MACHINE LEARNING FROM DATA
Fall 2018

Lab Session 2 – Feature selection: PCA and MDA

1. Goal 2
2. Instructions 2
3. Introduction and previous study 2
4. The Phoneme dataset 2
   4.1. Characteristics of the dataset 2
   4.2. Classification using all the features or a manually selected subset 3
5. Feature selection on a synthetic Gaussian dataset 4
6. Feature selection on the Phoneme dataset 5
   6.1. Dimensionality reduction using PCA 5
   6.2. Dimensionality reduction using MDA 6

1. Goal

The goal of this session is to

- Use and compare two methods for dimensionality reduction: principal component analysis (PCA) and multiple discriminant analysis (MDA or Fisher discriminant analysis)
- Test the two methods using synthetic and real datasets
- Compute and plot decision boundaries for three classes

2. Instructions

Getting the material:

- Download and uncompress the file ML_Lab2.zip

Handling your work:

- Answer the questions in the document Lab2_report_yourname.pdf
- Save the report and Matlab code (if necessary) in a folder, and upload the compressed folder (zip, rar).

3. Introduction and previous study

Read the slides corresponding to lecture 3: feature selection by dimensionality reduction using PCA and MDA.

For both methods we will build the total scatter matrix **S** and generate a transform matrix **W**.
For PCA, matrix **W** contains the eigenvectors of **S** associated with the largest eigenvalues.

For MDA, matrix **W** contains the eigenvectors associated with the largest eigenvalues of $\mathbf{S}_C^{-1}\mathbf{S}_B$, where $\mathbf{S}_B$ is the between-class scatter matrix and $\mathbf{S}_C$ is the sum of the covariance matrices of all the classes and measures the intra-class scatter.

4. The Phoneme dataset

4.1. Characteristics of the dataset

The Phoneme dataset can be found here: https://web.stanford.edu/~hastie/ElemStatLearn/data.html

This is the information provided by the authors:

```
These data arose from a collaboration between  Andreas Buja, Werner
Stuetzle and Martin Maechler, and we used as an illustration in the
paper on Penalized Discriminant Analysis by Hastie, Buja and
Tibshirani (1995), referenced in the text.

The data were extracted from the TIMIT database (TIMIT
Acoustic-Phonetic Continuous Speech Corpus, NTIS, US Dept of Commerce)
which is a widely used resource for research in speech recognition.  A
dataset was formed by selecting five phonemes for
classification based on digitized speech from this database.  The
phonemes are transcribed as follows: "sh" as in "she", "dcl" as in
```

```
"dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and
"ao" as the first vowel in "water".  From continuous speech of 50 male
speakers, 4509 speech frames of 32 msec duration were selected,
approximately 2 examples of each phoneme from each speaker.  Each
speech frame is represented by 512 samples at a 16kHz sampling rate,
and each frame represents one of the above five phonemes.  The
breakdown of the 4509 speech frames into phoneme frequencies is as
follows:

  aa   ao dcl   iy  sh
 695 1022 757 1163 872

From each speech frame, we computed a log-periodogram, which is one of
several widely used methods for casting speech data in a form suitable
for speech recognition.  Thus the data used in what follows consist of
4509 log-periodograms of length 256, with known class (phoneme)
memberships.

The data contain 256 columns labelled "x.1" - "x.256", a response
column labelled "g", and a column labelled "speaker" identifying the
different speakers.
```

Use Matlab script `lab2_phonemes.m`

Read the script and identify code sections for:

- Reading the dataset and selecting a subset of 64 features for each observation (from the original vector of size 256)
- Removing the sample mean for each observation
- Plotting the vectors (using all the 256 features, up to 8kHz)
- Dividing the dataset into training (70%) and test (30%) subsets. Note that the partition is random and it keeps the same proportion of training/test observations for each class
- Classifying the data using linear and quadratic classifiers
- Computing error probabilities and confusion matrices
- When using 2 features, showing a scatter plot and decision boundaries between class ‘aa’ and all the other classes.

4.2. Classification using all the features or a manually selected subset

Run the script `lab2_phonemes.m`
Answer the following 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.

Suppose that due to computational issues we can only use two features per observation. Considering the plots in Q1, manually select two features that seem to be the most discriminative.

Use the variable `v-coor`

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.

5. Feature selection on a synthetic Gaussian dataset

Now we will use script `lab2_gauss_main.m`.
This script generates training and test sets with 3-dimensional vectors (d=3) corresponding to three classes (c=3). The classes follow Gaussian distributions, with mean and covariance matrices specified in the code.
A linear and a quadratic classifier are used to classify the data, and error probabilities are computed. Next, the classification is repeated using two features and then using only one feature. The goal is to compare the performance of the classifiers as we reduce the number of features using PCA or MDA.

Run the script `lab2_gauss_main.m` .
Select the 'PCA' option and input one value (seed) to initialize the random number generator. Use SNR = 10dB. Different values of the seed parameter will produce different eigenvectors for the class covariance matrices.

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.

Repeat the analysis for MDA, using the same seed to generate exactly the same dataset.
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.

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

Repeat the previous analysis using a different SNR value, SNR= 0 dB:

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.

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.

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

From now on, and to check the effectivity of the previous techniques, we will use a Gaussian dataset where the three class centroids (i.e. the means) are colinear and the three classes have identical covariance matrices. Therefore, the alignment direction of the classes corresponds to the direction of minimum variance.
Edit the file `lab2_gauss_main.m` and change the instruction `lab2_gengauss` by `lab2_gengauss_al`

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?

Repeat the previous analysis for PCA and MDA and SNR = -5 dB:

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?

6. Feature selection on the Phoneme dataset

6.1. Dimensionality reduction using PCA

Previously we have reduced the dimensionality of the Phoneme dataset by observing the data and manually selecting features. In this section we will reduce dimensionality following a principal component analysis criterion. We will use the function pca.m which returns a projection matrix with the principal eigenvectors. Check that vectors are orthogonal.

Edit a new script named `lab2_phonemes_PCA.m`.
Read the complete dataset (256-dimensional vectors). Perform feature selection using PCA, and then apply a linear and a quadratic classifier. Compute the training and test classification error for the two classifiers when using a number of features d' ranging from 1 to 256 (d'=1:256).
Show the four error curves (training/test errors for LC and QC as a function of the dimension d') **in the same figure**, using different colours for the curves.

Note: use the training dataset to find the projection matrix and to train the classifie. The test dataset should only be used for evaluating the error probabilities.

You can use the 'help' command to review the documentation for *figure*, *plot, hold*, *legend, title, grid* and other useful functions

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).

6.2. Dimensionality reduction using MDA

In this section we will repeat the previous analysis, but using MDA instead of PCA for dimensionality reduction.

Modify the code in order to use MDA for a number of features d' ranging from 1 to *dmax* (d':1:*dmax*). The function `mda_ml.m` returns the projection matrix with the generalized eigenvectors.

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