Lucy Zhang
01/27/23
ID: 1229576766

Project 1: Density Estimation and Classification

Introduction:

In this project, I was able to build a image classifier for digit 0 and digit 1. As we run the code, we will load the trainset and test set for digit0 and digit1 respectively. Both trainset and test set are sub-dataset from the MNIST dataset. The MNIST dataset contains 70,000 images of handwritten digits, divided into 60,000 training images and 10,000 testing images. We assume that the prior probabilities are the same (P(Y=0) = P(Y=1) =0.5)

Given dataset:

- Number of samples in the training set: "0": 5000; "1": 5000.
- Number of samples in the testing set: "0": 980; "1": 1135

Extract Features:

Extract the following two features for each image for train sets and test sets. We assume that these two features are independent and that each image is drawn from a normal distribution, and result eight data arrays for train sets and test sets:

- Feature1: The average brightness of each image
- Feature2: The standard deviation of the brightness of each image

Train sets	Test sets
feature1_digit0	test_feature1_digit0
feature2_digit0	test_feature2_digit0
feature1_digit1	test_feature1_digit1
feature2_digit1	test_feature2_digit1

The Parameters:

Calculate all the parameters for the two-class naive bayes classifiers respectively, based upon the 2-D data from above.

Parameters for Digit 0	Est.	Parameters for Digit 1	Est.
Mean_of_feature1_for_digit0	44.10	Mean_of_feature1_for_digit1	19.36
Variance_of_feature1_for_digit0	114.00	Variance_of_feature1_for_digit1	31.48
Mean_of_feature2_for_digit0	87.33	Mean_of_feature2_for_digit1	61.34
Variance_of_feature2_for_digit0	101.30	Variance_of_feature2_for_digit1	82.80

Gaussian Naïve Bayes:

Since we get the NB classifiers' parameters from above, we implement their calculation formula according to their Mathematical Expressions. Then we use our implemented classifiers to classify/predict all the unknown labels of newly coming data points. By comparing the probability of digit 0 and probability of digit 1for each test element, we can predict if the image belongs to digit 0 or digit 1.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Accuracy:

After successfully predicted the labels for all the test data, now we need to calculate the accuracy of our predictions for test set for both digit0 and digit1 respectively. With the number of samples in each testing set already given, the accuracy result with correct predictions above listed below.

Testing Sets	Accuracy
Accuracy_for_digit0testset	91.73%
Accuracy_for_digit1testset	92.33%

Observation and Analysis:

While working on this project, I have practiced Naïve Baye Classifier with real data and observed the parameters for train set digit 0 and digit 1 are very different, such as average brightness of digit 0 image estimated 44.1 and 19.36 for digit 1. The parameters difference helped with calculating probability of digit 0 and probability of digit 1for each test element, and it effectively improved the testing accuracy. With accuracy all above 90%, I can say that Gaussian Naive Bayes Classifier works well with data in this project, and it can be used for a variety of other classification problems.