Text Classification – Naïve Bayes

CSCI 544 – Fall 2016
9/7/2016

Kallirroi Georgila

Classification

- Assign a category (class) to some data
- Binary (2 classes) classification vs. multi-class (more than 2 classes) classification
- Examples:
  - Face recognition
  - Optical character recognition
  - Handwriting recognition
  - Medical imaging
  - Speech recognition
  - Speaker recognition
  - Biometric classification

Text classification

- Assign a category (class) to a piece of text (sentence, document, etc.)
- Binary (2 classes) classification vs. multi-class (more than 2 classes) classification
- Examples:
  - Spam detection: spam, not spam
  - Sentiment analysis: positive, neutral, negative
  - Topic identification: politics, sports, entertainment, history, etc.
  - Authorship identification: Homer, Shakespeare, Austen, Dickens, James, Hemingway, McEwan, etc.
  - Language identification: English, Italian, Spanish, etc.
  - Dialogue act detection: question, answer, acknowledgement, information request, clarification request, repetition, etc.

Text classification (cont.)

- Use training data to learn a function that maps text input (a vector of features) into classes
  - The training data consists of a set of training examples
  - For each training example we have a feature vector and a target class

    $x_1, x_2, x_3, \ldots, x_n \rightarrow c$
  - Example: e-mail spam detection
    - Classes: *spam* vs. *not-spam*
    - Features: bag-of-words (all words in the e-mail message with counts but without accounting for order)
    - Training data: collection of e-mail messages marked as *spam* or *not-spam*
- Where does the training data come from?
  – Expert annotation, crowdsourcing, users' reports, etc.

Text classification (cont.)

- Training phase
  - Given a dataset (M training examples, K classes)

    $x_{11}, x_{12}, x_{13}, \ldots, x_{1n} \rightarrow c_1$

    $x_{21}, x_{22}, x_{23}, \ldots, x_{2n} \rightarrow c_2$

    $x_{31}, x_{32}, x_{33}, \ldots, x_{3n} \rightarrow c_3$

    …

    $x_{M1}, x_{M2}, x_{M3}, \ldots, x_{Mn} \rightarrow c_K$
  - The goal is to estimate $f(\mathbf{x}_m) = c_k$
- Testing phase
  - Given a set of features (not necessarily the same features we had for training) and the function $f(\mathbf{x}_m)$, find the most likely class (class with the highest probability for these features)

Naïve Bayes for text classification

- It is based on applying *Bayes'* theorem with *naïve* independence assumptions between features
- This is a very common baseline that performs surprisingly well in many tasks
- Features: bag-of-words (all words with counts but without accounting for order)
- For each class c_k compute $P(c_k|\text{bag-of-words})$ and pick the class with the highest probability

Naïve Bayes for text classification (cont.)

- Given a document d, what class does it belong to?
- Find the most likely class c_{pred}

$$c_{pred} = \arg\max_{c_k} P(c_k \mid d)$$

$$= \arg\max_{c_k} \frac{P(c_k)P(d \mid c_k)}{P(d)}$$

$$= \arg\max_{c_k} \frac{P(c_k)P(d \mid c_k)}{\sum_{k=1}^{K} P(c_k)P(d \mid c_k)}$$

Naïve Bayes for text classification (cont.)

$$c_{pred} = \arg\max_{c_k} \frac{P(c_k)P(d \mid c_k)}{\sum_{k=1}^{K} P(c_k)P(d \mid c_k)}$$

- How do we estimate $P(c_k)$?
- How do we estimate $P(d|c_k)$?
  - Naïve Bayes assumption: words are independent
  - If document d is L words long

    $P(d|c_k) = P(w_1|c_k)\, P(w_2|c_k)\, P(w_3|c_k) \ldots P(w_L|c_k)$

Note: the denominator (in the equation on the top) is the same for all classes and omitting it will not affect the comparison of classes

