Report

Problem Description:

For this homework, we applied three different methods for calculating the Naïve Bayes Classification accuracy/error.

Procedure:

The hw3.py code given to us serves to upload data points from a data set in Matlab and identify the sizes of the test and training sets. A ‘split’ is then performed on the data, dictating the sizes of our two sets for each particular run. Also, we can alter the amount of iterations of our code to obtain a more accurate answer for ourselves. The three different methods we used comprised of a full covariance matrix, a diagonal covariance matrix, and a spherical covariance matrix. Inside the implementations of each method, we build our covariance matrix and use the likelihood function to form resulting vectors. The vectors are then compared to the actual data in order to calculate the accuracy of the algorithm for different set ratios and iterations.

Results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Ratio** | **Training Set Error** | **Testing Set Error** |
| Full | 0.1 | 0.005 | 0.016 |
| Full | 0.2 | 0.005 | 0.017 |
| Full | 0.5 | 0.003 | 0.018 |
| Full | 0.8 | 0.002 | 0.023 |
| Full | 0.9 | 0.012 | 0.038 |
| Diagonal | 0.1 | 0.041 | 0.042 |
| Diagonal | 0.2 | 0.041 | 0.043 |
| Diagonal | 0.5 | 0.039 | 0.046 |
| Diagonal | 0.8 | 0.039 | 0.045 |
| Diagonal | 0.9 | 0.039 | 0.051 |
| Spherical | 0.1 | 0.076 | 0.075 |
| Spherical | 0.2 | 0.076 | 0.079 |
| Spherical | 0.5 | 0.077 | 0.081 |
| Spherical | 0.8 | 0.088 | 0.089 |
| Spherical | 0.9 | 0.065 | 0.083 |

We clearly see that Full is the most accurate while Spherical is the least accurate of the three methods. This is because full matrices are more rigorous to calculate and depends on more points of data whereas the other two methods are quicker to execute. This stems from the fact that the likelihood function for the full matrix requires more flops per iteration than diagonal or spherical methods. The Spherical model is useful when we have fewer data points and want to save running time. On the other hand, the diagonal model usually best balances time to run and accuracy of prediction. We also see that for 100 iterations, the result of each ratio converges to the numbers shown above.