Project proposal (due Feb 15)

- In general, we want to see that you have a clear goal in the project. The technical details can be described in a rough manner. In principle, you need to show what problem you want to study.
- **Introduction**: the problem has to be well-defined. What are the input and output. What is the problem you are interested in? What is the motivation? Why is the problem important? What is the state-of-the-art? What information is known? Why is it suitable for your task?
- **Related work**: put your work in context. Describe what has been done in previous work on the same or related subjects. What are the limitations of what you have done? What is the idea behind your work? What makes your work novel and different?
- **Datasets**: what data do you want to use? What is the size 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?
- **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.
- **Grading**: based on each section described above, 20 points per section. But as you can tell, they’re related to each other.
- Each group member will make separate submission with all group members’ names indicated.

Outline

- Text Categorization/Classification
- Naïve 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
- ...

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 $C = \{c_1, c_2, ..., c_J\}$
  - A training set of $m$ hand-labeled documents $(d_1, y_1), ..., (d_m, y_m)$, $y_i$ is in $C$
- **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
  - Neural networks
  - ...

Outline

- Text Categorization/Classification
  - Naïve Bayes
  - Evaluation
Naive Bayes Classifier

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

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

Naive 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

- seen: 2
- sweet: 1
- whimsical: 1
- recommend: 1
- happy: 1
- ...: ...

Bayes’ Rule Applied to Documents and Classes

- Naive Bayes Classifier (I)

\[
c_{MAP} = \operatorname{argmax}_{c \in C} P(c \mid d) = \operatorname{argmax}_{c \in C} \frac{P(d \mid c)P(c)}{P(d)}
\]

Naive Bayes Classifier (I)

\[
c_{MAP} = \operatorname{argmax}_{c \in C} P(c \mid d) = \frac{\operatorname{argmax}_{c \in C} P(d \mid c)P(c)}{P(d)}
\]

\[
c_{MAP} = \operatorname{argmax}_{c \in C} P(d \mid c)P(c)
\]

Bayes Rule

MAP = “maximum a posteriori” = most likely class

Droping the denominator
Naive Bayes Classifier (I)

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

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

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

**MAP** is "maximum a posteriori" = most likely class

Bayes Rule

Dropping the denominator

Why we can do this?

19

---

Naive Bayes Classifier (II)

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

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

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

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

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

20

---

Naive Bayes Classifier (IV)

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

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

\[
\prod_{c \in \text{class}} P(x_1, x_2, \ldots, x_n | c)
\]

\[O(\times | \times | \text{class}) \text{ parameters}
\]

\([\times]\) represents the maximum number of possible values for \(x\)

21

---

Applying Naive Bayes Classifiers to Text Classification

positions ← all word positions in test document

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

22

---

Naive Bayes Classifier

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

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

23

---
Learning for Naive Bayes Model

Parameter estimation

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

fraction of times word \( w \) appears among all words in documents of class \( c_j \)

Problem with Maximum Likelihood

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

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

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

\[
c_{\text{MLE}} = \text{argmax} \ \hat{P}(c) \prod_{j} \hat{P}(x_j | c)
\]

Laplace (add-1) smoothing for Naive Bayes

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

Learning the Naive Bayes Model

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

\[
\hat{P}(c_j) = \frac{\text{doccount}(C = c_j)}{N_{\text{doc}}}
\]
\[
\hat{P}(w | c_j) = \frac{\text{count}(w, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}
\]

Naive Bayes: Learning

- From training corpus, extract Vocabulary
  - Calculate \( P(c_j) \) terms
    - For each \( c_j \) in \( C \) do
      - \( \text{doc}_j \leftarrow \text{all docs with class } = c_j \)
      - \( P(c_j) \leftarrow |\text{doc}_j| / \text{total # documents} \)
  - Calculate \( P(w_j | c_j) \) terms
    - For each word \( w_j \) in Vocabulary
      - \( \text{Text} \leftarrow \text{single doc containing all } w_j \)
      - \( n_k \leftarrow \# \text{ of occurrences of } w_j \text{ in } \text{Text} \)
      - \( \hat{P}(w_j | c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha | \text{Vocabulary}|} \)
### Naive Bayes: Learning

- From training corpus, extract Vocabulary
- Calculate $P(c)$ terms
  - For each $c_j$ in $C$ do
    - $P(c_j) = \frac{\text{docs}_j}{\text{total # documents}}$
- Calculate $P(w_k | c)$ terms
  - For each word $w_k$ in Vocabulary
    - $n_{w_k,c_j} = \#$ of occurrences of $w_k$ in Text, $n_{c_j} = \#$ of docs with class $c_j$
    - $P(w_k | c_j) = \frac{n_{w_k,c_j} + \alpha}{n_{c_j} + \alpha | \text{Vocabulary}}$

A more general form: add-$\alpha$ smoothing!

### Naive Bayes and Language Modeling

- Naive bayes classifiers can use any sort of feature
  - URL, email address, dictionaries, network features
- But if, as in the previous slides
  - We use only word features
  - We use all of the words in the text (not a subset)
- Then
  - Naive bayes has an important similarity to language modeling.

