**MACHINE LEARNING FROM DATA**

**Fall 2018**

**Report: Lab Session 2 – Feature selection: PCA and MDA**

**Names:**

**Group:**

Instructions

* Answer the questions
* Save the report and upload the file

Questions

Q1. Include the plots of the phoneme spectra.

Q2. Include the error probabilities for the training and test sets obtained with the linear classifier (LC) and the quadratic classifier (QC), using all the features. Discuss the results.

Q3. Include the confusion matrices for the test set obtained with the linear classifier (LC) and the quadratic classifier (QC), using all the features. Discuss the results.

Q4. Which features would you choose? Show the error probabilities for the training and test sets obtained with the linear and the quadratic classifier. Compare with the previous case (using all features) and discuss the results.

Q5. Include the scatter plot and decision boundaries obtained between class ‘aa’ and all the other (four) classes. Discuss the results.

Q6: Complete the table with the training and test errors for the linear (LC) and the quadratic (QC) classifiers when using three, two and one feature, and SNR=10dB. In this case PCA is used for feature selection. Discuss the results. Analyze the scatter plots in two dimensions and in one dimension.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 3 features | | 2 features | | 1 feature | |
| Test | Train | Test | Train | Test | Train |
| LC |  |  |  |  |  |  |
| QC |  |  |  |  |  |  |

Q7: Complete the table with the training and test errors for the linear (LC) and the quadratic (QC) classifiers when using three, two and one feature, and SNR=10dB. In this case MDA is used for feature selection. Discuss the results. Analyze the scatter plots in two dimensions and in one dimension.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 3 features | | 2 features | | 1 feature | |
| Test | Train | Test | Train | Test | Train |
| LC |  |  |  |  |  |  |
| QC |  |  |  |  |  |  |

Q8: By rotating the figure observe that there are 2D projections where projected clusters are well separated while there are other projections where projected clusters overlap. Copy a pair of examples illustrating this point

Q9: Complete the table with the training and test errors for the linear (LC) and the quadratic (QC) classifiers when using three, two and one feature, and SNR= 0 dB. Use PCA for feature selection. Discuss the results. Analyze the scatter plots in two dimensions and in one dimension.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 3 features | | 2 features | | 1 feature | |
| Test | Train | Test | Train | Test | Train |
| LC |  |  |  |  |  |  |
| QC |  |  |  |  |  |  |

Q10: Complete the table with the training and test errors for the linear (LC) and the quadratic (QC) classifiers when using three, two and one feature, and SNR= 0 dB. Use MDA for feature selection. Discuss the results. Analyze the scatter plots in two dimensions and in one dimension.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 3 features | | 2 features | | 1 feature | |
| Test | Train | Test | Train | Test | Train |
| LC |  |  |  |  |  |  |
| QC |  |  |  |  |  |  |

Q11: observe by rotating the figure that there are 2D projections where projected clusters are well separated while there are other projections where projected clusters overlap. Copy a pair of examples illustrating this point

Q12. Find and write the three vectors corresponding to the class means. Give also the value of the seed used in your experiments.

Q13. Which is the rank of the matrix Sb? How many features can we use with MDA?

Q14. Complete a table with the training and test errors for the linear (LC) and the quadratic (QC) classifiers when using three, two and one feature, and SNR= -5 dB. Use PCA and MDA for feature selection. Discuss the results. In which cases is MDA clearly better than PCA?

Q15. Show the error curves for the linear and the quadratic classifier on the training and on the test set. Copy your code in Annex1

Q16. Discuss which dimension is the most adequate for the linear classifier and which is the best one for the quadratic classifier. Remember that it is important not to overfit on the training data (the test error should not be much larger than the training error).

Q17. Which is the maximum number of features dmax?

Q18. Show the error curves for the linear and the quadratic classifier on the training and on the test set. Copy your code in Annex2

Q19. Compare results and discuss the use of PCA and MDA for the Phoneme dataset

**Annex 1. Matlab code for Q15 (PCA)**

**Annex 2. Matlab code for Q18 (MDA)**