Natural Language Processing Applications

Week 4: Text classification

fit@hcmus
KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

fit@hcmus

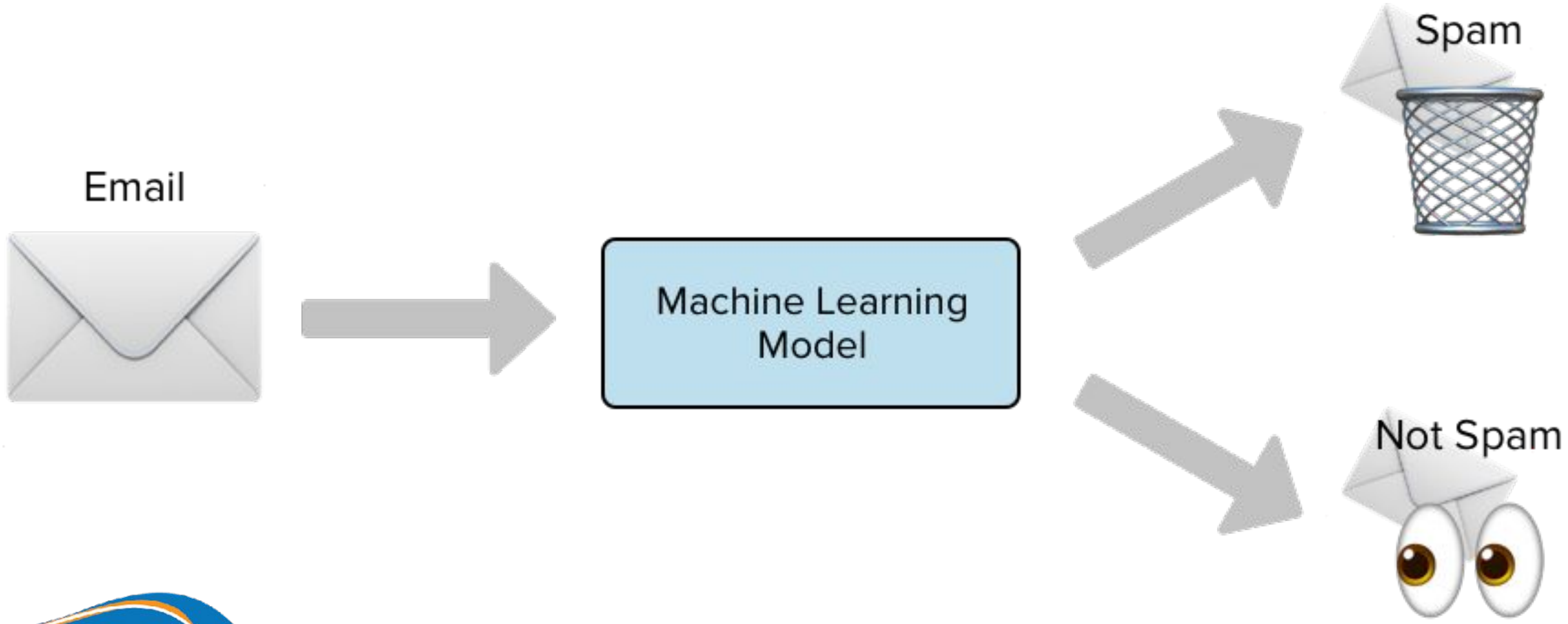
- Introduction
- Naive Bayes
- Evaluation

NLP Applications - Text classification

Introduction

Introduction

- Spam or not-spam?



fit@hcmus

Introduction (cont.)

