INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES ROUEN

IML
Exam 2019/2020

DÉPARTEMENT ARCHITECTURE DES SYSTÈMES D'INFORMATION

- Duration: 2h, *Access to handouts, notes and codes is granted*
- Deliverables: for each exercise name your main script following the rules below. Upload on Moodle an archive (same naming rules) containing the main script files and eventually your own libraries required for running the scripts
- How to name the main script? `lastname_firstname_exercise_num.ext`

Reminders Activate Python virtual environment and launch your notebook (or spyder) by running in the terminal

```
>>> source /opt/venv/spyder/bin/activate
>>> jupyter-notebook&
>>> spyder3&
```

To get the help of a function, see an example below.

```
from sklearn.preprocessing import StandardScaler
# to get the online help, type:
>>> ?StandardScaler
```

1 Customers segmentation (8 points)

A supermarket has collected data on the purchased items of customers. These data are related to the annual spending on fresh, frozen, milks products, etc or on the geographical location of the customers. The data are provided in the file `customers_data.csv` (with a header and delimiter being ","). The first 6 columns are continuous variables and the remaining columns are categorical (they represent the location of the supermarket's channel).

The objective is to segment these customers into K clusters.

1. Perform a statistical analysis of the variables. According to the boxplots do you need to normalize the data?
2. Write a program to achieve the clustering of the customers into $K = 5$ clusters using K-means method.
3. We want to visualize the obtained clusters.
   (a) Use PCA to project the data into 2D and represent the clusters in a two-dimensional plot with different symbols and colors.
   (b) Answer the same question using t-SNE method.
   (c) Which method gives the best visualization of your segmentation? Discuss the results.
4. We need to select the "optimal number" of clusters K. Propose a procedure to select the optimal K (the code is not required)

2 Commercial Block Detection in TV videos (12 points)

The objective is to automatically identify commercial blocks (advertisements) in TV news using machine learning methods. For this most approaches extract different features from the raw videos. Those features can be the fundamental frequencies of video sounds (as musics are more abundant in advertisements compare to speech in the news), time-frequency transform applied to the sounds, distribution of the difference between video frames, text information...

We have at disposal two files: the training data `commercial_train.csv` and the test data `commercial_test.csv` collected from a TV channel. The last column represents the **label +1/-1 (Commercials/Non Commercials)**. The remaining columns represent the features which descriptions are given in Table 1.

Type of Feature	Feature Index
Shot Length	0
Motion Distribution(Mean and Variance)	1 - 2
Frame Difference Distribution (Mean and Variance)	3 - 4
Short time energy (Mean and Variance)	5 - 6
ZCR (Mean and Variance)	7 - 8
Spectral Centroid (Mean and Variance)	9 - 10
Spectral Roll off (Mean and Variance)	11 - 12
Spectral Flux (Mean and Variance)	13 - 14
Fundamental Frequency (Mean and Variance)	**15 - 16**
Motion Distribution (40 bins)	17 - 57
Frame Difference Distribution (32 bins)	**58 - 90**
Text area distribution (15 bins Mean and 15 bins for variance)	**91 - 121**

Table 1: Features of commercials detection dataset

1. Load the data. Is the classification problem balanced?

In this first part the goal is to find the most important features to correctly classify commercials and non-commercials. For this, we will learn a linear logistic regression model with ℓ_2 penalty. We will investigate three types of features: **Fundamental Frequency**, **Frame Difference Distribution**, **Text area distribution**. These features are indicated in bold font in Table 1.

2. For each of these 3 types of features
   (a) extract the corresponding data (columns)
   (b) learn accordingly your optimal logistic regression model (you should highlight how the best model is selected)
   (c) evaluate its classification accuracy on the test set

Remark: given the matrix $X \in \mathcal{R}^{n \times d}$, we can extract the samples corresponding to (for instance) fundamental frequency in the following way:

```
import numpy as np
# extracting columns 15-16 (fundamental frequency)
index_col = np.arange(15,17)
data_train = X[:, index_col]
```

3. Which type of feature provides the best results? Justify your answer.

In the remainder, you will only proceed with the best features you have previously identified. At the second step of the exercise, we want to improve the classification accuracy by looking for a non-linear SVM model.

4. Learn a non-linear SVM with gaussian kernel.
   (a) highlight the selection of the optimal model
   (b) evaluate the classification accuracy on the test set. Show the confusion matrix.

5. Have you improved the detection rate of commercials? Discuss the obtained results.