Machine Learning External Lab Exam

Set 1 – Lab Exam

PART A – Image Classification using Fashion MNIST (Total: 22 marks)

Dataset Description:

You are provided with a file named set_1_dataset_1.csv, which contains the Fashion MNIST dataset. This is a collection of grayscale images of clothing items, such as shirts, shoes, dresses, etc. Each image is 28 pixels wide and 28 pixels tall, totaling 784 pixels per image. The first column in the dataset contains the label (a number from 0 to 9) indicating the type of clothing. Example: 0 = T-shirt/top, 1 = Trouser, 2 = Pullover … up to 9 = Ankle boo

The remaining 784 columns represent the pixel values of the image, in row-major order (left to right, top to bottom). Each pixel value ranges from 0 (black) to 255 (white).

1. Load the dataset set_1_dataset_1.csv. (No marks)
2. Print the shape of the dataset. (1 mark)
3. Print the unique classes and their counts. (2 marks)
4. Display one image by reshaping it to 28 x 28 and print its label. (2 marks)
5. Display 6 images using subplot with their labels. (2 marks)
6. Prepare features (X) and labels (y), split the dataset (15% test), and print the shape of training and testing sets. (3 marks)
7. Train a suitable classification model. (2 marks)
8. Predict on the test data. (1 mark)
9. Print the confusion matrix. (1 mark)
10. Print the classification report. (1 mark)
11. Display 6 test images with actual and predicted labels. (2 marks)
12. Apply GridSearchCV on one parameter. (3 marks)
13. Print the best score and best parameter from GridSearchCV. (2 marks)

PART B – Regression using Automobile MPG Dataset (Total: 18 marks)

Dataset Description:

You are given a CSV file named set_1_dataset_2.csv, which contains data on automobiles. This dataset includes various specifications of cars such as: Horsepower, Weight, Number of cylinders, Model, year, Displacement, Acceleration. The target variable is mpg (miles per gallon), which indicates the fuel efficiency of a car.

14. Load the dataset set_1_dataset_2.csv. (No marks)
15. Print the data types of each column. (1 mark)
16. Print the number of missing values in each column. (1 mark)
17. Drop rows with missing output labels (mpg column). (1 mark)
18. Impute the missing values if any (mean/mode). (2 marks)
19. Encode categorical variables. (2 marks)

20. Plot a histogram of the horsepower column. (2 marks)
21. Plot a scatter plot of weight and horsepower vs mpg in two subplots. (2 marks)
22. Prepare features (X) and label (y), split the dataset (15% test). (2 marks)
23. Train a regression model. (2 marks)
24. Print the R^2 value on the test data. (2 marks)
25. Predict the regression coefficients. (1 mark)

Set 1 – Viva Questions (20 Marks)

1. The K in K-Nearest Neighbors stands for the number of _____.
2. A hyperplane in 2D space is a _____.
3. A classification model trained using a labeled dataset is an example of _____ learning.
4. A probability distribution over multiple classes is produced by the _____ function in logistic regression.
5. In Naive Bayes, joint probability is the probability of _____.
6. The main advantage of Decision Trees is their ability to handle both _____ and categorical data.
7. The output of a regression model is typically a _____ value.
8. The support vectors in an SVM are the data points that lie _____ to the margin.
9. Data preprocessing includes steps like dealing with missing values, scaling, and _____.
10. In feature scaling, _____ normalization brings features into a 0-1 range.
11. The Gaussian Naive Bayes classifier assumes features follow a _____ distribution.
12. One common preprocessing method to deal with non-numerical features is _____ encoding.
13. Increasing the number of neighbors in KNN can lead to smoother but potentially less _____ models.
14. Information gain in Decision Trees is calculated by subtracting weighted entropies from the _____ .
15. In unsupervised learning, K-Means attempts to minimize the _____ within clusters.
16. A classifier that can handle nonlinear decision boundaries using kernels is _____.
17. When comparing models, the metric _____ is preferred for regression tasks.
18. The basic building block of an artificial neural network is the _____.
19. A neuron triggers an output signal only if the combined input exceeds a certain _____.
20. To implement XOR logic using neural networks, we need a network with at least _____ layers.