Naïve Bayes for Text Classification

Text Classification

- Task of assigning a given document to one of a fixed set of classes, on the basis of the text it contains.

- Naïve Bayes models are often used for this task.
  - Query variable is the document category, and the evidence variables are the presence or absence of each word in the language.

- How such a model can be constructed, given as **training data** a set of documents that **have been** assigned to categories?

Bernoulli Model

For each class c, $P(c)$ is estimated as the fraction of all the "training" documents that are of class c.

$$P(c) = \frac{N_c}{N}$$

$P(t|c)$ is estimated as the fraction of documents of class c that contain term t.

$$P(t|c) = \frac{N_{c,t}}{N_c}$$

Bernoulli Model (cont’d)

Now we can use Naïve Bayes for classifying a new document *d*:

Estimate:

$$P(c \mid d) = \alpha * P(c) * \Pi_{t \in d} P(t \mid c) * \Pi_{t \notin d} (1 - P(t \mid c))$$

Produce as classification result:

$$c_{map} = \text{argmax}_c \; P(c|d)$$

map: maximum a posteriori

Bernoulli Model (cont'd)

To avoid number overflow, we operate on the logs of probabilities:

$$\log P(c|d) \; = \log \alpha + \log P(c) + \sum\nolimits_{t \in d} \log P(t|c) + \sum\nolimits_{t \notin d} \log(1 - P(t|c))$$

To avoid the zero frequency problem we do:

$$P(t|c) = \frac{N_{c,t} + 1}{N_c + 2}$$

Bernoulli Model (cont'd)

```
TRAINBERNOULLINB(ℂ, 𝔻)
  V ← EXTRACTVOCABULARY(𝔻)
  N ← COUNTDOCS(𝔻)
  for each c ∈ ℂ
  do N_c ← COUNTDOCSINCLASS(𝔻, c)
      prior[c] ← N_c / N
      for each t ∈ V
      do N_ct ← COUNTDOCSINCLASSCONTAININGTERM(𝔻, c, t)
          condprob[t][c] ← (N_ct + 1)/(N_c + 2)
  return V, prior, condprob
```

```
APPLYBERNOULLINB(ℂ, V, prior, condprob, d)
  V_d ← EXTRACTTERMSFROMDOC(V, d)
  for each c ∈ ℂ
  do score[c] ← log prior[c]
      for each t ∈ V
      do if t ∈ V_d
            then score[c] += log condprob[t][c]
            else  score[c] += log(1 − condprob[t][c])
  return arg max_{c∈ℂ} score[c]
```

Bernoulli Model Example

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Priors: P(c) = 3/4 and P(c) = 1/4

Conditional probabilities:
P(Chinese|c) = (3 + 1)/(3 + 2) = 4/5
P(Japan|c) = P(Tokyo|c) = (0 + 1)/(3 + 2) = 1/5
P(Beijing|c) = P(Macao|c) = P(Shanghai|c) = (1 + 1)/(3 + 2) = 2/5

P(Chinese|-c) = (1 + 1)/(1 + 2) = 2/3
P(Japan|c-) = P(Tokyo|-c) = (1 + 1)/(1 + 2) = 2/3
P(Beijing|-c) = P(Macao|-c) = P(Shanghai|-c) = (0 + 1)/(1 + 2) = 1/3

Denominators are (3 + 2) and (1 + 2) because there are three documents in c and one document in c and because the constant we add is 2 – there are two cases to consider for each term, occurrence and nonoccurrence.

Then, we get:
$P(c|d_5) \propto P(c) \cdot P(\text{Chinese}|c) \cdot P(\text{Japan}|c) \cdot P(\text{Tokyo}|c) \cdot (1 - P(\text{Beijing}|c)) \cdot (1 - P(\text{Shanghai}|c)) \cdot (1 - P(\text{Macao}|c)) =$
$3/4 \cdot 4/5 \cdot 1/5 \cdot 1/5 \cdot (1-2/5) \cdot (1-2/5) \cdot (1-2/5) \approx 0.005$
$P(c|d_5) \propto 1/4 \cdot 2/3 \cdot 2/3 \cdot 2/3 \cdot (1-1/3) \cdot (1-1/3) \cdot (1-1/3) \approx 0.022$
Classifier assigns the test document to c = not-China.
When looking only at binary occurrence and not at term frequency, Japan and Tokyo are indicators for c (2/3 > 1/5) and the conditional probabilities of Chinese for c and -c are not different enough (4/5 vs. 2/3) to affect the classification decision.

