# Naïve Bayes classifier

Introduction to Machine Learning Experiment 1

Spring 2018

# Goal

- Implement a Naïve Bayes classifier and test it on a real dataset
- Have basic ideas about:
  - How to implement and apply a machine learning algorithm on a practical dataset
  - How to evaluate its performance
  - How to analyze your results
- NOTE: all these parts are very important, and should be included in your code/report.

# Naïve Bayes classifier

- Assume that: $P(y|x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$
- Training:
  - Estimate $P(y)$ and $P(x_i|y)$
- Test:
  - Output $\hat{y} = \text{argmax}_y P(y) \prod_{i=1}^{n} P(x_i|y)$

# Data

- The Chinese E-mail Data set: (https://plg.uwaterloo.ca/~gvcormac/treccorpus06/)
- Aims at detecting spams.
- We provide a utf-8 and cut version on http://learn.tsinghua.edu.cn/
- Each e-mail is a text file.

# Data (cont.)

- Files:
  - ./data/: data files, 64620 files
  - ./data_cut/: data files (cut), 64620 files
  - ./label/index (./full/index): labels, each row contains a label (spam, ham) and a relative path.

# How to Evaluate the Performance

- Train your classifier on training set and test it performance on test set
- At least report the accuracy:
  - $Accuracy = \frac{\text{number of correctly classified records}}{\text{number test records}}$
- You are welcome to learn about, and then use other evaluation metrics (e.g. precision, recall or F1)

# How to Analyze Your Results

- What is the issue that you encounter?
- How do you address the issue?
  - how do you design the experiment?
  - how do you modify the algorithm?
- Does your solution work or not?
  - Does the classification performance improve?
- And finally try to explain why your solution works (or why it does not)

# Issue 1: the impact of the size of training set

- How does the size of training set influence the classification performance?
- Suggested solution:
  - Sample 5%, 50% and 100% from the whole training set to train your model
  - Repeat the random sampling (5 times) and report min/max/average accuracy

# Issue 2: zero-probabilities

- Suppose on training set, no records with $x_i = k, y = c$
- Then $\hat{P}(y = c|x_1, \dots, x_i = k, \dots, x_n) = 0$
- (why is this an issue? When does it likely to happen?)
- Possible solution:
  - Smoothing: $\hat{P}(x_i = k|y = c) = \frac{\#\{y=c, x_i=k\}+\alpha}{\#\{y=c\}+M\alpha}$

# Issue 3: specific features

- Are there any specific features expect for bag-of-words?
- Possible:
  - Sender ends with .edu?
  - Time?
  - Client/Software/Mailer?
  - 130-xxxx-xxxx?

# Requirement

- Implement a Naïve Bayes classifier (30% of the overall score)
- Address all 3 mentioned issues:
  - Issue 1 (30%)
  - Issue 2&3 (2*20%=40%)
- NOTE: the score is not based on the performance (i.e. the accuracy) of your model, but how you implement the algorithm, evaluate its performance and analyze your results.

# Submission

- A zipped file that contains:
- Source Code:
  - No limit to program language
  - With necessary comments
  - Make sure the TA can understand/run your code
- README
  - A text file (in utf8 encoding)
  - Includes your name, student id and contact information (the TA may give you feedbacks if you submit before the deadline:) )
- Report
  - A PDF file
  - Experiment design/results/analysis/discussion
  - Don’t just copy&paste the source code

# Deadline & other Information

- Deadline: 2018.03.29 Thursday 23:59
  - Upload the zipped file, with your name and student id in the filename, to learn.tsinghua.edu.cn
  - Late submissions WILL NOT BE ACCEPTED
    - In fact, they will. But the final score will be at most the original score*0.8.
- Contact the TA:
  - shisy13@outlook.com
  - 18505555325

# Reference

- http://scikit-learn.org/stable/modules/naive_bayes.html (for smoothing and Gaussian estimation for continuous attributes)