- Is the author male or female?
  - By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
  - Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

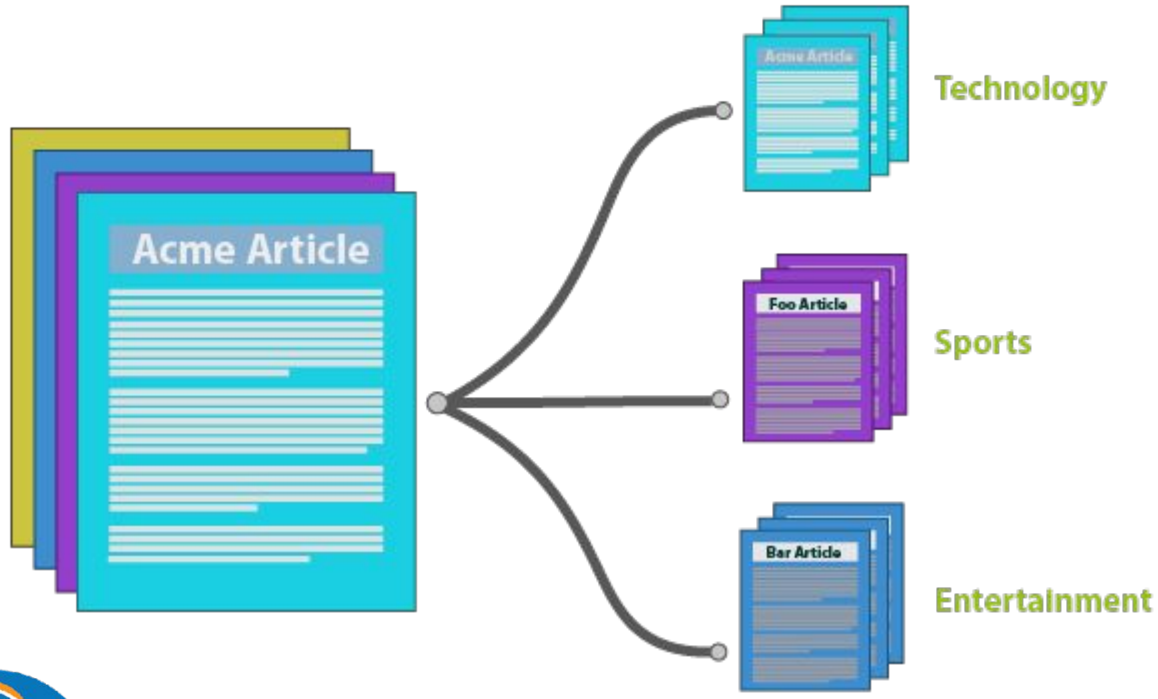
Introduction (cont.)

- Is a film good or bad?
  - Extremely disappointed 
  - Full of famous celebrities, good acting, role-playing 
  - The best film I've ever watched. 
  - I regret wasting my time on watching this film. 

fit@hcmus

Introduction (cont.)

- What genre is the article?



Introduction (cont.)

- Problem definition:
  - Input:
    - Document d
    - A fixed set of categories $C = \{c_1, c_2, \ldots, c_j\}$
  - Output:
    - A predicted category $c \in C$

Introduction (cont.)

- Classification approach:
  - Rule-based:
    - Combination of terms and features
      - spam: black-list-address OR (“dollars" AND “have been selected")
    - The result can be high
      - If the rules are defined clearly by the experts.
      - Build/maintain the rules is costly.

Introduction (cont.)

- Classification approach:
    - Machine-based:
        - Input:
            - Document d
            - A fixed set of categories $C = \{c_1, c_2, \ldots, c_j\}$
            - Training datasets consist of m tagged documents $(d_1, c_1), \ldots, (d_m, c_n)$
        - Output:
            - Classifier ɣ:d -> c

Introduction (cont.)

- Classifiers:
  - Naive Bayes
  - Logistic Regression
  - Support Vector Machine
  - k-Nearest Neighbors
  - Conditional Random Field
  - ...

NLP Applications - Text classification

NAIVE BAYES

Naïve Bayes

- Introduction:
  - Simple classification method based on Bayes's Theorem.
  - Use the simple text representation method: Bag of words - BOW

Naïve Bayes (cont.)