Spam filtering with naïve Bayes classification

- Users create labeled data for free by tagging their own e-mails, thus training data is abundant
- Each sentence in the training data below is regarded as a document

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary (12 distinct words in total in our training data)

click
for
pharmacy
free
time
today
online
link
no
good
is
money

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

$$P(c_k) = \frac{count(c_k)}{M}$$

count(c_k): number of documents of class c_k

M: total number of documents

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

P(pharmacy|spam) = 3/9 = 1/3

$$P(w_l \mid c_k) = \frac{count(w_l, c_k)}{count(w, c_k)}$$

$count(w_l, c_k)$: number of times the word w_l appears in documents of class c_k

$count(w, c_k)$: total number of words in documents of class c_k

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
$\neg$SPAM	free time today
SPAM	online pharmacy link
$\neg$SPAM	no free time
$\neg$SPAM	free good pharmacy
SPAM	pharmacy free link
$\neg$SPAM	for time today
$\neg$SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P($\neg$spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|$\neg$spam) = 1/15

$$P(w_l \mid c_k) = \frac{count(w_l, c_k)}{count(w, c_k)}$$

count(w_l, c_k): number of times the word w_l appears in documents of class c_k

count(w, c_k): total number of words in documents of class c_k

Spam filtering with naïve Bayes classification (cont.)

Msg = "pharmacy for pharmacy"
Classify Msg as spam or ¬spam

$$c_{pred} = \arg\max_{c_k} P(c_k \mid d) = \arg\max_{c_k} \frac{P(c_k)P(d \mid c_k)}{\sum_{k=1}^{K} P(c_k)P(d \mid c_k)}$$

$$P(spam \mid Msg) = \frac{P(spam)P(Msg \mid spam)}{P(spam)P(Msg \mid spam) + P(\neg spam)P(Msg \mid \neg spam)}$$

$$P(\neg spam \mid Msg) = \frac{P(\neg spam)P(Msg \mid \neg spam)}{P(spam)P(Msg \mid spam) + P(\neg spam)P(Msg \mid \neg spam)}$$

if P(spam|Msg) > P(¬spam|Msg) then Msg is classified as spam
else if P(spam|Msg) < P(¬spam|Msg) then Msg is classified as ¬ spam
else cannot decide
Note: the denominator is the same for all classes and omitting it will not affect the comparison of classes

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|¬spam) = 1/15

Msg = "pharmacy for pharmacy"

$$P(spam \mid Msg) = \frac{P(spam)P(Msg \mid spam)}{P(spam)P(Msg \mid spam) + P(\neg spam)P(Msg \mid \neg spam)}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|¬spam) = 1/15

Msg = "pharmacy for pharmacy"

$$P(spam \mid Msg) = \frac{\frac{3}{8} P(Msg \mid spam)}{\frac{3}{8} P(Msg \mid spam) + \frac{5}{8} P(Msg \mid \neg spam)}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|¬spam) = 1/15

P(for|spam) = 1/9

P(for|¬spam) = 1/15

Msg = "pharmacy for pharmacy"

$$P(spam \mid Msg) = \frac{\frac{3}{8} P(Msg \mid spam)}{\frac{3}{8} P(Msg \mid spam) + \frac{5}{8} P(Msg \mid \neg spam)}$$

$$P(Msg \mid spam) = P(pharmacy \mid spam) P(for \mid spam) P(pharmacy \mid spam)$$

$$= \frac{1}{3} \times \frac{1}{9} \times \frac{1}{3} = \frac{1}{81}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|¬spam) = 1/15

P(for|spam) = 1/9

P(for|¬spam) = 1/15

Msg = "pharmacy for pharmacy"

$$P(spam \mid Msg) = \frac{\frac{3}{8} \times \frac{1}{81}}{\frac{3}{8} \times \frac{1}{81} + \frac{5}{8} P(Msg \mid \neg spam)}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
$\neg$SPAM	free time today
SPAM	online pharmacy link
$\neg$SPAM	no free time
$\neg$SPAM	free good pharmacy
SPAM	pharmacy free link
$\neg$SPAM	for time today
$\neg$SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P($\neg$spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|$\neg$spam) = 1/15

