Text Classification and Naïve Bayes

The Task of Text Classification

(Reading: MR+S Ch13)
Is this spam?

Subject: Important notice!
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.


Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

© Stanford University. All Rights Reserved.
Who wrote which Federalist papers?

- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods
Male or female author?

1. 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...

2. 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...

Positive or negative movie review?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.
What is the subject of this article?

MEDLINE Article

MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...
Text Classification

• Assigning subject categories, topics, or genres
• Spam detection
• Authorship identification
• Age/gender identification
• Language Identification
• Sentiment analysis
• ...

Text Classification: definition

• **Input:**
  - a document $d$
  - a fixed set of classes $C = \{c_1, c_2, ..., c_J\}$

• **Output:** a predicted class $c \in C$
Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive
Classification Methods: Supervised Machine Learning

• **Input:**
  - a document \( d \)
  - a fixed set of classes \( C = \{c_1, c_2, \ldots, c_J\} \)
  - A training set of \( m \) hand-labeled documents \( (d_1, c_1), \ldots, (d_m, c_m) \)

• **Output:**
  - a learned classifier \( \gamma: d \rightarrow c \)
Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naïve Bayes
  - Logistic regression
  - Support-vector machines
  - k-Nearest Neighbors

- ...
Text Classification and Naïve Bayes

Naïve Bayes (I)
Naïve Bayes Intuition

- Simple ("naïve") classification method based on Bayes rule
- Relies on very simple representation of document
  - Bag of words
The bag of words representation

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.
The bag of words representation

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.
The bag of words representation: using a subset of words

\[ \gamma(\text{}) = C \]
The bag of words representation

\[
\gamma(n) = c
\]

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>great</td>
<td>2</td>
</tr>
<tr>
<td>love</td>
<td>2</td>
</tr>
<tr>
<td>recommend</td>
<td>1</td>
</tr>
<tr>
<td>laugh</td>
<td>1</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Bag of words for document classification

Test document

- parser
- language
- label
- translation

Machine Learning

- learning
- training
- algorithm
- shrinkage
- network

NLP

- parser
- tag
- training
- translation
- language

Garbage Collection

- garbage
- collection
- memory
- optimization
- region

Planning

- planning
- temporal
- reasoning
- plan
- language

GUI

- ...

...
Text Classification and Naïve Bayes

Formalizing the Naïve Bayes Classifier
Bayes’ Rule Applied to Documents and Classes

- For a document $d$ and a class $c$

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$
Naïve Bayes Classifier (I)

\[ c_{MAP} = \arg \max_{c \in C} P(c \mid d) \]

\[
= \arg \max_{c \in C} \frac{P(d \mid c)P(c)}{P(d)}
\]

\[
= \arg \max_{c \in C} P(d \mid c)P(c)
\]

MAP is “maximum a posteriori” = most likely class

Bayes Rule

Dropping the denominator
Naïve Bayes Classifier (II)

\[ c_{MAP} = \arg\max_{c \in C} P(d \mid c) P(c) \]

\[ = \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n \mid c) P(c) \]
Naïve Bayes Classifier (IV)

\[ c_{MAP} = \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n \mid c)P(c) \]

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

\[ O(|X|^n \cdot |C|) \] parameters

How often does this class occur?

We can just count the relative frequencies in a corpus
Multinomial Naïve Bayes Independence Assumptions

\[ P(x_1, x_2, \ldots, x_n | c) \]

- **Bag of Words assumption**: Assume position doesn’t matter
- **Conditional Independence**: Assume the feature probabilities \( P(x_i | c_j) \) are independent given the class \( c \).

\[ P(x_1, \ldots, x_n | c) = P(x_1 | c) \cdot P(x_2 | c) \cdot P(x_3 | c) \cdot \ldots \cdot P(x_n | c) \]
Multinomial Naïve Bayes Classifier

\[ c_{MAP} = \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n \mid c)P(c) \]

\[ c_{NB} = \arg\max_{c \in C} P(c) \prod_{x \in X} P(x \mid c) \]
Applying Multinomial Naive Bayes Classifiers to Text Classification

$$ c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i \mid c_j) $$

positions ← all word positions in test document
Text Classification and Naïve Bayes

Naïve Bayes: Learning
Learning the Multinomial Naïve Bayes Model

• First attempt: maximum likelihood estimates
  • simply use the frequencies in the data

\[
\hat{P}(c_j) = \frac{N(C = c_j)}{N}
\]

\[
\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}
\]
Parameter estimation

\[
\hat{P}(X_i = w | c_j) = \text{fraction of times word } w \text{ appears among all words in documents of topic } c_j
\]

- Create mega-document for topic \( j \) by concatenating all docs in this topic
  - Use frequency of \( w \) in mega-document
Problem with Maximum Likelihood

• What if we have seen no training documents with the word *fantastic* and classified in the topic *positive* (*thumbs-up*)?

\[
\hat{P}(x_i = "fantastic" \mid C = pos) = \frac{N(x_i = "fantastic", C = pos)}{N(C = pos)} = 0
\]

• Zero probabilities cannot be conditioned away, no matter the other evidence!

\[
c_{\text{MAP}} = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)
\]
Smoothing the Naïve Bayes estimates

- Laplace (add-1):
  \[ \hat{P}(x_i | c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k} \]

- Applied to text:
  \[ \hat{P}(w | c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|} \]
Multinomial Naïve Bayes: Learning