- BOW

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

I **love** this movie! It's **sweet**, but with **satirical** humor. The dialogue is **great** and the adventure scenes are **fun…** It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it to just about anyone. I've seen it **several** times, and I'm always **happy** to see it **again** whenever I have a friend who hasn't seen it yet.

Naïve Bayes (cont.)

- BOW

```
x love xxxxxxxxxxxxxxxxx sweet
xxxxxxx satirical xxxxxxxxxx
xxxxxxxxxxx great xxxxxxx
xxxxxxxxxxxxxxxxxxx fun  xxxx
xxxxxxxxxxxxx whimsical xxxx
romantic xxxx  laughing
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx recommend xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx several xxxxxxxxxxxxxxxxx
xxxxx  happy xxxxxxxxx again
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxx
```

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

Naïve Bayes (cont.)

- Naïve Bayes **classifier**
  - Given document d and category c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naïve Bayes (cont.)

- Naïve Bayes **classifier**
  - Given document d and category c

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

MAP is "maximum a posteriori" = most likely class

Bayes Rule

Dropping the denominator

Naïve Bayes (cont.)

- Naïve Bayes **classifier**
  - Given document d and category c

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c)P(c)$$

Document d represented as features x1..xn

Naïve Bayes (cont.)

- Naïve Bayes **classifier**
  - Given document d and category c

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}}\, P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

How often does this class occur?

O(|X|^n•|C|) parameters

Could only be estimated if a very, very large number of training examples was available.

We can just count the relative frequencies in a corpus

Naïve Bayes (cont.)

- Naïve Bayes **classifier**
  - Given document d and category c $P(x_1, x_2, \ldots, x_n \mid c)$
  - BOW assumption: The order of words is not important.
  - Conditional Independence: Suppose that probabilities of features $P(x_i|c)$ are conditionally independent.

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Naïve Bayes (cont.)

- Naïve Bayes **classifier**
  - Given document d and category c

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

fit@hcmus

Naïve Bayes (cont.)

- Training: Maximum likelihood Estimation (MLE)
  - Use frequency in the training data

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Naïve Bayes (cont.)

- Problem of MLE:
  - Terms do not appear in training corpus.

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

  - Then the final result equals to 0

$$c_{MAP} = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Naïve Bayes (cont.)

- Problem of MLE:
  - Solve by applying smoothing-1 method (Laplace)

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c)}{\sum_{w \in V} \left(count(w, c) \right)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

Naïve Bayes (cont.)

- Training

- From training corpus, extract *Vocabulary*

- Calculate $P(c_j)$ terms
  - For each c_j in C do

    $docs_j \leftarrow$ all docs with class $= c_j$

    $$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

    $$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Naïve Bayes (cont.)

- As a Language model

Model pos

0.1 I

0.1 love

0.01this

0.05fun

0.1 film

Model neg

0.2 I

0.001 love

0.01this

0.005 fun

0.1 film

I	love	this	fun	film
0.1	0.1	0.01	0.05	0.1
0.2	0.001	0.01	0.005	0.1

$P(s|pos) > P(s|neg)$

Naïve Bayes (cont.)

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

- Example:

$$\hat{P}(c) = \frac{N_c}{N} \qquad \hat{P}(w|c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Priors:
P(c) = ¾
P(j) = ¼

Conditional Probabilities:

P(Chinese|c) = (5+1) / (8+6) = 6/14 = 3/7
P(Tokyo|c) = (0+1) / (8+6) = 1/14
P(Japan|c) = (0+1) / (8+6) = 1/14
P(Chinese|j) = (1+1) / (3+6) = 2/9
P(Tokyo|j) = (1+1) / (3+6) = 2/9
P(Japan|j) = (1+1) / (3+6) = 2/9

Choosing a class:

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

NLP Applications - Text classification

Evaluation

fit@hcmus

Evaluation

	correct	not correct
selected	tp	fp
not selected	fn	tn

- Precision: The percentage of right items
- Recall: The percentage of chosen items
- F-Measure: The harmonic mean of recall and precision.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

=> F = 2PR/(P+R) khi β = 1