P(for|spam) = 1/9

P(for|$\neg$spam) = 1/15

Msg = "pharmacy for pharmacy"

$$P(spam \mid Msg) = \frac{\frac{1}{216}}{\frac{1}{216} + \frac{5}{8} P(Msg \mid \neg spam)}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM click for pharmacy
¬SPAM free time today
SPAM online pharmacy link
¬SPAM no free time
¬SPAM free good pharmacy
SPAM pharmacy free link
¬SPAM for time today
¬SPAM time is money

Vocabulary size: 12
P(spam) = 3/8
P(¬spam) = 5/8
P(pharmacy|spam) = 1/3
P(pharmacy|¬spam) = 1/15
P(for|spam) = 1/9
P(for|¬spam) = 1/15

Msg = “pharmacy for pharmacy”

$$P(spam \mid Msg) = \frac{\frac{1}{216}}{\frac{1}{216} + \frac{5}{8} P(Msg \mid \neg spam)}$$

$$P(Msg \mid \neg spam) = P(pharmacy \mid \neg spam) P(for \mid \neg spam) P(pharmacy \mid \neg spam)$$
$$= \frac{1}{15} \times \frac{1}{15} \times \frac{1}{15} = \frac{1}{3375}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|¬spam) = 1/15

P(for|spam) = 1/9

P(for|¬spam) = 1/15

Msg = "pharmacy for pharmacy"

$$P(spam \mid Msg) = \frac{\frac{1}{216}}{\frac{1}{216} + \frac{5}{8} \times \frac{1}{3375}}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM click for pharmacy
¬SPAM free time today
SPAM online pharmacy link
¬SPAM no free time
¬SPAM free good pharmacy
SPAM pharmacy free link
¬SPAM for time today
¬SPAM time is money

Vocabulary size: 12
P(spam) = 3/8
P(¬spam) = 5/8
P(pharmacy|spam) = 1/3
P(pharmacy|¬spam) = 1/15
P(for|spam) = 1/9
P(for|¬spam) = 1/15

Msg = "pharmacy for pharmacy"

$$P(spam|Msg) = \frac{\frac{1}{216}}{\frac{1}{216} + \frac{1}{5400}}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

$P(\text{spam}) = 3/8$

$P(\neg\text{spam}) = 5/8$

$P(\text{pharmacy}|\text{spam}) = 1/3$

$P(\text{pharmacy}|\neg\text{spam}) = 1/15$

$P(\text{for}|\text{spam}) = 1/9$

$P(\text{for}|\neg\text{spam}) = 1/15$

Msg = "pharmacy for pharmacy"

$P(\text{spam}|\text{Msg}) = 25/26$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

$P(\text{spam}) = 3/8$

$P(\neg\text{spam}) = 5/8$

$P(\text{pharmacy}|\text{spam}) = 1/3$

$P(\text{pharmacy}|\neg\text{spam}) = 1/15$

$P(\text{for}|\text{spam}) = 1/9$

$P(\text{for}|\neg\text{spam}) = 1/15$

Msg = “pharmacy for pharmacy”

$P(\text{spam}|\text{Msg}) = 25/26$

What happens if Msg = “time for pharmacy”?

