CSE-575 Project-1 Report

In this Project I have implemented a naive bayes classifier for image classification. I was given two datasets for training one belonging to digit zero and the other was digit one from MNIST dataset. Both had 5000 training data and had prior probabilities of 0.5. Task 1 indicated that the images were drawn from a normal distribution, so I had to use gaussian likelihood estimation, it also said to create to features to use for classification. Feature 1 was the average brightness of each image which I calculated by NumPy’s mean function along the axes 1 and 2 as each 2-D image had brightness values in their matrix. After that was done for each training set, I calculated the standard deviation for the brightness of each image. In task 2 I have calculated the values required as following:

(No.1) Mean of feature1 for digit0: 44.1835244898

(No.2) Variance of feature1 for digit0: 116.58047493

(No.3) Mean of feature2 for digit0: 87.4005523816

(No.4) Variance of feature2 for digit0: 102.510631096

(No.5) Mean of feature1 for digit1: 19.4063058673

(No.6) Variance of feature1 for digit1: 31.1762082116

(No.7) Mean of feature2 for digit1: 61.4137659884

(No.8) Variance of feature2 for digit1: 82.3016868451

In task 3, I have calculated the scores for each classifier by using the formula: prior probability \* feature1-likelihood\*feature2\*likelihood. The prior is calculated by using the training data and although test data wasn’t divided equally, as we had 5000 training data for both digits, the prior probability was 5000/10000 = 0.5 for each. I have used the gaussian likelihood formula and implemented it as: 1/numpy.sqrt(2\*math.pi\*varianceOfFeatureForDigit) \* numpy.exp (-(averagePixelTestArray1- meanOfFeatureForDigit)\*\*2/(2\*varianceOfFeatureForDigit)). I took the log likelihood as to avoid underflow on any small values and compared all testing data’s scores for digit one and digit zero likelihood scores. For the last part I have created two matrices, one for test0 and one for test1. I have then calculated for each image which class it belongs to comparing its log likelihood scores. After the calculations I found the accuracy for each test set as the following:

Accuracy\_for\_digit0testset: 0.917346938776

Accuracy\_for\_digit1testset: 0.923348017621

Overall, the project was a great one for me as it clarified many questions, I had about naïve bayes classifiers with normally distributed data. The hardest part for me in this project was to get the gaussian likelihood formula right but in the end, I was able to successfully implement the project.

Barlas YARDIMCI - 1223149330