NAME: **RAHUL BIRWADKAR**
MATRICULATION NO.: **11037364**
DATE OF SUBMISSION: **27/04/2024**

Hochschule Heidelberg

Task 1 : Handwritten Digit Classification.

Introduction

In this project, we are going to see the apply Gaussian Naive bayes approach for handwriting digit classification. In this project we are calculation Confusion matrix, Micro average precision, Micro average recall, Accuracy, and F1 score.

Implementation and Methodology

For the implementation of the project, VS code IDE is used, and Python programming language is used.

- Firstly, Import the following libraries.
  - Numpy
    - NumPy stands for Numerical Python supports large arrays and matrices and can write advanced arithmetic operations that operate on arrays
  - Pandas
    - Pandas is a Python library used for working with data sets.
  - Sklearn
    - Imports the Gaussian Naive Bayes classifier from the Scikit-learn library.

- Upload MNIST training and testing datasets using Pandas library.
- X_train, Y_train and X_test, Y_test used to sperate the features and labels.
- Initializes a Gaussian Naive Bayes classifier (gnb).
- Fits the classifier to the training data (x_train, y_train).
- Uses the trained classifier to predict labels for the testing data (x_test), storing the predictions in y_pred.
- Calculates the confusion matrix using the actual labels (y_test) and the predicted labels (y_pred).
- Calculates accuracy, micro-average precision, micro-average recall, and F1 score using the confusion matrix.

Results:

```
Confusion Matrix:
 [[ 870    0    3    5    2    5   31    1   35   28]
 [   0 1079    2    1    0    0   10    0   38    5]
 [  79   25  266   91    5    2  269    4  271   20]
 [  32   39    6  353    2    3   51    8  409  107]
 [  19    2    5    4  168    7   63    7  210  497]
 [  71   25    1   20    3   44   40    2  586  100]
 [  12   12    3    1    1    7  895    0   26    1]
 [   0   15    2   10    5    1    5  280   39  670]
 [  13   72    3    7    3   11   12    4  648  201]
 [   5    7    3    6    1    0    1   13   18  955]]
accuracy: 0.5558555855585559
micro_Avg_precision: 0.6865108998951719
micro_Avg_recall: 0.5484735723231241
f1 score: 0.6097778460775847
```

Figure 1: Output

Questions

- How do you handle features with a constant value across all images in the dataset?

- Features with a constant value provide absolutely no predictive power for a machine learning model. In other words, they don't help the model distinguish between different classes or determine any kind of pattern.
- The first step is to find these constant features. Calculate the variance of each feature. Features with a variance of zero are constant.Use techniques to remove the columns representing constant features.

- How do you handle features with a constant value across all images belonging to a given class?

- Select only the images belonging to the specific class for which you want to handle constant features.
- Calculate the standard deviation of each feature within the subset of images belonging to the chosen class. If the standard deviation is zero, it indicates that the feature has a constant value across all images in that class.
- Delete the constant features from the subset of data corresponding to the chosen class. This ensures that these features do not influence the classification process for that particular class.

- Do you, and if so, how do you the prevent arithmetic underflow or overflow?

- Arithmetic underflow or overflow can occur when performing calculations with very small or very large numbers, respectively. In the context of Gaussian Naive Bayes,

where probabilities are multiplied together, underflow can occur when multiplying many small probabilities. To prevent this, you can work with log probabilities instead of raw probabilities. Taking the logarithm of probabilities allows you to sum probabilities instead of multiplying them, which mitigates the risk of underflow.

- How do you explain the low classification accuracy obtained by applying the Gaussian naive Bayes approach in this particular context?

- Gaussian Naive Bayes assumes that features are independent and follow a Gaussian distribution. If these assumptions do not hold true in the dataset, such as when features are correlated or do not follow a Gaussian distribution, the model may not capture the underlying patterns effectively, leading to low accuracy.

References:

[1] How To Import Numpy As Np - GeeksforGeeks
[2] Pandas Introduction (w3schools.com)
[3] Gaussian Naive Bayes - GeeksforGeeks
[4] gnjatovic.info: The Milan Gnjatovic Website