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|¬spam) = 1/15

P(for|spam) = 1/9

P(for|¬spam) = 1/15

P(time|spam) = 0

P(time|¬spam) = 4/15

Msg = "time for pharmacy"

$$P(spam \mid Msg) = \frac{\frac{3}{8} P(Msg \mid spam)}{\frac{3}{8} P(Msg \mid spam) + \frac{5}{8} P(Msg \mid \neg spam)}$$

$$P(Msg \mid spam) = P(time \mid spam) P(for \mid spam) P(pharmacy \mid spam)$$

$$= 0 \times \frac{1}{9} \times \frac{1}{3} = 0$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels	
SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

P(spam) = 3/8

P(¬spam) = 5/8

P(pharmacy|spam) = 1/3

P(pharmacy|¬spam) = 1/15

P(for|spam) = 1/9

P(for|¬spam) = 1/15

P(time|spam) = 0

P(time|¬spam) = 4/15

Msg = "time for pharmacy"

$$P(spam \mid Msg) = \frac{\frac{3}{8} \times 0}{\frac{3}{8} \times 0 + \frac{5}{8} P(Msg \mid \neg spam)}$$

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

$P(\text{spam}) = 3/8$

$P(\neg\text{spam}) = 5/8$

$P(\text{pharmacy}|\text{spam}) = 1/3$

$P(\text{pharmacy}|\neg\text{spam}) = 1/15$

$P(\text{for}|\text{spam}) = 1/9$

$P(\text{for}|\neg\text{spam}) = 1/15$

$P(\text{time}|\text{spam}) = 0$

$P(\text{time}|\neg\text{spam}) = 4/15$

Msg = “time for pharmacy”

$P(\text{spam}|\text{Msg}) = 0$

Is this classification good?

Spam filtering with naïve Bayes classification (cont.)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12

$P(spam) = 3/8$

$P(\neg spam) = 5/8$

$P(pharmacy|spam) = 1/3$

$P(pharmacy|\neg spam) = 1/15$

$P(for|spam) = 1/9$

$P(for|\neg spam) = 1/15$

$P(time|spam) = 0$

$P(time|\neg spam) = 4/15$

Msg = “time for pharmacy”

$P(spam|Msg) = 0$

We need “smoothing”, e.g., add-one smoothing

Add-one smoothing

Computing $P(c_k)$, e.g., P(spam) or P(¬spam)

Same formula with or without smoothing (we assume that we have enough documents in our training data for each class so no smoothing is required)

$$P(c_k) = \frac{count(c_k)}{M}$$

$count(c_k)$: number of documents of class c_k
M: total number of documents

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12
P(spam) = 3/8
P(¬spam) = 5/8

Add-one smoothing (cont.)

Computing $P(w_l|c_k)$, e.g., P(pharmacy|spam) or P(pharmacy|¬spam)

Without smoothing

$$P(w_l \mid c_k) = \frac{count(w_l, c_k)}{count(w, c_k)}$$

With smoothing

$$P(w_l \mid c_k) = \frac{count(w_l, c_k) + 1}{count(w, c_k) + V}$$

count(w_l, c_k): number of times the word w_l appears in documents of class c_k
count(w, c_k): total number of words in documents of class c_k
V: vocabulary size (number of distinct words in our training data)

Documents with labels

SPAM	click for pharmacy
¬SPAM	free time today
SPAM	online pharmacy link
¬SPAM	no free time
¬SPAM	free good pharmacy
SPAM	pharmacy free link
¬SPAM	for time today
¬SPAM	time is money

Vocabulary size: 12
P(spam) = 3/8
P(¬spam) = 5/8
Without smoothing:
P(pharmacy|spam) = 3/9 = 1/3
P(time|spam) = 0
With smoothing:
P(pharmacy|spam)=(3+1)/(9+12)=4/21
P(time|spam)=(0+1)/(9+12)=1/21

Add-one smoothing (cont.)

