Student ID: 1204599352

The data provided for the project is a set of multi-dimensional arrays that contains data about the brightness of the individual pixels for images. There are two training sets: a set of images representing digit 0, and a set of images representing digit 1. There are 5000 images in each of the training sets that are 28x28, providing a total of 784 pixels per image. The pixels for each image were transformed into two features, average and standard deviation, that are easier to manipulate and use to represent the images. The average and standard deviation of the brightness values for all 784 pixels in an image are used to identify the images. It was assumed that the features are independent and that each image is taken from a normal distribution. It was also assumed that the prior probability of an image being part of the two classes was 0.5. The average and variance of the two features for each image set were calculated and reported in the submission. These values are below:

1. Mean of feature 1 (average) for digit 0: 44.2244505102
2. Variance of feature 1 (average) for digit 0: 115.002954019
3. Mean of feature 2 (standard deviation) for digit 0: 87.4560676165
4. Variance of feature 2 (standard deviation) for digit 0: 100.965049403
5. Mean of feature 1 (average) for digit 1: 19.3828806122
6. Variance of feature 1 (average) for digit 1: 31.7518393062
7. Mean of feature 2 (standard deviation) for digit 1: 61.3648740226
8. Variance of feature 2 (standard deviation) for digit 1: 83.0522156466

Using these features, we want to be able to classify/predict some test values as either the digit 0 or the digit 1. Due to our assumption regarding the normal distribution, we can use the formula from the lecture to define our PDF function that will determine the probability that a test data point will be of the class digit 0 or digit 1 based on the average and standard deviation of the brightness values of the test image. The formula used can be seen below:

Diagram, schematic

Description automatically generated

The formula asks the question, what is the probability that feature x is of the class digit 0 or the class digit 1? Mu (μ) is value 1, 3, 5, or 7 above and sigma (σ) is value 2,4,6, or 8 above.

For some test image x, we want to know how likely the image is part of one of our classes. This forms the likelihood function p(x|θ) where θ is a class. In order to achieve this, we want to maximize the likelihood function by finding the class that returns the largest probability value. Using our assumption of the independence of the features, we can determine that the probability of a class y is the multiplication of the probability of the features and the prior probability. This can be seen in the formula below:

For an example test image, the probability of feature 1 (average) can be found by plugging in value 1 above for μ and value 2 above for σ. X is the average pixel brightness value for the test image. The probability of feature 2 (standard deviation) can be found by plugging in value 3 above for μ and value 4 above for σ. X is the standard deviation of the pixel brightness values for the test image. These probability values are multiplied by the assumed prior probability of 0.5 to determine the estimated probability that the test image is of the class digit 0. This is repeated for the digit 1 class. The class with the maximum probability value is determined to be the class for the test image.

The above process was performed for each of the test images provided. There were 980 test images for digit 0 and 1135 test images for digit 1. Each of these test images was classified as either digit 0 or digit 1 and compared against the ground truth value provided. In order to determine the accuracy of the model created, the below formula was used:

For the digit 0 test set, the accuracy was 0.9173469387755102, while the digit 1 test set had an accuracy of 0.9233480176211454. These accuracy values appear to be reasonable for the given data. The prior probability values used were not 100% exact given the uneven distribution of test values between the sets, so this can be improved for future attempts at this experiment. Additionally, pixel brightness may not be the best option for classifying an image as 0 or 1. Many other digits can have similar average or standard deviation values for pixel brightness, which may result in possible issues even though this model showed around 90% accuracy.