Natural Language Processing

Text classification

Yulia Tsvetkov

yuliats@cs.washington.edu
Some questions from the previous lecture

- NLP for sign languages
Announcements

● HW1 is out today, please start early!
Dear Colleague,

Account: yuliats@cs.washington.edu

Good news. Due to many requests, the submission deadline has been extended to 10 March 2022 (It is firm day).

We would like to invite you to submit a paper to 10. European Conference on Renewable Energy Systems (ECRES). **ECRES 2022 will be held hybrid mode, the participants can present their papers physically or online.** The event is going to be organized in Istanbul/Turkey under the technical sponsorship of Istanbul Medeniyet University and many international institutions. The conference is highly international with the participants from all continents and more than 40 countries.

The submission deadline and special and regular issue journals can be seen in [ecres.net](https://ecres.net)

There will be keynote speakers who will address specific topics of energy as you would see at [ecres.net/keynotes.html](https://ecres.net/keynotes.html)

[CLICK FOR PAPER SUBMISSION](https://ecres.net)

All accepted papers will be published in a special Conference Proceedings under a specific ISBN. Besides, the extended versions will be delivered to reputable journals indexed in SCI, E-SCI, SCOPUS, and EBSCO. You can check our previous journal publications from [ecres.net](https://ecres.net). Please note that the official journal of the event, Journal of Energy Systems (jes@iuu.edu.tr) is also indexed in SCOPUS.
Spam classification

Dear Colleague,

Account: voliana@wa.washington.edu

Good news! Due to many requests, the submission deadline has been extended to 10 March 2022 (it is firm date).

We would like to invite you to submit a paper to the European Conference on Renewable Energy Systems (ECRES). ECRES 2023, the participants can present their papers physically or virtually to be organized in Istanbul/Turkey under the technical coordination of Medeniyet University and many international institutions. The conference is international with the participants from all continents and more than 40 countries.

The submission deadline and specific and regular issue journals can be seen in ecres.net

There will be keynote speakers who will address specific topics of energy as you would see at ecres.net/keynotes.html

CLICK FOR PAPER SUBMISSION

All accepted papers will be published in a special Conference Proceedings under a specific ISBN. Besides, the extended versions will be delivered to reputable journals indexed in SCI, E-SCI, SCOPUS, and EBSCO. You can check our previous journal publications from ecres.net. Please note that the official journal of the event Journal of Energy Systems (derganark.org/jes) is also indexed in SCOPUS.

---

Invitation to present at the February 2022 Wikimedia Research Showcase

Emily Lenacik: emily@wikimedia.org

Hi Emily,

My name is Emily Lenacik and I am a member of the Research team at the Wikimedia Foundation. On Tuesday, February 22, 2022, I will present your research on social biases on Wikipedia at our Research Showcase in February 2022. This topic fits into our theme for this showcase, which is gaps and biases on Wikipedia.

The Wikimedia Research Showcase is a monthly, public lecture series where Foundation staff present research to an engaged audience. We invite team members and external researchers across wikimedia related to Wikipedia, Wikimedia, peer review, and open-source software. We focus on topics and projects that we think our audience—a global community of academic researchers, Wikipedia editors, and technology staff—would find interesting and relevant to their work.

Research Showcase presentations are generally 20 minutes long, with an additional 5 minutes for questions and discussion. Each session includes two presenters to every showcase. Most presenters choose to use slides to present their work.

The February showcase takes place on the 23rd at 8:00AM Pacific / 17:00 Pacific. Both sessions will be live streamed on the Wikipedia Foundation’s YouTube channel and also archived for later viewing on the Wikimedia Foundation’s YouTube channel.

If this date does not work for you, but you are still interested in giving a showcase, please let us know. We can discuss other options.

I hope to get a chance to see your work presented at the Research Showcase!

Sincerely,
Emily

---
Language ID

Аяны замд тур зогсон тэнгэрийн байдлыг ажиглаад хедлех зуур гутал дор шинэээн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот хүрээ тийш цас орвол орно л биз гэсэн хэнэггүй бодол маань хеде талд, говийн ээрэм хэндийд, малын бэлчээрт, малдцын хотонд болохоор солигдож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике Србије Ивица Дачић честитао је каякашкици златне медаље у олимпијској дисциплини K-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.


