MACHINE LEARNING FROM DATA

Lab Session 2 – Feature selection - PCA

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 the Phoneme dataset 4
5.1. Dimensionality reduction using PCA 4

1. Goal

The goal of this session is to use principal component analysis (PCA) for feature selection or dimensionality reduction

2. Instructions

- Download and uncompress the file **Mlearn_Lab2.zip**
- Answer the questions in the document **Mlearn_Lab2_report_surnames.doc**

3. Introduction and previous study

Read the slides corresponding to lecture 3: feature selection by dimensionality reduction.

Read and run the code in the Colab Notebook **Mlearn_lab2_1_IntroPCA.ipynb** to understand the use of PCA for dimensionality reduction.

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.
```

Open Colab Notebook **Mlearn_lab2_2_Phoneme.ipynb**

Read the notebook 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 classes.

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

Run the code in **Mlearn_lab2_2_Phoneme.ipynb**

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. Discuss the results.

5. Feature selection on the Phoneme dataset

5.1. 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.

Edit a new Colab Notebook named **Mlearn_lab2_3_PhonemePCA_surname.ipynb**

Write the code to:
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 `nfeat` ranging from 1 to 256.
Show the four error curves (training/test errors for LC and QC as a function of the dimension `nfeat`) **in the same figure**, using different colours for the curves.

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

Q6. Show the error curves for the linear and the quadratic classifier on the training and on the test set.

Q7. 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).