Bernoulli Model Problem

- When classifying a test document, the Bernoulli model uses binary occurrence information, ignoring the number of occurrences.

- As a result, the Bernoulli model typically makes many mistakes when classifying long documents.

For example, a document could have 1000 occurrences of China, and only two occurrences of Tokyo and Japan, and the classifier assigns the document to class "not-China".

Multinomial Model

For each class c, $P(c)$ is estimated as the fraction of all the "training" documents that are of class c.

$$P(c) = \frac{N_c}{N}$$

$P(t|c)$ is estimated as as the relative frequency of term t in documents belonging to class c.

$$P(t|c) = \frac{T_{c,t}}{\sum_{t \in V} T_{c,t}}$$

V is the vocabulary

$T_{c,t}$ is the number of occurrences of t in training documents from class c, including multiple occurrences of a term in a document.

Multinomial Model (cont'd)

Now we can use Naïve Bayes for classifying a new document d:

Estimate:

$$P(c \mid d) = \alpha * P(c) * \Pi_{t \in d} P(t \mid c)$$

Produce as classification result:

$$c_{map} = \text{argmax}_c \; P(c|d)$$

map: maximum a posteriori

Multinomial Model (cont'd)

To avoid number overflow, we operate on the logs of probabilities:

$$\log P(c|d) = \log \alpha + \log P(c) + \sum_{t \in d} \log P(t|c)$$

To avoid the zero frequency problem we do:

$$P(t|c) = \frac{T_{c,t} + 1}{\sum_{t \in V}(T_{c,t} + 1)} = \frac{T_{c,t} + 1}{\left(\sum_{t \in V} T_{c,t}\right) + |V|}$$

Multinomial Model (cont'd)

```
TrainMultinomialNB(C, D)
V ← ExtractVocabulary(D)
N ← CountDocs(D)
for each c ∈ C
do N_c ← CountDocsInClass(D, c)
    prior[c] ← N_c / N
    text_c ← ConcatenateTextOfAllDocsInClass(D, c)
    for each t ∈ V
    do T_ct ← CountTokensOfTerm(text_c, t)
    for each t ∈ V
    do condprob[t][c] ← (T_ct + 1) / Σ_t' (T_ct' + 1)
return V, prior, condprob
```

```
ApplyMultinomialNB(C, V, prior, condprob, d)
W ← ExtractTokensFromDoc(V, d)
for each c ∈ C
do score[c] ← log prior[c]
    for each t ∈ W
    do score[c] += log condprob[t][c]
return arg max_{c∈C} score[c]
```

Multinomial Model Example

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Priors: P(c) = 3/4 and P(c) = 1/4

Conditional probabilities:
P(Chinese|c) = (5 + 1)/(8 + 6) = 6/14 = 3/7
P(Tokyo|c) = P(Japan|c) = (0 + 1)/(8 + 6) = 1/14

P(Chinese|-c) = (1 + 1)/(3 + 6) = 2/9
P(Tokyo|-c) = P(Japan|-c) = (1 + 1)/(3 + 6) = 2/9

Denominators are (8 + 6) and (3 + 6)
because the lengths of textc and textc are 8 and 3, respectively, and
because the constant we add is 6 (vocabulary consists of six terms).

We then get:
$P(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003.$
$P(c|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001.$
Thus, the classifier assigns the test document to c = China.
The three occurrences of the positive indicator Chinese in d_5 outweigh
the occurrences of the two negative indicators Japan and Tokyo.

FEATURE SELECTION

What is it?

- Feature selection is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification.
- **Two main purposes**:
  - First, it makes training and applying a classifier more efficient by decreasing the size of the effective vocabulary.
  - Second, feature selection often increases classification accuracy by eliminating **noise features**.

A noise feature is one that, when added to the document representation, increases the classification error on new data.

