CS 6120/CS4120: Natural Language Processing
Instructor: Prof. Lu Wang
College of Computer and Information Science
Northeastern University
Webpage: www.ccs.neu.edu/home/luwang

Announcement

• Assignment 1 is released. Due on Oct 10, 11:59pm, on Blackboard.

Project Proposal

• Length: 1 page (or more if necessary).
  • Single space if MS word is used. Or you can choose latex templates, e.g. https://www.acm.org/publications/proceedings/template or http://icml.cc/2015/?page_id=151.

• Introduction: the problem has to be well-defined. What are the input and output. Why this is an important problem to study.

• Related work: put your work in context. Describe what has been done in previous work on the same or related subject. And why what you propose to do here is novel and different.

• Datasets: what data do you want to use? What is the use of it? What information is contained? Why is it suitable for your task?

• Methodology: what models do you want to use? You may change the model as the project goes, but you may want to indicate some type of models that might be suitable for your problem. Is it a supervised learning problem or unsupervised? What classifiers can you start with? Are you making improvements? You don’t have to be crystal clear on this section, but it can be used to indicate the direction that your project goes to.

• Evaluation: what metrics do you want to use for evaluating your models?

Sample proposal and reports

• www.ccs.neu.edu/home/luwang/courses/cs6120_fa2018/cs6120_fa2018.html
• Sample projects from Stanford NLP course
  • http://web.stanford.edu/class/cs224n/reports.html
• Finding teammates on Piazza!

Outline

• Text Categorization/Classification
• Naive Bayes
• Evaluation

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.
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...

Text Classification

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

Classification Methods:

Hand-coded rules
- Rules based on combinations of words or other features
- spam: blacklist 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
- Any kind of classifier
  - Naive Bayes
  - Logistic regression
  - Support-vector machines
  - k-Nearest Neighbors
Outline

- Text Categorization/Classification
- Naïve Bayes
- Evaluation

Naïve Bayes Intuition

- Simple (“naïve”) classification method based on Bayes rule
- Relies on very simple representation of document
- Bag of words

Naïve Bayes Classifier

The Bag of Words Representation

The bag of words representation

<table>
<thead>
<tr>
<th>seen</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sweet</td>
<td>1</td>
</tr>
<tr>
<td>whimsical</td>
<td>1</td>
</tr>
<tr>
<td>recommend</td>
<td>1</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
</tbody>
</table>

... ... ...

Bayes’ Rule Applied to Documents and Classes

- For a document \( d \) and a class \( c \)

\[
P(c | d) = \frac{P(d | c)P(c)}{P(d)}
\]
Bayes’ Rule Applied to Documents and Classes

• For a document \(d\) and a class \(c\)

\[
P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}
\]

NB is a generative model! (We will talk about it later.)

**Naïve Bayes Classifier (I)**

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

MAP is “maximum a posteriori” = most likely class

**Bayes Rule**

**Dropping the denominator**

**Why we can do this?**

**Naïve Bayes Classifier (II)**

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

**Document \(d\) represented as features \(x_1, \ldots, x_n\)**

**Naïve Bayes Classifier (IV)**

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

**O(\mid X \mid n \cdot \mid C \mid) parameters**

**Bag of Words assumption:** Assume position doesn’t matter

**Conditional Independence:** Assume the feature probabilities \(P(x_i \mid c)\) are independent given the class \(c\).

\[
P(x_1, x_2, \ldots, x_n \mid c) = P(x_1 \mid c) \cdot P(x_2 \mid c) \cdot P(x_3 \mid c) \cdot \ldots \cdot P(x_n \mid c)
\]
Multinomial Naive 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

positions ← all word positions in test document

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

Learning for Naive Bayes Model

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

\[ \hat{P}(c_j) = \frac{\text{doccount}(C = c_j)}{N_{\text{doc}}} \]

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

Problem with Maximum Likelihood

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

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

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

\[ c_{MAP} = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c) \]
Calculate
From training corpus, extract Vocabulary

Each class = a unigram language model

Each class = a unigram language model

Each class = a unigram language model

Multinomial Naïve Bayes: Learning

Naïve Bayes and Language Modeling

Naïve Bayes as a Language Model
An Example

Choosing a class:

$P(c|d_5) = \frac{1}{4} \times \frac{2}{9} \times \frac{2}{9} \approx 0.0001$

Choosing a class:

$P(c|d_5) = \frac{1}{4} \times \frac{1}{14} \times \frac{1}{14} \approx 0.0003$

Summary: Naive Bayes is Not So Naive

- Very Fast, low storage requirements
- Robust to Irrelevant Features
  - Irrelevant Features cancel each other without affecting results
- Very good in domains with many equally important features
- Optimal if the independence assumptions hold:
  - If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification

Outline

- Text Categorization/Classification
- Naïve Bayes
- Evaluation

Evaluation

The 2-by-2 contingency table (or confusion matrix)

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>not correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp (true positive)</td>
<td>fp (false positive)</td>
</tr>
<tr>
<td>not selected</td>
<td>fn (false negative)</td>
<td>tn (true negative)</td>
</tr>
</tbody>
</table>

For example,

- Which set of documents are related to the topic of NLP?
- Which set of documents are written by Shakespeare?
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>

Precision and recall

- **Precision**: % of selected items that are correct, \( \frac{tp}{tp + fp} \)
- **Recall**: % of correct items that are selected, \( \frac{tp}{tp + fn} \)

A combined measure: F-measure or F-score

- A combined measure that assesses the P/R tradeoff is F measure
  (weighted harmonic mean):
  \[
  F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} \frac{(\beta^2 + 1)PR}{\beta^2 P + R}
  \]
- People usually use balanced F1 measure
  - i.e., \( \alpha = \frac{1}{2} \), \( F = 2PR/(P + R) \)

Text Classification Evaluation

- 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 $c_2$?
- Example: 90 wheat documents incorrectly assigned to poultry

<table>
<thead>
<tr>
<th>Docs in test set</th>
<th>Assigned $c_1$</th>
<th>Assigned $c_2$</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>2</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>34</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>True interest</td>
<td>0</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:
  $$\sum_{j \neq i} c_{ij}$$

Precision:
- Fraction of docs assigned class $i$ that are actually about class $i$:
  $$\sum_{j \neq i} c_{ij}$$

Accuracy:
- (1 - error rate)
  - Fraction of docs classified correctly:
    $$\sum_{i} c_{ii}$$

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

- Macroaveraged precision: $(0.5 + 0.9)/2 = 0.7$
- Microaveraged precision: $100/120 = .83$

Development Test Sets and Cross-validation

Metric: P/R/F1 or Accuracy

- Unseen test set
  - Avoid overfitting ('tuning to the test set')
  - More conservative estimate of performance
- Cross-validation over multiple splits
  - Handle sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Development Test Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set</td>
<td>Dev Test</td>
<td>Test Set</td>
</tr>
<tr>
<td>Training Set</td>
<td>Dev Test</td>
<td>Test Set</td>
</tr>
</tbody>
</table>