Documents with labels

SPAM click for pharmacy
¬SPAM free time today
SPAM online pharmacy link
¬SPAM no free time
¬SPAM free good pharmacy
SPAM pharmacy free link
¬SPAM for time today
¬SPAM time is money

Vocabulary size: 12
P(spam) = 3/8
P(¬spam) = 5/8
With smoothing:
P(pharmacy|spam) = 4/21
P(pharmacy|¬spam) = 2/27
P(for|spam) = 2/21
P(for|¬spam) = 2/27
P(time|spam) = 1/21
P(time|¬spam) = 5/27

Msg = "time for pharmacy"

$$P(spam \mid Msg) = \frac{\frac{3}{8} P(Msg \mid spam)}{\frac{3}{8} P(Msg \mid spam) + \frac{5}{8} P(Msg \mid \neg spam)}$$

$$P(Msg \mid spam) = P(time \mid spam) P(for \mid spam) P(pharmacy \mid spam)$$
$$= \frac{1}{21} \times \frac{2}{21} \times \frac{4}{21} = \frac{8}{9261}$$

Evaluation

Accuracy

- Out of all predictions, what fraction was correct?

$$accuracy = \frac{count(correctly\_classified\_documents)}{count(documents)}$$

Evaluation (cont.)

- Precision of class c_k
  - Out of the documents *predicted* to be of class c_k, what fraction was *actually* of class c_k?

$$precision(c_k) = \frac{count(correctly\_classified\_as\_c_k)}{count(classified\_as\_c_k)}$$

- Recall of class c_k
  - Out of all the documents that *actually* belong in class c_k, what fraction did we find?

$$recall(c_k) = \frac{count(correctly\_classified\_as\_c_k)}{count(belongs\_in\_c_k)}$$

Evaluation (cont.)

- F-score: combining precision and recall

$$F_1(c_k) = \frac{2 \times precision(c_k) \times recall(c_k)}{precision(c_k) + recall(c_k)}$$

Evaluation (cont.)

Actual	Predicted
SPAM	SPAM
¬SPAM	¬SPAM
¬SPAM	SPAM
¬SPAM	¬SPAM
SPAM	¬SPAM
¬SPAM	SPAM

Accuracy = 3/6 = 1/2

Evaluation (cont.)

Actual	Predicted
SPAM	SPAM
¬SPAM	¬SPAM
¬SPAM	SPAM
¬SPAM	¬SPAM
SPAM	¬SPAM
¬SPAM	SPAM

Accuracy = 1/2

Precision(spam) = 1/3

Evaluation (cont.)

Actual	Predicted
SPAM	SPAM
¬SPAM	¬SPAM
¬SPAM	SPAM
¬SPAM	¬SPAM
SPAM	¬SPAM
¬SPAM	SPAM

Accuracy = 1/2

Precision(spam) = 1/3

Recall(spam) = 1/2

Evaluation (cont.)

Actual	Predicted
SPAM	SPAM
$\neg$SPAM	$\neg$SPAM
$\neg$SPAM	SPAM
$\neg$SPAM	$\neg$SPAM
SPAM	$\neg$SPAM
$\neg$SPAM	SPAM

Accuracy = 1/2

Precision(spam) = 1/3

Recall(spam) = 1/2

F_1(spam) = 2x0.33x0.5/(0.33+0.5)
= 0.33/0.83 = 0.4

What happens if the vocabulary is very large?

- Some probabilities become very low resulting in underflow
  - Especially P(unknown_word|c_k)
- To avoid this problem we can use logarithms
- Below we omit the denominators from the equations because the denominator is the same for all classes

$$c_{pred} = \arg\max_{c_k} P(c_k) P(d \mid c_k) = \arg\max_{c_k} P(c_k) \prod_{l=1}^{L} P(w_l \mid c_k)$$

$$c_{pred} = \arg\max_{c_k} [\log P(c_k) + \log P(d \mid c_k)] = \arg\max_{c_k} [\log P(c_k) + \sum_{l=1}^{L} \log P(w_l \mid c_k)]$$

Beyond bag-of-words

- Features are not limited to words
- In naïve Bayes we have the independence assumption between features
- Other classification methods, e.g., support vector machines, maximum entropy models, etc., do not force us to assume that features are independent and often result in better accuracies

Reference

- C. D. Manning, P. Raghavan, and H. Schütze. Introduction to Information Retrieval, Cambridge University Press, 2008

  Chapter on Text Classification and Naïve Bayes

  http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf