**ISTANBUL TECHNICAL UNIVERSITY**

**COMPUTER ENGINERING DEPARTMENT**

**BLG 527E MACHINE LEARNING**

**CRN: 13817**

**Instructor: Zehra Çataltepe**

**Homework #2**

**November 6, 2017**

**Tuğrul Yatağan**

**504161551**

**Running Code**

q1.r and q2.r R files can be directly run. They search "optdigits.tra" and "optdigits.tes.txt" files in the same directory. If R files and data files are different directory, “trainingFile” and “testFile” variables must be set accordingly. It took about 30 minutes for q1.r and about 5 minutes for q2.r to run.

**Answers**

**Q1a-b)**

Steps for multivariate analysis:

1. Eliminate feature column in training data set if whole column is 0 or fixed value to be sure within-class scatter matrix is invertible.
2. Calculate d-dimensional mean vectors per class. Create meanij matrix.

, with i = 0,1,2,3,...,9

1. Calculate common variance matrix

where covariance matrix

1. Get diagonal elements of common variance matrix as vector s
   1. Use vector s as it is if common covariance matrix and Σ is diagonal
   2. Use square root mean of s if common covariance matrix) and Σ=s2I for some s>0
2. Discriminant function:
3. Run discriminant function for dataset and get prediction results.
4. Create k X k confusion matrix.

**Q1c)**

1. Training error of Q1a:

8.579649 %

1. Test error of Q1a:

10.68447 %

1. Training error of Q1b:

8.16113 %

1. Test error of Q1b:

10.62883 %

1. Test error for each class for Q1a:

1.117, 22.413, 13.812, 5.389, 6.077, 5.494, 3.333, 7.978, 19.883, 21.134

1. Test error for each class for Q1b:

1.129, 22.285, 10.285, 5.263, 6.250, 7.065, 3.846, 8.602, 16.867, 23.414

1. Confusion matrix of the Q1a:

truth

prediction 0 1 2 3 4 5 6 7 8 9

0 177 0 1 0 0 0 1 0 0 0

1 0 135 6 0 4 0 5 0 21 3

2 0 21 156 2 0 0 0 0 1 1

3 0 0 0 158 0 1 0 0 1 7

4 1 0 1 0 170 1 0 3 0 5

5 0 1 0 3 0 172 0 0 2 4

6 0 4 0 0 0 1 174 0 1 0

7 0 0 1 7 3 0 0 173 1 3

8 0 8 10 6 3 0 1 2 137 4

9 0 13 2 7 1 7 0 1 10 153

1. Confusion matrix of the Q1b:

truth

prediction 0 1 2 3 4 5 6 7 8 9

0 175 0 1 0 0 0 1 0 0 0

1 0 136 6 0 8 0 3 0 19 3

2 0 16 157 1 0 0 0 0 1 0

3 0 1 2 162 0 0 0 0 1 5

4 2 0 0 0 165 1 0 3 0 5

5 1 1 0 2 0 171 0 2 3 4

6 0 5 0 0 0 1 175 0 1 0

7 0 0 2 7 3 0 0 170 1 3

8 0 7 5 5 4 0 2 2 138 3

9 0 16 4 6 1 9 0 2 10 157

Training data sets gives better error than test sets as expected. Q1b (Σi = Σ (common covariance matrix) and Σ=s2I for some s>0) assumption gives best test error.

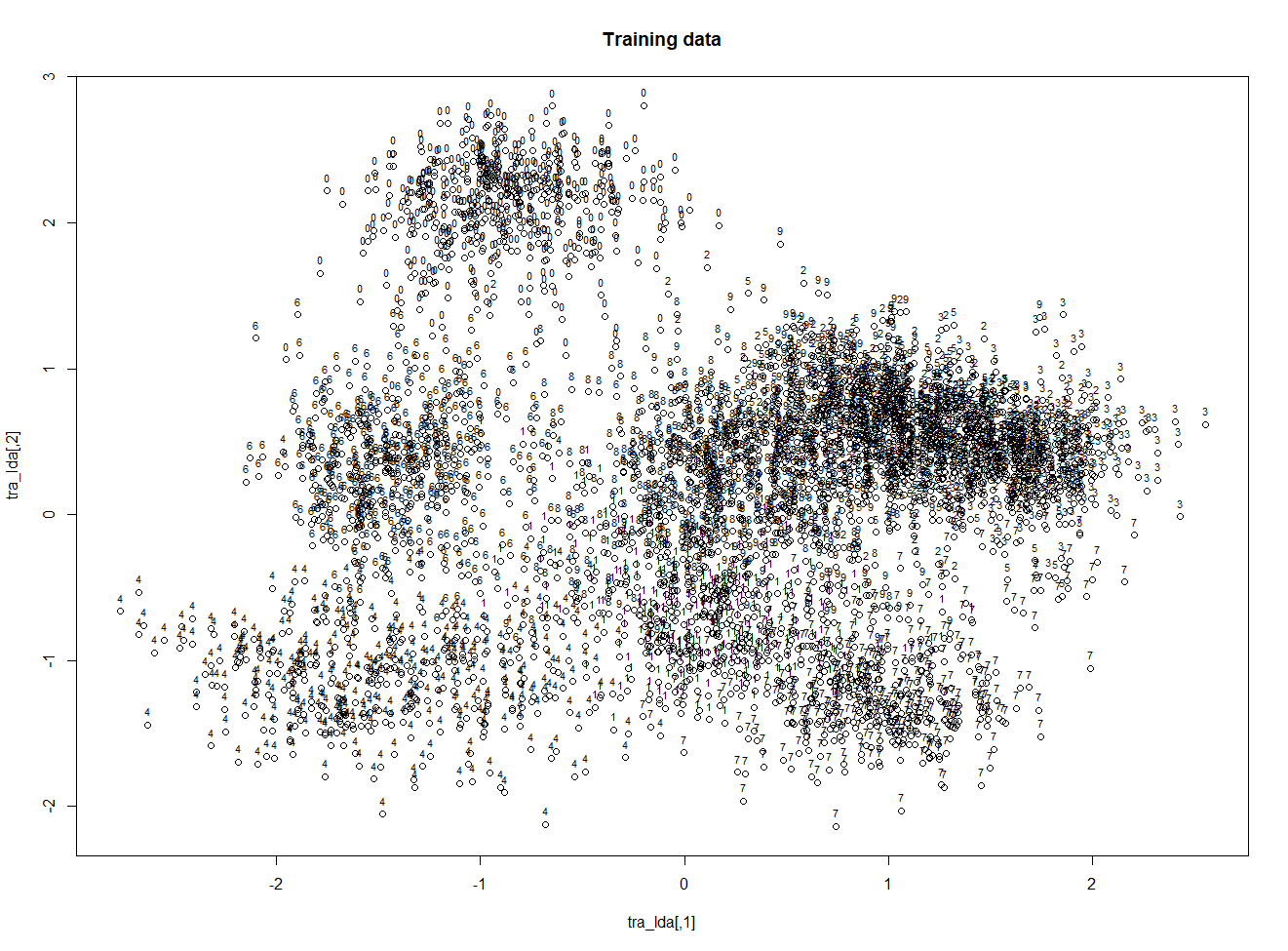
Classes (1 & 2) and (1 & 8) have confused each other most.

