Project 1: Density Estimation and Classification Report

Student ID:1228855799

1.Introduction

the MNIST dataset contains 70,000 images of handwritten digits, divided into 60,000 training images and 10,000 testing images. I use only a part of images for digit “0” and digit “1” in this question. I have the following statistics for the given dataset:

Number of samples in the training set: "0": 5000;"1": 5000.

Number of samples in the testing set: "0": 980; “1": 1135.

I assume that the prior probabilities are the same (P(Y=0) = P(Y=1) =0.5), although you may have noticed that these two digits have different numbers of samples in testing sets.

2.Task1

I use the free open-source Python library SciPy to obtain data: training data set 0, training data set 1, test data set 0, test data set 1.I need to first extract features from your original trainset to convert the original data arrays to 2-Dimensional data points.

Previous assumptions: these two features are independent and that each image is drawn from a normal distribution.

Feature1: The average brightness of each image (average all pixel brightness values within a whole image array)

Feature2: The standard deviation of the brightness of each image (standard deviation of all pixel brightness values within a whole image array)

NumPy directly calculates the average and standard deviation from training data set 0 and training data set 1.

3.Task2

I need to calculate all the parameters for the two-class naive bayes classifiers respectively, based upon the 2-D data points you generated in task 1.NumPy directly calculates the mean and variance from data obtained from task 1.

1	Mean of feature1 for digit0	44.2101193878	2	Variance of feature1 for digit0	115.195094866
3	Mean of feature2 for digit0	87.4238343014	4	Variance of feature2 for digit0	101.725589801
5	Mean of feature1 for digit1	19.3429196429	6	Variance of feature1 for digit1	31.1836615196
7	Mean of feature2 for digit1	61.3104678976	8	Variance of feature2 for digit1	82.7949939856

4.task3

So now we take the mean and standard deviation of the entire training dataset. Then we divide the data into two sets, one set containing only digit 0 and another set containing digit 1. Now we calculate the gaussian naïve bayes, because it is a typical assumption if the values are continuous associated with each class are distributed according to normal distribution. We use this formula to predict class probability:

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

We already know the average and standard deviations, so we implement these values in the formula. We are applying the formula of each data set on the number 0 set and the number 1 set respectively. Now, we use test data to test the data set to predict whether the value will be classified and add the number 0 or the number 1 to the list if the probability of the number 0 and the number 1. This will continue to be used for the complete testing dataset, each of which will be classified as 0 or 1 and will be stored in the list.

5.task4

After analyzing and predicting the data we will now check the accuracy of the model by comparing the output predicted by the model that is stored in the list and the labels given in testing dataset.

1	accuracy digit 0	0.9173469387755102
2	accuracy digit 1	0.9356828193832599

6.Summary

This project mainly builds the model through test data, and then uses the model for data prediction to obtain the prediction results. By comparing the prediction results with the test data, the accuracy of the final prediction is obtained. The project has two prerequisites. The project is characterized by independence, and the project conforms to the normal distribution. Use training data to extract features, and then estimate the two-dimensional normal distribution parameters of each number. Note: You will have two distributions, one for each number. Naïve Bayes classifies the test data using estimation distribution. Report the accuracy of the classification of "0" and "1" in the test set.