> **E.g.** Suppose a rare term, say arachnocentric, has no information about a class, say China, but all instances (say two) of arachnocentric happen to occur in China documents in our training set. Then the learning method might produce a classifier that misassigns test documents containing arachnocentric to China. Such an incorrect generalization from an accidental property of the training set is called OVERFITTING.

Mutual Information (MI)

- For a given class c, we compute a utility measure A(t, c) for each term of the vocabulary and select the *k* terms that have the highest values of A(t, c)
  - All other terms are discarded and not used in classification
- A common method is to compute A(t, c) as the expected mutual information (MI) of term t and class c.
  - MI measures how much information the presence/absence of a term contributes to making the correct classification decision on c. Formally:

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$

$$\begin{aligned} I(U;C) = & \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_{0.}N_{.1}} \\ & + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0.}N_{.0}} \end{aligned}$$

Observe: $\frac{NN_{11}}{N_{1.}N_{.1}} = \frac{\frac{N_{11}}{N}}{\frac{N_{1.}}{N}.\frac{N_{.1}}{N}}$

Ns are counts of documents that have the values of e_t and e_c that are indicated by the two subscripts. E.g.,
N_{10} is the number of docs that contain t ($e_t = 1$) and are not in c ($e_c = 0$).
$N1. = N_{10} + N_{11}$ is the number of documents that contain t ($e_t = 1$) and we count docs independent of class ($e_c \in \{0, 1\}$).
$N = N_{00} + N_{01} + N_{10} + N_{11}$ is the total number of documents.

Interpretation

- Mutual information measures how much information – in the information theoretic sense – a term contains about the class.
- If a term's distribution is the same in the class as it is in the collection as a whole, then $I(U; C) = 0$.
- MI reaches its maximum value if the term is a perfect indicator for class membership, that is, if the term is present in a document if and only if the document is in the class.

Example: Reuters-RCV1

- Collection with roughly 1 GB of text.
- Covers a wide range of international topics.
- Consists of about 800,000 documents sent over the Reuters newswire during a 1-year period between August 20, 1996, and August 19, 1997.
- Typical document:

REUTERS

You are here: Home > News > Science > Article

Go to a Section: U.S. International Business Markets Politics Entertainment Technology Sports Oddly Enough

Extreme conditions create rare Antarctic clouds

Tue Aug 1, 2006 3:20am ET

Email This Article | Print This Article | Reprints

[-] Text [+]

SYDNEY (Reuters) - Rare, mother-of-pearl colored clouds caused by extreme weather conditions above Antarctica are a possible indication of global warming, Australian scientists said on Tuesday.

Known as nacreous clouds, the spectacular formations showing delicate wisps of colors were photographed in the sky over an Australian meteorological base at Mawson Station on July 25.

Example

Consider the class poultry and the term export in Reuters-RCV1.
The counts are as follows:

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\mathsf{export}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\mathsf{export}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

$$
\begin{aligned}
I(U;C) &= \frac{49}{801{,}948}\log_2\frac{801{,}948\cdot 49}{(49+27{,}652)(49+141)} \\
&\quad + \frac{141}{801{,}948}\log_2\frac{801{,}948\cdot 141}{(141+774{,}106)(49+141)} \\
&\quad + \frac{27{,}652}{801{,}948}\log_2\frac{801{,}948\cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)} \\
&\quad + \frac{774{,}106}{801{,}948}\log_2\frac{801{,}948\cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)} \\
&\approx 0.0001105
\end{aligned}
$$

Features with high mutual information scores for six Reuters-RCV1 classes.

UK

london	0.1925
uk	0.0755
british	0.0596
stg	0.0555
britain	0.0469
plc	0.0357
england	0.0238
pence	0.0212
pounds	0.0149
english	0.0126

China

china	0.0997
chinese	0.0523
beijing	0.0444
yuan	0.0344
shanghai	0.0292
hong	0.0198
kong	0.0195
xinhua	0.0155
province	0.0117
taiwan	0.0108

poultry

poultry	0.0013
meat	0.0008
chicken	0.0006
agriculture	0.0005
avian	0.0004
broiler	0.0003
veterinary	0.0003
birds	0.0003
inspection	0.0003
pathogenic	0.0003

coffee

coffee	0.0111
bags	0.0042
growers	0.0025
kg	0.0019
colombia	0.0018
brazil	0.0016
export	0.0014
exporters	0.0013
exports	0.0013
crop	0.0012