Nestrankarski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je vlada predsednika Donald Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo kongresno potrjene vojaške pomoći Ukrajini zaradi političnih razlogov. Predstavniški dom kongresa je prav zaradi tega sprožil ustavno obbožbo proti Trumpu.
Language ID

Аяны замд тур зогсон тэнгэрийн байдлыг ажиглаад хедех зуур гутал дор шинэхэн орсон  цас шаржигнэн дуугарч байв. Цасны тухай бодол сонин юм. Хотын түүхийг өрөлөөр байр л биз гэээн хэнэгтүй бодол маань хэдэн тахал гэндэлд, малын бэлчээрт, малчдын хотонд болохоор солиг дож эргэцүүлэн бодох нь хачин. Цас хэр орсон бол?

Београд, 16. јун 2013. године – Председник Владе Републике Србије наказало је крајашки цин вел мадаље у олимпиади Висини К-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.

Nestankarski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje, da je vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo kreditov, pridružene vojaške pomoci Ukrajini zaradi političnih razlogov. Predstavniški delavniki ZDA vendar vendar zaradi tega sprožil ustavno obtožbo proti Trumpu.
Sentiment analysis

By John Neal
This review is from: Accoutrements Horse Head Mask (Toy)
When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and gave me a list of suggested places to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.

By Christine E. Torok
Verified Purchase (What's this?)
First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside...
By John Neal
This review is from: Accoutrements Horse Head Mask (Toy)

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and sent me to the sugary places to my past. Since then, I've had a life of peace and safety knowing that my old identity is forever obscured by this life-saving item.

By Christine E. Torok
Verified Purchase (What's this?)

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweat, floating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside.
Topic classification

MeSH Subject Category Hierarchy
- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …
By 1925 Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam.

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 the greatest assets...

Fact verification: trustworthy or fake?

Have Covid19? Drink bleach!

Detecting COVID-19-Related Fake News Using Feature Extraction

Suleman Khan, Saqib Hakak, N. Deepa, B. Prabadevi, Kapal Dev and Silvia Trelova
Text classification

- We might want to categorize the content of the text:
  - Spam detection (binary classification: spam/not spam)
  - Sentiment analysis (binary or multiway)
    - movie, restaurant, product reviews (pos/neg, or 1-5 stars)
    - political argument (pro/con, or pro/con/neutral)
    - Topic classification (multiway: sport/finance/travel/etc)
  - Language Identification (multiway: languages, language families)
  - ...

- Or we might want to categorize the author of the text (authorship attribution)
  - Human- or machine generated?
  - Native language identification (e.g., to tailor language tutoring)
  - Diagnosis of disease (psychiatric or cognitive impairments)
  - Identification of gender, dialect, educational background, political orientation (e.g., in forensics [legal matters], advertising/marketing, campaigning, disinformation)
  - ...

We’ll consider alternative models for classification

- Text classification
  - Rule-based
  - Probabilistic
    - Generative models
    - Discriminative models
      - Naïve Bayes
      - Linear models
        - Multinomial logistic regression
          (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron
We’ll consider alternative models for classification

- Rule-based
- Probabilistic
  - Generative models
  - Discriminative models
    - Naïve Bayes
    - Linear models
      - Multinomial logistic regression (aka MaxEnt)
    - Non-linear models
      - Multilayer perceptron

Text classification
def classify_sentiment(document):
    for word in document:
        if word in {"good", "wonderful", "excellent"}:
            return 5
        if word in {"bad", "awful", "terrible"}:
            return 1
Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.
Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

→ hard to identify a priori which words are informative (and what information they carry!)
Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

→ hard to identify a priori which words are informative (and what information they carry!)

Sentiment: It’s not life-affirming, it’s vulgar, it’s mean, but I liked it.
Rule-based classification: challenges

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

→ hard to identify a priori which words are informative (and what information they carry!)

Sentiment: It’s not life-affirming, it’s vulgar, it’s mean, but I liked it.

→ word order matters, but hard to encode in rules!
Rule-based classification: challenges

**Sentiment:** Half submarine flick, half ghost story, all in one a criminally neglected film.

→ hard to identify a priori which words are informative (and what information they carry!)

