistical modeling. The `id` column serves as a unique identifier and does not contribute analytically to classification tasks; it is usually dropped during preprocessing .

Note: Please don’t use inbuilt functions

Naive Bayes Programming Questions

1. Load the Breast Cancer Wisconsin (Diagnostic) dataset and preprocess it by removing irrelevant columns and encoding the target variable.

2. Implement a function that calculates the prior probabilities of each class (malignant and benign).

3. Estimate the mean and variance of each feature for each class assuming a Gaussian distribution (Maximum Likelihood Estimation).

4. Write a Gaussian likelihood function to compute the likelihood of feature values given the class-specific mean and variance.

5. Implement a Naive Bayes classifier that uses the Gaussian likelihoods and class priors to make predictions.

6. Evaluate the Naive Bayes classifier using accuracy, precision, recall, F1-score, and a confusion matrix.

7. Plot ROC curves and calculate AUC scores for the Naive Bayes classifier.

MLE and MAP Programming Questions

1. Write a function to calculate and visualize the correlation between features in the dataset and remove highly correlated features (threshold > 0.9).

2. Implement a function that calculates the prior probabilities of each class (malignant and benign) in the dataset.

3. Implement a function that estimates the class-conditional mean and variance of each feature using Maximum Likelihood Estimation (MLE).

4. Implement a Gaussian likelihood function that computes the probability density of a feature value given the estimated mean and variance.

5. Write a function that performs classification using the MLE approach, where each sample is assigned to the class with the highest likelihood.

6. Write a function that performs classification using the Maximum A Posteriori (MAP) approach, combining prior probabilities and Gaussian likelihoods to compute posteriors.

7. Evaluate both MLE and MAP classifiers using accuracy, confusion matrix, precision, recall, and F1-score.

8. Plot the distributions and Gaussian fits of selected features (e.g., radius_mean) for both classes to visualize the effectiveness of your model.