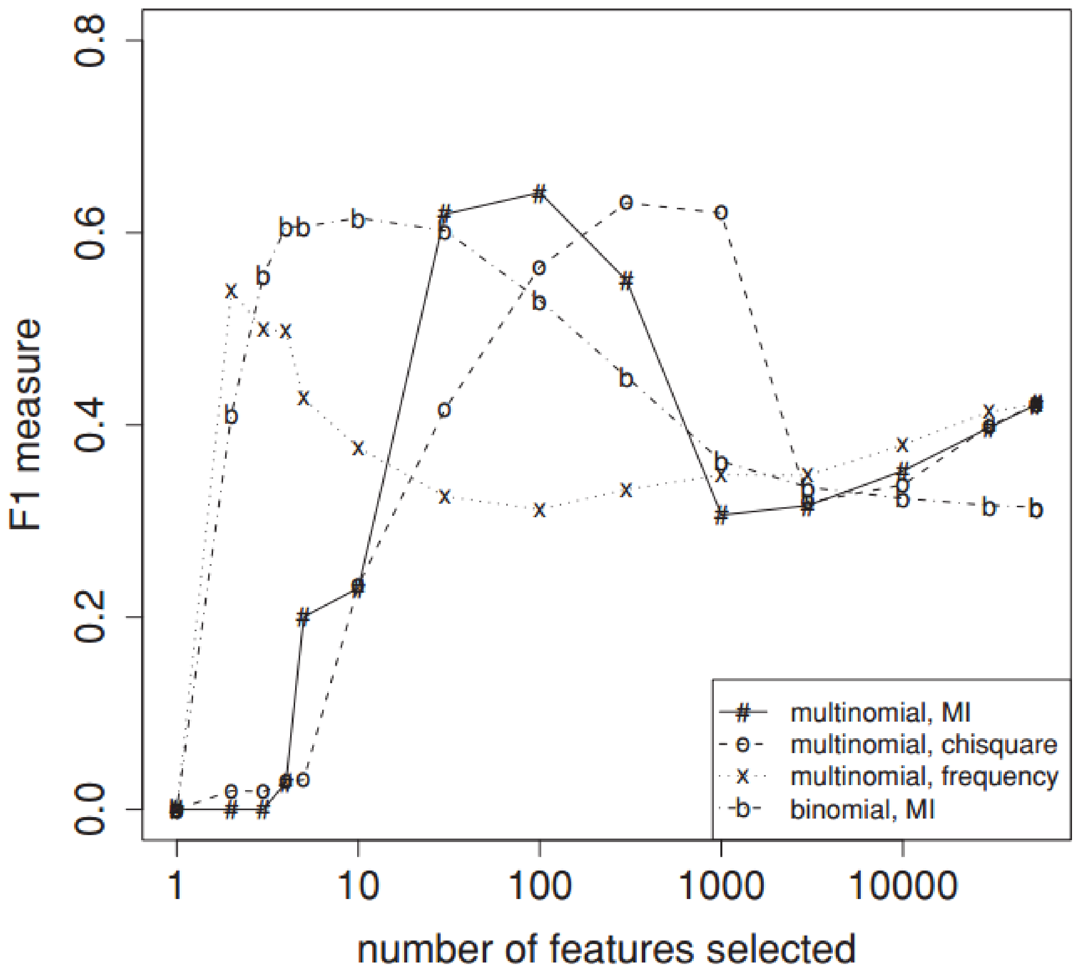
elections

election	0.0519
elections	0.0342
polls	0.0339
voters	0.0315
party	0.0303
vote	0.0299
poll	0.0225
candidate	0.0202
campaign	0.0202
democratic	0.0198

sports

soccer	0.0681
cup	0.0515
match	0.0441
matches	0.0408
played	0.0388
league	0.0386
beat	0.0301
game	0.0299
games	0.0284
team	0.0264

Effect of feature set size on accuracy for multinomial and Bernoulli models in Reuters-RCV1



For Assignment 2

- Implement the multinomial Bayes model, and do not do feature selection.