**Sentiment:** It’s not life-affirming, it’s vulgar, it’s mean, but I liked it.

→ word order matters, but hard to encode in rules!

**Language ID:** All falter, stricken in kind. “LINGERIE SALE”

→ simple features can be misleading!
Rule-based classification

But don’t forget: if you don’t have access to data, speaker intuition and a bit of coding get you pretty far!
We’ll consider alternative models for classification

- Text classification
  - Rule-based
  - Probabilistic
    - Generative models
    - Discriminative models
      - Naïve Bayes
      - Linear models
        - Multinomial logistic regression
          (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron
Learning-based classification

Goal: pick the function $f$ that does “best” on training data
Classification: learning from data

- Supervised
  - labeled examples
    - Binary (true, false)
    - Multi-class classification (politics, sports, gossip)
    - Multi-label classification (#party #FRIDAY #fail)

- Unsupervised
  - no labeled examples

- Semi-supervised
  - labeled examples + non-labeled examples

- Weakly supervised
  - heuristically-labeled examples
Where do datasets come from?

<table>
<thead>
<tr>
<th>Human institutions</th>
<th>Noisy labels</th>
<th>Expert annotation</th>
<th>Crowd workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Government proceedings</td>
<td>Domain names</td>
<td>Treebanks</td>
<td>Question answering</td>
</tr>
<tr>
<td>Product reviews</td>
<td>Link text</td>
<td>Biomedical corpora</td>
<td>Image captions</td>
</tr>
</tbody>
</table>
Supervised classification
Training, validation, and test sets

- Training set
- Validation set
- Test set

- Labeled data
- Unlabeled data

Inference
Classification: features (measurements)

- Perform measurements and obtain features

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Apple</th>
<th>Orange</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diameter</td>
<td>4.2</td>
<td>5.2</td>
</tr>
<tr>
<td>Weight</td>
<td>212</td>
<td>315</td>
</tr>
<tr>
<td>Softness</td>
<td>3.4</td>
<td>5.7</td>
</tr>
<tr>
<td>Color</td>
<td>1332</td>
<td>4567</td>
</tr>
</tbody>
</table>

- diameter, weight, softness, color
Supervised classification: formal setting

- Learn a **classification model** from labeled data on
  - properties ("**features**") and their importance ("**weights**")
- \( X \): set of attributes or features \( \{x_1, x_2, \ldots, x_n\} \)
  - e.g. fruit measurements, or word counts extracted from an input documents
- \( y \): a "class” label from the label set \( Y = \{y_1,y_2,\ldots,y_k\} \)
  - e.g., fruit type, or spam/not spam, positive/negative/neutral
Supervised classification: formal setting

- Learn a **classification model** from labeled data on
  - properties ("features") and their importance ("weights")
- **X**: set of attributes or features \( \{x_1, x_2, \ldots, x_n\} \)
  - e.g. fruit measurements, or word counts extracted from an input document
- **y**: a "class" label from the label set \( Y = \{y_1, y_2, \ldots, y_k\} \)
  - e.g., fruit type, or spam/not spam, positive/negative/neutral

Given data samples \( \{x_1, x_2, \ldots, x_n\} \) and corresponding labels \( Y = \{y_1, y_2, \ldots, y_k\} \)

We **train** a function \( f: x \in X \rightarrow y \in Y \) (the model)
Supervised classification: formal setting

- Learn a **classification model** from labeled data on
  - properties ("features") and their importance ("weights")

- **X**: set of attributes or features \{x_1, x_2, \ldots, x_n\}
  - e.g. fruit measurements, or word counts extracted from an input document

- **y**: a “class” label from the label set \(Y = \{y_1, y_2, \ldots, y_k\}\)
  - e.g., fruit type, or spam/not spam, positive/negative/neutral

- At inference time, apply the model on new instances to predict the label
Text classification – feature extraction

What can we measure over text? Consider this movie review:

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 just to 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 before.
Text classification – feature extraction

What can we measure over text? Consider this movie review:

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 just to 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 before.
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 just to 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 before.