### Each class = a unigram language model

- Assigning each word: $P(\text{word} | c)$
- Assigning each sentence: $P(\text{sentence} | c) = \prod_i P(\text{word}_i | c)$

<table>
<thead>
<tr>
<th>Class</th>
<th>pos</th>
<th>Model pos</th>
<th></th>
<th>Model neg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>I</td>
<td>love</td>
<td>this</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>0.05 fun</td>
<td>P(sentence</td>
<td>pos) = 0.0000005</td>
<td></td>
<td>P(sentence</td>
</tr>
<tr>
<td>0.1 film</td>
<td>P(sentence</td>
<td>pos) = 0.0000005</td>
<td></td>
<td>P(sentence</td>
</tr>
</tbody>
</table>

### Naive Bayes as a Language Model

- Which class assigns the higher probability to $s$?

<table>
<thead>
<tr>
<th>Class</th>
<th>pos</th>
<th>Model pos</th>
<th></th>
<th>Model neg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>I</td>
<td>love</td>
<td>this</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>0.05 fun</td>
<td>P(sentence</td>
<td>pos) = 0.0000005</td>
<td></td>
<td>P(sentence</td>
</tr>
<tr>
<td>0.1 film</td>
<td>P(sentence</td>
<td>pos) = 0.0000005</td>
<td></td>
<td>P(sentence</td>
</tr>
</tbody>
</table>

### An Example

<table>
<thead>
<tr>
<th>Docs</th>
<th>1/4</th>
<th>1/8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>Beijing Chinese</td>
<td>Chinese Shanghai</td>
</tr>
<tr>
<td>1</td>
<td>Chines</td>
<td>Beijing Chinese</td>
</tr>
<tr>
<td>2</td>
<td>Chinese</td>
<td>Chinese Shanghai</td>
</tr>
<tr>
<td>3</td>
<td>Chinese</td>
<td>Macao</td>
</tr>
<tr>
<td>4</td>
<td>Chinese</td>
<td>Tokyo Japan Chinese</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>pos</th>
<th>Model pos</th>
<th></th>
<th>Model neg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>I</td>
<td>love</td>
<td>this</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>0.05 fun</td>
<td>P(sentence</td>
<td>pos) = 0.0000005</td>
<td></td>
<td>P(sentence</td>
</tr>
<tr>
<td>0.1 film</td>
<td>P(sentence</td>
<td>pos) = 0.0000005</td>
<td></td>
<td>P(sentence</td>
</tr>
</tbody>
</table>

### Choosing a class:

- $P(c | d_5) = \frac{P(d_5 | c) \cdot P(c)}{\sum_c P(d_5 | c) \cdot P(c)}$ = 0.003

### Conditional Probabilities:

<table>
<thead>
<tr>
<th>Class</th>
<th>pos</th>
<th>Model pos</th>
<th></th>
<th>Model neg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>English</td>
<td>I</td>
<td>love</td>
<td>this</td>
</tr>
<tr>
<td>0.1</td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>0.01</td>
</tr>
<tr>
<td>0.05 fun</td>
<td>P(sentence</td>
<td>pos) = 0.0000005</td>
<td></td>
<td>P(sentence</td>
</tr>
<tr>
<td>0.1 film</td>
<td>P(sentence</td>
<td>pos) = 0.0000005</td>
<td></td>
<td>P(sentence</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Training</th>
<th>Validation</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chinese Beijing Chinese</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>2</td>
<td>Chinese Chinese Shanghai</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>3</td>
<td>Chinese Macao</td>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>Tokyo Japan Chinese</td>
<td>c</td>
<td>c</td>
</tr>
</tbody>
</table>

### Priors:

- $P(c) = \frac{N_c}{N}$
- $P(w_k | c) = \frac{\text{count}(w_k, c) + \alpha}{\text{count}(c) + \alpha | \text{Vocabulary}}$

### Choosing a class:

- $P(c | d_5) = \frac{P(d_5 | c) \cdot P(c)}{\sum_c P(d_5 | c) \cdot P(c)}$ = 0.003
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</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</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, tp/(tp+fp)
• Recall: % of correct items that are selected, 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):
  \[ F = \frac{1}{\frac{1}{P} + (1-\alpha)\frac{1}{R}} \]
- People usually use balanced F1 measure
  - i.e., \( \alpha = \frac{1}{2} \), \( F = \frac{2PR}{P+R} \)

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

Confusion matrix \( c \)

- For each pair of classes \( c_1, c_2 \) how many documents from \( c_1 \) were incorrectly assigned to \( c_2 \)?
  - \( c_{12} \): 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>34</td>
<td>3</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>

Text Classification Evaluation

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

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:
  \[ \sum_{j \neq i} c_{ij} \]
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.

<table>
<thead>
<tr>
<th>Class 1</th>
<th>Class 2</th>
<th>Micro Avg. Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth: yes</td>
<td>10</td>
<td>Truth: no</td>
</tr>
<tr>
<td>Classifier: yes</td>
<td></td>
<td>Classifier: yes</td>
</tr>
<tr>
<td>Classifier: no</td>
<td>20</td>
<td>Classifier: no</td>
</tr>
<tr>
<td>Truth: yes</td>
<td>100</td>
<td>Truth: no</td>
</tr>
<tr>
<td>Classifier: yes</td>
<td>10</td>
<td>Classifier: no</td>
</tr>
</tbody>
</table>

- Macroaveraged precision: \((0.5 + 0.9)/2 = 0.7\)
- Microaveraged precision: \(100/120 = 0.83\)

Development Test Sets and Cross-validation

**Metric**: P/R/F1 or Accuracy

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/tuning/held-out Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Dev Test</td>
<td></td>
</tr>
<tr>
<td>Training Set</td>
<td>Dev Test</td>
<td></td>
</tr>
<tr>
<td>Dev Test</td>
<td>Training Set</td>
<td>Test Set</td>
</tr>
</tbody>
</table>