ASU ID: 1217060520

Name: Tiriac Ioana-Raluca

SML Project 1 Report
- Density Estimation & Classification –

1. Objective

Given a subset of the MNIST dataset in form of 2 train and 2 test sets - one for digit 0 and digit 1 – train two classifiers to predict the correct label – 0 or 1 given an image as input and calculate the accuracy of these classifiers. The trainsets each contain 5000 images of dimension 28 x 28 and the test sets contain 980 and 1135 images of the same dimension.

Task 1: Extracting two features – mean & standard deviation from the given trainset (a subset of 5000 images from MNIST dataset) and creating a 2D data points array

Task 2: Calculating parameters for two-class naïve bayes classifiers

(No.1) Mean of feature1 for digit0
(No.2) Variance of feature1 for digit0
(No.3) Mean of feature2 for digit0
(No.4) Variance of feature2 for digit0
(No.5) Mean of feature1 for digit1
(No.6) Variance of feature1 for digit1
(No.7) Mean of feature2 for digit1
(No.8) Variance of feature2 for digit1

Task 3: Calculation of Naïve Bayes classifiers for digit 0 and digit 1 using the PDF formula of Gaussian Distribution and calculating the predictions for a test set of digit 0 (with correct label 0) and test set of digit 1 (with correct label 1).

Task 4: Calculation of the accuracy of the predicted labels

2. Task 1

The features have been calculated by iterating each trainset and using NumPy's mean() and std() functions. The asked 2D array was created using append() function on the axis= 0. I've created a general *extract_features(test set)* function to be invoked with both train sets and test sets and to return a 2D feature array.

It resulted in a features array, looking like this containing 2D data points.

```
[[ 47.34566327  92.43247208]
 [ 40.35841837  83.9385578 ]…]
```

Task 2

For task2 a general *calculate_mean_var()* function was defined taking in as a parameter an array and returning mean() and var() – calculated with NumPy of that array. So, for both trainsets – first the function was invoked using the first feature column (MEAN) and then using the second feature column (STD).

The resulted parameters were:

```
Mean feature1 digit0:  44.2031344388
Variance feature1 digit0:  116.286767271
Mean feature2 digit0:  87.4110447613
Variance feature2 digit0:  102.431947094
Mean feature1 digit1:  19.340167602
Variance feature1 digit1:  31.1393121632
Mean feature2 digit1:  61.3049091784
Variance feature2 digit1:  82.2398612549
```

Task 3

The NB classifiers were implemented using the 8 parameters above and the PDF function for the Gaussian Distribution formula. Two classifiers were implemented, one for digit 0 and one for digit 1, each one using their respective parameters.

In order to calculate the predictions for the given test sets, each containing 980 and 1135 images with dimensions 28 x 28 respectively, the mean and standard deviation features had to be extracted like the train sets – by using the already defined *extract_features (data set)* function. This resulted in each train set having a 2D array, made up of 2D data points, each data point having as feature 1: mean and feature2: standard deviation.

The classifiers' calculation uses the formula *prior_prob_digit * prob_feature1 * prob_feature2* having as input 2 features. Now by feeding the two features of each data point of a test set into each classifier two probabilities are being calculated. The higher probability indicates the predicted label of the image in the test set: if it was the digit 0 classifier generating it, then it's the 0 label, otherwise it's 1.

Task 4

The accuracy of the predicted labels was calculated by using the formula:
# of correctly predicted labels/ total length of test set (correctly predicted labels + incorrectly predicted ones).
The results in % are:

```
Accuracy for digit 0 prediction:  91.73469387755102
Accuracy for digit 1 prediction:  92.33480176211454
```