(almost) the entire lexicon

<table>
<thead>
<tr>
<th>word</th>
<th>count</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>love</td>
<td>10</td>
<td>0.0007</td>
</tr>
<tr>
<td>great</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>recommend</td>
<td></td>
<td></td>
</tr>
<tr>
<td>laugh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>happy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>several</td>
<td></td>
<td></td>
</tr>
<tr>
<td>boring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Types of textual features

- **Words**
  - content words, stop-words
  - punctuation? tokenization? lemmatization? lowercase?
- **Word sequences**
  - bigrams, trigrams, n-grams
- **Grammatical structure, sentence parse tree**
- **Words’ part-of-speech**
- **Word vectors**
- …
We’ll consider alternative models for classification

- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
    - Discriminative models
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron

- Naïve Bayes
Generative and discriminative models

- **Generative model**: a model that calculates the probability of the input data itself

  \[ P(X, Y) \]

  joint

- **Discriminative model**: a model that calculates the probability of a latent trait given the data

  \[ P(Y | X) \]

  conditional
Generative and discriminative models

- Generative text classification: Learn a model of the joint $P(X, y)$, and find

$$\hat{y} = \arg \max_{\tilde{y}} P(X, \tilde{y})$$

- Discriminative text classification: Learn a model of the conditional $P(y | X)$, and find

$$\hat{y} = \arg \max_{\tilde{y}} P(\tilde{y} | X)$$
We’ll consider alternative models for classification

- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
      - Naïve Bayes
    - Discriminative models
      - Linear models
        - Multinomial logistic regression
          (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron
Generative text classification: naïve Bayes

- Simple (naïve) classification method
  - based on the Bayes rule
- Relies on very simple representation of a documents
  - bag-of-words, no relative order
- A good baseline for more sophisticated models

Naïve Bayes

Sentiment analysis: movie reviews

- Given a document $d$ (e.g., a movie review)
- Decide which class $c$ it belongs to: positive, negative, neutral
- Compute $P(c \mid d)$ for each $c$
  - $P(\text{positive} \mid d)$, $P(\text{negative} \mid d)$, $P(\text{neutral} \mid d)$
  - select the one with max $P$
Bag-of-Words (BOW)

- Given a document $d$ (e.g., a movie review) – how to represent $d$?
Possible representations for text

- **Bag-of-Words (BOW)**
  - Easy, no effort required
  - Variable size, ignores sentential structure

- **Hand-crafted features**
  - Full control, can use NLP pipeline, class-specific features
  - Over-specific, incomplete, makes use of NLP pipeline

- **Learned feature representations**
  - Can learn to contain all relevant information
  - Needs to be learned
Naïve Bayes

- Given a document \( d \) and a class \( c \), Bayes’ rule:

\[
P(c|d) = \frac{P(d|c)P(c)}{P(d)}
\]
Naïve Bayes

- Given a document \( d \) and a class \( c \), Bayes’ rule:

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

\[
P(\text{‘positive’}|d) \propto P(d|\text{‘positive’})P(\text{‘positive’})
\]

\[
\text{likelihood} \quad \text{prior}
\]
Naïve Bayes

- Given a document $d$ and a class $c$, Bayes’ rule:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

$$P(\text{‘positive’}|d) \propto P(d|\text{‘positive’})P(\text{‘positive’})$$
Naïve Bayes

- Given a document $d$ and a class $c$, Bayes’ rule:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

$$P(\text{‘positive’}|d) \propto P(d|\text{‘positive’})P(\text{‘positive’})$$

likelihood
Naïve Bayes independence assumptions

\[ P(w_1, w_2, \ldots, w_n | c) \]

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

\[
P(w_1, w_2, \ldots, w_n | c) = P(w_1 | c) \times P(w_2 | c) \times P(w_3 | c) \times \ldots \times P(w_n | c)
\]
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!
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!

$$P(d|c) = P(w_1, w_2, \ldots, w_n|c) = \prod_i P(w_i|c)$$

**bag of words (BOW)**
Generative text classification: Naïve Bayes

\[ C_{NB} = \arg\max_c P(c|d) = \arg\max_c \frac{P(d|c)P(c)}{P(d)} \propto \]

\[ \arg\max_c P(d|c)P(c) = \]

\[ \arg\max_c P(w_1, w_2, \ldots, w_n|c)P(c) = \]

\[ \arg\max_{c_j} P(c_j) \prod_i P(w_i|c) \]

