ASSESSMENT KNOWING TRUE CLASS

The below probability of error and Confusion Matrix are generated by tallying the class of each vector in test set, computed from Take Home 1 and their true class which follows the recurring pattern 2-3-1-3-1-2..

P(Error) = 0.18

Confusion Matrix, based on test set

$$\begin{matrix} 3802 & 490 & 708 \\ 134 & 4795 & 71 \\ 1121 & 198 & 3681 \end{matrix}$$

SEPARATING HYPERPLANES

The explanation of the algorithm used can be found in HoKashyapAlgorithm.pdf file.

Using Hyperplane constructed from class 1 and 2(in that order in Confusion Matrix) and making use of data from these two classes in training set the Confusion Matrix Obtained is this.

$$\begin{matrix} 3360 & 1640 \\ 3097 & 1903 \end{matrix}$$

The ω_{12} is given by $\begin{matrix} 0.971 \\ 0.1 \\ 0.069 \\ 0.074 \\ -0.025 \end{matrix}$

Using Hyperplane constructed from class 2 and 3(in that order in Confusion Matrix) and making use of data from these two classes in training set the Confusion Matrix Obtained is this.

$$\begin{matrix} 3818 & 1182 \\ 926 & 4074 \end{matrix}$$

The ω_{23} is given by $\begin{matrix} -3.154 \\ 0.146 \\ 0.208 \\ 0.141 \\ 0.213 \end{matrix}$

Using Hyperplane constructed from class 1 and 3(in that order in Confusion Matrix) and making use of data from these two classes in training set the Confusion Matrix Obtained is this.

$$\begin{matrix} 2436 & 2564 \\ 1166 & 3834 \end{matrix}$$

The ω_{13} is given by $\begin{matrix} -2.917 \\ 0.12 \\ 0.227 \\ 0.143 \\ 0.108 \end{matrix}$

For classifying each vector in test set all the 3 hyper-planes are used. The 4 hyper-planes divide the 4D space into 7 regions. A vector is classified depending upon in which region it lies.

Confusion Matrix when classification is performed on S_T using the hyperplanes

$$\begin{matrix} 2273 & 137 & 2590 \\ 2661 & 1358 & 981 \\ 1157 & 61 & 3782 \end{matrix}$$

And the P(Error) is 0.5058

k-NNR

1-NNR

For every vector, x_i in S_T compute its distance from each of the 15000 vectors in H. Classify x_i into that class where the nearest vector in H with respect to x_i falls. Mahalanobis distance measure is used in computing the distance.

Confusion Matrix for k = 1 computed on S_T

$$\begin{matrix} \mathbf{3269} & \mathbf{550} & \mathbf{1181} \\ \mathbf{390} & \mathbf{4419} & \mathbf{191} \\ \mathbf{1291} & \mathbf{269} & \mathbf{3440} \end{matrix}$$

For k = 1, P(Error) = 0.26 computed on S_T

3-NNR

For every vector x_i in S_T find out the 3 most nearest vectors in H(by computing the Mahalanobis distances and picking out 3 smallest ones). Classify x_i into that class where majority of these 3 vectors falls.

Confusion Matrix for k = 3 computed on S_T

$$\begin{matrix} 3532 & 562 & 906 \\ 368 & 4543 & 89 \\ 1294 & 254 & 3452 \end{matrix}$$

For k=3, P(Error) = 0.23 computed on S_T

5-NNR

For every vector x_i in S_T find out the 5 most nearest vectors in H(by computing the Mahalanobis distances and picking out 5 smallest ones). Classify x_i into that class where majority of these 5 vectors falls.

Confusion Matrix for k = 5 computed on S_T

$$\begin{matrix} 3550 & 602 & 848 \\ 274 & 4651 & 75 \\ 1285 & 295 & 3420 \end{matrix}$$

For k=5, P(Error) = 0.23 computed on S_T

PCA

Steps followed:-

1. Compute the mean of all vectors in H and subtract from each vector the mean to form a new set whose expected value is 0.
2. Covariance Matrix on the new set is computed.
3. Eigen Vectors and Eigen Values of the covariance matrix is computed.
4. Eigen Values are reverse sorted and corresponding Eigen Vectors are arranged as well, as Eigen Values and Eigen Vectors always move in pair. The Eigen Vectors are of unit magnitude.
5. Eigen Vectors corresponding to top two Eigen Values are selected because by doing so we will be able to preserve the top two features in which variance among the vectors is maximum.
6. New data set with reduced dimension is derived by multiplying the training set with 0 mean to the matrix of top two Eigen Vectors.
7. On this new data set, Bayesian Classification is applied to compute density function of the three classes and based on that the confusion matrix and error probability on training set is computed

Confusion Matrix Based on H(Reality Check)

$$\begin{matrix} 2268 & 525 & 2207 \\ 80 & 4838 & 82 \\ 1823 & 448 & 2729 \end{matrix}$$

Based on H, P(Error) = 0.344

8. Subtract the same mean from each vectors in test data and reduce the dimensionality to 2 by multiplying it with the matrix of top two Eigen Vectors.
9. Use the vectors form the resultant matrix obtained from the previous step and the density function computed in step 7 to determine class of each vector in test set.

Confusion Matrix Based on S_T

$$\begin{matrix} 2263 & 551 & 2186 \\ 95 & 4818 & 87 \\ 1843 & 453 & 2705 \end{matrix}$$

Based on S_T, P(Error) = 0.3476

The probability of error, computed from test set has gone up significantly. When all the four features were used the probability was 0.18 but now it is 0.3476 which is almost double.