INF436 Machine Learning / Homework 1

Filtering spam messages

Jae Yun JUN KIM*

February 5, 2019

Due: Before the next lab session.
Evaluation: Interrogation during the next lab session about:

- code (in group of up to 3 people)
- (theoretical, practical) questions (individual)

Remark:

- Only groups of one/two/three people accepted. Forbidden groups of larger number of people.
- No late homework will be accepted.
- No plagiarism. If plagiarism happens, both the "lender" and the "borrower" will have a zero.
- Code yourself from scratch. No homework will be considered if you solve the problem using any ML library.
- Do thoroughly all the demanded tasks.
- Study the theory for the interrogation.

---

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. Implement the Naive Bayes from scratch, 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 your results.

*ECE Paris Graduate School of Engineering, 37 quai de Grenelle 75015 Paris, France; jae-yun.jun-kim@ece.fr