Bayes rule
same denominator
representation
conditional independence
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

$$
C_{NB} = \arg\max_{c_j} P(c_j) \prod_i P(w_i|c)
$$

$$
C_{NB} = \arg\max_{c_j} \log(P(c_j)) + \sum_i \log(P(w_i|c))
$$

- Model is now just max of sum of weights
Learning the multinomial naïve Bayes

- How do we learn (train) the NB model?
Learning the multinomial naïve Bayes

- How do we learn (train) the NB model?
- We learn $P(c)$ and $P(w_i|c)$ from training (labeled) data

$$C_{NB} = \arg\max_{c_j} \log(P(c_j)) + \sum_i \log(P(w_i|c))$$
Parameter estimation

- Parameter estimation during training
- Concatenate all documents with category $c$ into one mega-document
- Use the frequency of $w_i$ in the mega-document to estimate the word probability

$$C_{NB} = \arg \max_{c_j} \log(P(c_j)) + \sum_i \log(P(w_i|c))$$

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

$$\hat{P}(w_i|c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$
Parameter estimation

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

- fraction of times word \( w_i \) 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**?
Problem with Maximum Likelihood

- What if we have seen no training documents with the word “fantastic” and classified in the topic positive?

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

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

  \[
  \arg\max_{c_j} P(c_j) \prod_i P(w_i|c)
  \]
Laplace (add-1) smoothing for naïve Bayes

\[
\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j) + 1}{\sum_{w \in V}(\text{count}(w, c_j) + 1)}
\]
Laplace (add-1) smoothing for naïve Bayes

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

\[
= \frac{\text{count}(w_i, c_j) + 1}{(\sum_{w \in V} (\text{count}(w, c_j)) + |V|}
\]
Multinomial naïve Bayes : learning

- From training corpus, extract *Vocabulary*
- Calculate $P(c_j)$ terms
  - For each $c_j$ do
    - $\text{docs}_j \leftarrow$ all docs with class $= c_j$
    - $P(c_j) \leftarrow \frac{|\text{docs}_j|}{\text{total # documents}}$
Multinomial naïve Bayes: learning

- From training corpus, extract *Vocabulary*
- Calculate $P(c_j)$ terms
  - For each $c_j$ do
    - $docs_j \leftarrow$ all docs with class $= c_j$
    - $P(c_j) \leftarrow \frac{|docs_j|}{\text{total \# documents}}$
- Calculate $P(w_i | c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word $w_i$ in *Vocabulary*
    - $n_i \leftarrow$ # of occurrences of $w_i$ in $Text_j$
    - $P(w_i | c_j) \leftarrow \frac{n_i + \alpha}{n + \alpha |Vocabulary|}$
### Example

<table>
<thead>
<tr>
<th>Doc</th>
<th>Words</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1</td>
<td>Chinese Beijing Chinese</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Chinese Chinese Shanghai</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Chinese Macao</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Tokyo Japan Chinese</td>
</tr>
<tr>
<td>Test</td>
<td>5</td>
<td>Chinese Chinese Chinese Tokyo Japan</td>
</tr>
</tbody>
</table>
Example

\[ \hat{P}(c) = \frac{N_c}{N} \]

Priors:
\[ P(c) = \frac{3}{4} \quad \frac{1}{4} \]

<table>
<thead>
<tr>
<th>Doc</th>
<th>Words</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1  Chinese Beijing Chinese</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>2  Chinese Chinese Shanghai</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>3  Chinese Macao</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>4  Tokyo Japan Chinese</td>
<td>j</td>
</tr>
<tr>
<td>Test</td>
<td>5  Chinese Chinese Chinese Tokyo Japan</td>
<td>?</td>
</tr>
</tbody>
</table>
**Example**

\[
\hat{P}(c) = \frac{N_c}{N}
\]

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

**Priors:**
- \( P(c) = \frac{3}{4} \)
- \( P(j) = \frac{1}{4} \)

**Conditional Probabilities:**
- \( P(\text{Chinese} \mid c) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7} \)
- \( P(\text{Tokyo} \mid c) = \frac{0+1}{8+6} = \frac{1}{14} \)
- \( P(\text{Japan} \mid c) = \frac{0+1}{8+6} = \frac{1}{14} \)
- \( P(\text{Chinese} \mid j) = \frac{1+1}{3+6} = \frac{2}{9} \)
- \( P(\text{Tokyo} \mid j) = \frac{1+1}{3+6} = \frac{2}{9} \)
- \( P(\text{Japan} \mid j) = \frac{1+1}{3+6} = \frac{2}{9} \)
Example

\[ \hat{P}(c) = \frac{N_c}{N} \]

