Project 1 - Density Estimation and Classification

Rishikesh Iyer ID: 1227881670

January 29, 2024

1 Project Summary

This project involved extracting two sets of data samples, finding the mean and standard deviation of each of the samples, and then finding the mean and variance of each of those, for a total of 8 parameters. These parameters will then be used for Naïve Bayes classification in order to predict the label for two different data sets of testing data of 980 for 0 and 1135 for 1.

2 Task 1

This task involved extracting the training data samples into 2 dimensional arrays of the mean and standard deviation for each of the data samples from the data sets. I did this by simply creating two empty arrays and then appending the mean and standard deviation in their own array for each of the samples. This extraction was done for both of the datasets.

3 Task 2

With the two dimensional arrays containing the averages and standard deviations of the samples, the next task was to retrieve the mean and variance for each of the averages and standard deviations. This was done by using the imported numpy module to calculate the mean and variance. The command for this was numpy.mean(array, axis=0) and numpy.std(array, axis=0), which will retrieve the mean and the standard deviation for the same indexed elements in all of the arrays. Next, we use the numpy.mean and numpy.var methods to each of the mean and variances. This provides the mean of the averages and standard deviations of both datasets and the variances of the averages and standard deviations of both datasets, resulting in the 8 parameters required for the Naïve Bayes classifiers to predict the unknown labels in the testing sets.

4 Task 3

Now with the 8 parameters, we plug them into the Gaussian probability density function to predict the labels. The formula for this is f(x) = $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, where x is the input datapoints from the test sets, μ is the mean and σ is retrieved by taking the square root of variance from task 2. The test datasets were also extracted into 2 dimensional arrays in the same manner as the train sets containing the mean and variance of the test data samples. As shown in lecture, for each of the two datasets, the mean and the standard deviation for the samples were input in the probability density function, and the prediction is done by computing and comparing which of the probabilities were higher. This is done by multiplying the probability density functions with each of the standard variation and mean as input variables by the prior probability of each of the digits, which is 0.5 for both datasets. The higher probability computed for each of the inputs is predicted to be the label of either the 0 or 1 digits.

5 Task 4

The last part involves counting the number of samples predicted to be in the corresponding dataset divided by the number of samples in the dataset. This is the accuracy of the probability density function.

6 Conclusion

The accuracy for the 0 data samples was 91 and the accuracy for the 1 digit data samples was 92, showing that the accuracy for the predictions was very high and that the Naïve Bayes Classifier model is a good model to accurately predict labels for a dataset.