Comp 551 – Applied Machine Learning

Programming Assignment 2

Sheldon Benard - 260618386

# Data

The data generated for the assignment be found in the **DS1** and **DS2** directories. **DS1** is synthetic data using a shared covariance matrix, and **DS2** is synthetic data using a blend of 3 covariance matrices.

# Set 1 – Shared Covariance Matrix

For data set 1, probabilistic LDA and Nearest Neighbor Classification were used, and each of their performance measures are highlighted below:

## Probabilistic LDA

Table - LDA Performance

|  |  |
| --- | --- |
| Performance Measure | Value |
| Accuracy | 0.955833333333 |
| Recall | 0.96 |
| Precision | 0.952066115702 |
| F-Measure | 0.95601659751 |

Table 2 - LDA Coefficients

|  |  |
| --- | --- |
| Coefficient | Value |
| w0 | -28.5636438698 |
| w | 0: [-14.97958970904708]  1: [9.0896749013246954]  2: [6.3206668629207439]  3: [3.4597761029438026]  4: [10.146994330221869]  5: [4.2188263350499344]  6: [-18.265797747673204]  7: [25.063182529460299]  8: [30.419582109649703]  9: [-9.346558085086432]  10: [13.840284382503954]  11: [13.516834721154789]  12: [-16.684278842601241]  13: [-13.75113699664751]  14: [6.1327979825897998]  15: [-13.66858669031415]  16: [-31.075980710459266]  17: [6.8884320363175853]  18: [0.88534748424387821]  19: [5.1596672086296262] |

## Nearest Neighbor Classification

|  |  |
| --- | --- |
| Performance Measure | Value |
| Accuracy | 0.544166666667 |
| Recall | 0.553333333333 |
| Precision | 0.543371522095 |
| F-Measure | 0.548307184145 |

Table 3 - kNN Performance (k = 47)

As one can see, Probabilistic LDA outperformed the Nearest Neighbor Classifier by a significant margin. For most k values, the performance indicators hovered around the 52% - 54% range, with a slight edge going to higher k values. The most accurate Nearest Neighbor Classification performance occurred at k = 47, with an accuracy of 54.4%.

# Set 2 – Hybrid of 3 Covariance Matrices

Similarly, for data set 2, probabilistic LDA and Nearest Neighbor Classification were used, and each of their performance measures are highlighted below:

## Probabilistic LDA

Table 4 - LDA Performance

|  |  |
| --- | --- |
| Performance Measure | Value |
| Accuracy | 0.515833333333 |
| Recall | 0.526666666667 |
| Precision | 0.515497553018 |
| F-Measure | 0.521022258862 |

Table 5 – LDA Coefficients

|  |  |
| --- | --- |
| Coefficient | Value |
| w0 | 0.0576514093926 |
| w | 0: [-0.0022240563725057783]  1: [-0.0055789097410681998]  2: [-0.0027194834650929034]  3: [0.014551468683540278]  4: [-0.052670541086871719]  5: [0.063582391488058235]  6: [-0.036255056729106144]  7: [-0.012813378777203205]  8: [-0.010377617774236916]  9: [-0.049340719065959493]  10: [-0.012347150055690453]  11: [0.035082618976222613]  12: [-0.020998269767416103]  13: [0.068007298472954669]  14: [-0.010435103349685587]  15: [0.016446395582976156]  16: [-0.039968549068858855]  17: [0.021512193718622674]  18: [-0.01998561904617599]  19: [0.0031784602777031601] |

## Nearest Neighbor Classification

|  |  |
| --- | --- |
| Performance Measure | Value |
| Accuracy | 0.52 |
| Recall | 0.483333333333 |
| Precision | 0.521582733813 |
| F-Measure | 0.501730103806 |

Table 6 - kNN Performance (k = 83)

Clearly, Probabilistic LDA performed much worse with data set 2. The hybrid of 3 Multivariate Normal Distributions, each with their own means and covariance matrix, was unbecoming for the linear decision boundary that Probabilistic LDA proposed. Further, the Nearest Neighbor Classifier performed quite similarly with both data sets, with its performance being slightly worse for data set 2.

On the difference in performance between the 2 classifiers, if the covariance matrix is shared (data set 1), LDA performs very well (all performance measures above 95%) compared to the Nearest Neighbor Classifier (with performance measures between 54% and 56%).

On the similarities in performance between the 2 classifiers, if the covariance matrix is not shared (data set 2), LDA and Nearest Neighbor Classifier perform similarly (with performance measures around the 51% to 52%).