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

**Priors:**

\[ P(c) = \frac{3}{4}, \quad \frac{1}{4} \]

**Conditional Probabilities:**

\[
\begin{align*}
P(\text{Chinese} \mid c) &= \frac{(5+1)}{(8+6)} = \frac{6}{14} = \frac{3}{7} \\
P(\text{Tokyo} \mid c) &= \frac{(0+1)}{(8+6)} = \frac{1}{14} \\
P(\text{Japan} \mid c) &= \frac{(0+1)}{(8+6)} = \frac{1}{14} \\
P(\text{Chinese} \mid j) &= \frac{(1+1)}{(3+6)} = \frac{2}{9} \\
P(\text{Tokyo} \mid j) &= \frac{(1+1)}{(3+6)} = \frac{2}{9} \\
P(\text{Japan} \mid j) &= \frac{(1+1)}{(3+6)} = \frac{2}{9}
\end{align*}
\]

**Choosing a class:**

\[
\begin{align*}
P(c \mid d5) &\propto 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \\
&\approx 0.0003
\end{align*}
\]

\[
\begin{align*}
P(j \mid d5) &\propto 1/4 \times (2/9)^3 \times 2/9 \times 2/9 \\
&\approx 0.0001
\end{align*}
\]
Summary: naïve Bayes is not so naïve

- Naïve Bayes is a probabilistic model
- Naïve because it assumes features are independent of each other for a class
- 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
  - Decision Trees suffer from fragmentation in such cases – especially if little data
- 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
  - But we will see other classifiers that give better accuracy
Classification evaluation

- **Contingency table**: model’s predictions are compared to the correct results
  - a.k.a. confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>actual pos</th>
<th>actual neg</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>predicted pos</strong></td>
<td>true positive (tp)</td>
<td>false positive (fp)</td>
</tr>
<tr>
<td><strong>predicted neg</strong></td>
<td>false negative (fn)</td>
<td>true negative (tn)</td>
</tr>
</tbody>
</table>

![Confusion matrix diagram](image)

- true positive (tp)
- false positive (fp)
- false negative (fn)
- true negative (tn)
Classification evaluation

- Borrowing from Information Retrieval, empirical NLP systems are usually evaluated using the notions of **precision** and **recall**
Classification evaluation

- Precision ($P$) is the proportion of the selected items that the system got right in the case of text categorization
  - it is the % of documents classified as “positive” by the system which are indeed “positive” documents
- Reported per class or average

Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} = \frac{tp}{tp + fp}
Classification evaluation

- Recall (R) is the proportion of actual items that the system selected in the case of text categorization
  - it is the % of the “positive” documents which were actually classified as “positive” by the system
- Reported per class or average

\[
\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = \frac{tp}{tp + fn}
\]
Classification evaluation

- We often want to trade-off precision and recall
  - typically: the higher the precision the lower the recall
  - can be plotted in a precision-recall curve
- It is convenient to combine P and R into a single measure
  - one possible way to do that is F measure

\[
F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2P + R} \quad \text{for } \beta=1, \quad F_1 = \frac{2PR}{P+R}
\]
Classification evaluation

- Additional measures of performance: accuracy and error
  - accuracy is the proportion of items the system got right
  - error is its complement

\[
\text{accuracy} = \frac{tp + tn}{tp + fp + tn + fn}
\]
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.
Classification common practices

- Divide the training data into $k$ folds (e.g., $k=10$)
- Repeat $k$ times: train on $k-1$ folds and test on the holdout fold, cyclically
- Average over the $k$ folds’ results
K-fold cross-validation

Diagram showing how data is split into folds for cross-validation.
K-fold 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**
  - Handles sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance

![Diagram of K-fold cross-validation](image)
Next class

- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
    - Discriminative models
      - Naïve Bayes
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron
Readings

- Eis 2
- J&M III 4