**Q2a)**

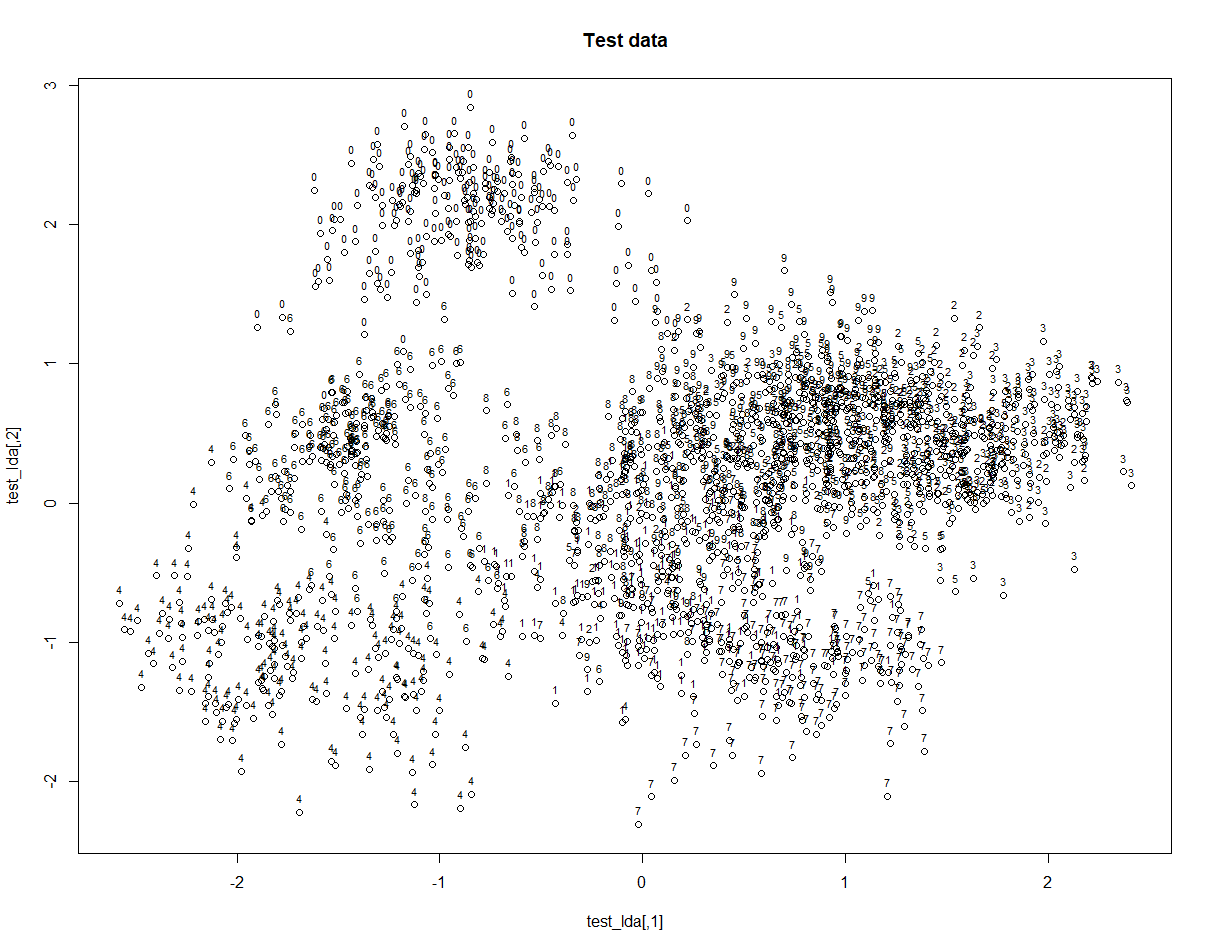
1. Eliminate feature column in training data set if whole column is 0 or fixed value to be sure within-class scatter matrix is invertible.
2. Calculate d-dimensional mean vectors per class. Create meanij matrix.

, with i = 0,1,2,3,...,9

1. Calculate scatter matrices
   1. Calculate within-class scatter matrix SW
   2. Calculate between-class scatter matrix SB
2. Solving the generalized eigenvalue problem for the matrix
3. Choosing 2 eigenvectors with the largest eigenvalues
4. Transforming the samples onto the new subspace for training and test data set by calculated eigenvectors from training data set.
5. Visualize training data set and test data set LDA results.
   1. Training data result:



* 1. Test data result:



**Q3)**

LDA gives better for dimensionality reduction over PCA since, PCA doesn’t use class information to reduce dimension.

We are interested in the matrix W that maximizes:

The largest eigenvectors is the solution, and PCA within-class scatter matrix SW gives bigger value over LDA which leads to worse dimension reduction.