Machine Learning — Programming Assignment

News classification

Claudio Cusano

Deadline 26–6–2020, 16.00

1 Problem definition and data

We want to build a system able to automatically classify news into four categories: “business”, “health”, “entertainment” and “science”. For each news we only know the publisher and the title. The information are stored in text files, one line per each news article, in the format `class|publisher|title`. The following lines are an excerpt of the data set that has been collected for the problem:

```
entertainment|Reuters|A Minute With: Kermit and Miss Piggy bicker about lines and love
business|Reuters|UPDATE 2-Interpublic revenue gets a boost from UK, beats estimate
business|NASDAQ|Level 3 to Buy TW Telecom for $5.7 Billion -- 2nd Update
entertainment|Examiner.com|'Game of Thrones' discussion of 'The Lion and the Rose' and who is the killer
business|CNET|Lenovo earnings soar on solid PC, smartphone sales
entertainment|Examiner.com|Dr. Oz: Lea Michele's sugar-free diet, slimming drink, weight loss breakfast
business|Fox News|Zillow CEO on buying up competitor Trulia amid rush of corporate mergers
business|Fox News|Iran's President Rouhani underscores his outreach to West, moderate policies at  ...
entertainment|Fox News|Man sues British Airways after booking mistake sends him Grenada instead of  ...
```

The data set has been split into training, validation and test sets including 10 000, 1000 and 1000 lines, respectively. Data is distributed as compressed text files, that you can read as follows:

```python
import gzip

f = gzip.open("train.txt.gz", "rt")
for line in f:
    klass, publisher, title = line.split("|")
    # ...
```

2 Assignment

As a programming assignment you are expected to:

1. design and implement a feature extraction procedure;

2. implement, train and evaluate one or more classification models;

3. use suitable data processing and visualization techniques to analyze the behavior of the trained model(s).

All the above should be implemented as scripts in the Python programming language. Any machine learning library (included `pvml`) can be used. Should your computer struggle to process all the data, you can start working on a reduced version of the training set and, before submitting, use as much data as you can (in that case remember to document how large the subset is for each result obtained).

3 Report

Prepare a report of two to four pages documenting all your work. Provide detailed instructions on how to reproduce the results. The report must be in the PDF format. Include your name in the report and conclude the document with the following statement: "I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher."

Make a ZIP archive with the report and the Python scripts, and send it by e-mail before the deadline. To keep the size of the submission manageable, do not include files containing the original data, the features etc.

Deadline for the submission: Friday, June 26 at 16.00.