INF519 Machine Learning 2: Homework 2
Filtering spam messages

Jae Yun JUN KIM*

October 6, 2017

Evaluation: Upload your report and python script(s) (in group of up to 3 people) to the course website available in the campus.ece.fr.

Due: Thursday October 12, till 23h55m
Remark:

- No late homework will be accepted.
- No plagiarism. If plagiarism happens, both the "lender" and the "borrower" will have a zero.

Similar to the `Example 4 (filtering spam emails)` given in the `Practice 2` document, this time you are asked to filter spam messages. These messages are stored in `messages.txt` file and can be downloaded from the course website. Each line in this file corresponds to spam (or ham) label and a message.

1 Tasks

1. Divide the data in two groups: training and test examples.
2. Parse both the training and test examples to generate both the spam and ham data sets.
3. Generate a dictionary from the training data.
4. Extract features from both the training data and test data.
5. Using both Naive Bayes and SVM, fit the respective models to the training data.
6. Make predictions for the test data.
7. Measure the spam-filtering performance for each approach through the confusion matrix.
8. Discuss the obtained results.

2 Deliverables

1. A brief report that contains:
   - Description of the problem
   - Methods
   - Results
   - Discussion on the results
2. Python script(s).
3. Data files.

Important Note: Send these files zipped with your names as follows:

inf519_2017_hw2_studentsNames.zip

*ECE Paris Graduate School of Engineering, 37 quai de Grenelle CS71520 75 725 Paris 15, France; jae-yun.jun-kim@ece.fr