- 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 $Text_j$
    - $P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{N + \alpha |\text{Vocabulary}|}$
\[ \hat{P}(c) = \frac{N_c}{N} \]

\[ \hat{P}(w | c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|} \]

**Priors:**

\[ P(c) = \frac{3}{4} \]

\[ P(j) = \frac{1}{4} \]

**Conditional Probabilities:**

\[ P(\text{Chinese} | c) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7} \]

\[ P(\text{Tokyo} | c) = \frac{0+1}{8+6} = \frac{1}{14} \]

\[ P(\text{Japan} | c) = \frac{0+1}{8+6} = \frac{1}{14} \]

\[ P(\text{Chinese} | j) = \frac{1+1}{3+6} = \frac{2}{9} \]

\[ P(\text{Tokyo} | j) = \frac{1+1}{3+6} = \frac{2}{9} \]

\[ P(\text{Japan} | j) = \frac{1+1}{3+6} = \frac{2}{9} \]

**Choosing a class:**

\[ P(c \mid d_5) \propto \frac{3}{4} \times \left(\frac{3}{7}\right)^3 \times \frac{1}{14} \times \frac{1}{14} \]

\[ \approx 0.0003 \]

\[ P(j \mid d_5) \propto \frac{1}{4} \times \left(\frac{2}{9}\right)^3 \times \frac{2}{9} \times \frac{2}{9} \]

\[ \approx 0.0001 \]
Text Classification and Naïve Bayes

Text Classification: Evaluation
The 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

Accuracy = \( \frac{tp + tn}{tp + fp + fn + tn} \)
Precision and recall

- **Precision**: % of selected items that are correct
- **Recall**: % of correct items that are selected

\[
P = \frac{tp}{tp + fp}
\]
\[
R = \frac{tp}{tp + fn}
\]

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
A combined measure: F

• A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

\[
F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}
\]

• People usually use balanced F1 measure
  • i.e., with \( \beta = 1 \):

\[
F1 = \frac{2PR}{P + R}
\]

(the original slide has been slightly modified)
More Than Two Classes: Sets of binary classifiers

- Dealing with any-of or multivalue classification
  - A document can belong to 0, 1, or >1 classes.

- For each class \( c \in C \)
  - Build a classifier \( \gamma_c \) to distinguish \( c \) from all other classes \( c' \in C \)

- Given test doc \( d \),
  - Evaluate it for membership in each class using each \( \gamma_c \)
  - \( d \) belongs to any class for which \( \gamma_c \) returns true
More Than Two Classes: Sets of binary classifiers

- **One-of or multinomial classification**
  - Classes are mutually exclusive: each document in exactly one class

- For each class \( c \in C \)
  - Build a classifier \( \gamma_c \) to distinguish \( c \) from all other classes \( c' \in C \)

- Given test doc \( d \),
  - Evaluate it for membership in each class using each \( \gamma_c \)
  - \( d \) belongs to the one class with maximum score
Confusion matrix c

- For each pair of classes \(<c_1,c_2>\) how many documents from \(c_1\) were incorrectly assigned to \(c_2\)?

- \(c_{3,2}\): 90 wheat documents incorrectly assigned to poultry

<table>
<thead>
<tr>
<th>Docs in test set</th>
<th>Assigned UK</th>
<th>Assigned poultry</th>
<th>Assigned wheat</th>
<th>Assigned coffee</th>
<th>Assigned interest</th>
<th>Assigned trade</th>
</tr>
</thead>
<tbody>
<tr>
<td>True UK</td>
<td>95</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>True poultry</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True wheat</td>
<td>10</td>
<td>90</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True coffee</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>34</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>True interest</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>True trade</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>14</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>
Per class evaluation measures

Recall:
Fraction of docs in class $i$ classified correctly:

$$\frac{C_{ii}}{\sum_{j} C_{ij}}$$

Precision:
Fraction of docs assigned class $i$ that are actually about class $i$:

$$\frac{C_{ii}}{\sum_{j} C_{ji}}$$

Accuracy:
Fraction of docs classified correctly:

$$\frac{\sum_{i} C_{ii}}{\sum_{j} \sum_{i} C_{ij}}$$
Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
- **Macroaveraging**: Compute performance for each class, then average.
- **Microaveraging**: Collect decisions for all classes, compute contingency table, evaluate.
### Micro- vs. Macro-Averaging: Example

<table>
<thead>
<tr>
<th></th>
<th>Class 1</th>
<th></th>
<th>Class 2</th>
<th></th>
<th>Micro Ave. Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier: yes</td>
<td>10</td>
<td>10</td>
<td>90</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Classifier: no</td>
<td>10</td>
<td>970</td>
<td>10</td>
<td>890</td>
<td>20</td>
</tr>
</tbody>
</table>

- Macroaveraged precision: \((0.5 + 0.9)/2 = 0.7\)
- Microaveraged precision: \(100/120 = .83\)
- Microaveraged score is dominated by score on common classes
Underflow Prevention: log space

- Multiplying lots of probabilities can result in floating-point underflow.
- Since \( \log(xy) = \log(x) + \log(y) \)
  - Better to sum logs of probabilities instead of multiplying probabilities.
- Class with highest un-normalized log probability score is still most probable.

\[
c_{NB} = \arg\max_{c_j \in C} \log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i \mid c_j)
\]

- Model is now just max of sum of weights