INF519 Machine Learning 2 / Homework

Filtering spam messages using Naïve Bayes

Jae Yun JUN KIM*

October 6, 2021

Due:

- **Group 2**: Before Wednesday October 13 2021 8h30.

- **Group 1**: Before Wednesday October 13 2021 14h.

Evaluation: Submit your actual code (no image capture) and explanation of the code and results through campus.ece.fr

Remark:

- Only groups of two or three people accepted (preferably three). Forbidden groups of fewer or larger number of people.

- Submit your homework before the due time. Otherwise the penalty system explained in the syllbus file will be applied.

- 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 questions.

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 Naïve Bayes from scratch, and fit it to the training data.

6. Make predictions for the test data.

7. Measure the spam-filtering performance for each approach through the confusion matrix, precision, and recall.

8. Discuss your results.

---

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