Chapter 6:
Text Classification
Text Categorization and related Tasks
Classification

Goal: Assign ‘objects’ from a universe to two or more *classes* or *categories*

Examples:

<table>
<thead>
<tr>
<th>Problem</th>
<th>Object</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sense Disambiguation</td>
<td>Word/Doc.</td>
<td>The word’s senses</td>
</tr>
<tr>
<td>Tagging</td>
<td>Words</td>
<td>POS/NE</td>
</tr>
<tr>
<td>Spam Mail Detection</td>
<td>Document</td>
<td>spam/not spam</td>
</tr>
<tr>
<td>Author identification</td>
<td>Document</td>
<td>Authors</td>
</tr>
<tr>
<td>Text Categorization</td>
<td>Document</td>
<td>Topic</td>
</tr>
<tr>
<td>Information retrieval</td>
<td>Document</td>
<td>Relevant/not relevant</td>
</tr>
</tbody>
</table>
Spam/junk/bulk Emails

- The messages you spend your time with just to delete them
  - Spam: do not want to get, unsolicited messages
  - Junk: irrelevant to the recipient, unwanted
  - Bulk: mass mailing for business marketing (or fill-up mailbox etc)

Classification task: decide for each e-mail whether it is spam/not-spam
They agreed that Mrs. X should only hear of the departure of the family, without being alarmed on the score of the gentleman's conduct; but even this partial communication gave her a great deal of concern, and she bewailed it as exceedingly unlucky that the ladies should happen to go away, just as they were all getting so intimate together.

Gas looming through the fog in divers places in the streets, much as the sun may, from the spongey fields, be seen to loom by husbandman and ploughboy. Most of the shops lighted two hours before their time—as the gas seems to know, for it has a haggard and unwilling look. The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest near that leaden-headed old obstruction, appropriate ornament for the threshold of a leaden-headed old corporation, Temple Bar.
They agreed that Mrs. X should only hear of the departure of the family, without being alarmed on the score of the gentleman's conduct; but even this partial communication gave her a great deal of concern, and she bewailed it as exceedingly unlucky that the ladies should happen to go away, just as they were all getting so intimate together.
bemoaned it as exceedingly unlucky
Did you mean: "bewailed it as exceedingly unlikely"

Jane Austen: Pride and Prejudice, Chapter XXI of Volume I (Chap. 21)
... and she bewailed it as exceedingly unlucky that the ladies should happen to go away, just as they were all getting so intimate together. ...
www.pemberley.com/janeinfo/ppv1n21.html - 19k - Cached - Similar pages

Chapter XXI, Austen, Jane: 1317. Pride and Prejudice, Vol. II ...
... and she bewailed it as exceedingly unlucky that the ladies should happen to go away just as they were all getting so intimate together. ...
www.bartleby.com/303/2/21.html - 33k - Cached - Similar pages

Pride and Prejudice
File Format: PDF/Adobe Acrobat
concern, and she bewailed it as exceedingly unlucky that the ladies should happen to go away just as they were all getting so intimate together. ...
www.bartleby.com/ebook/acrobat/3032.pdf - Similar pages

Pride and Prejudice
... and he bewailed it as exceedingly unlucky that the gentleman should happen to go away, just as they were all getting so intimate together. ...
www.lifeisgood.com/ppnchapters/19_21.html - 35k - Cached - Similar pages
• Gas looming through the fog in divers places in the streets, much as the sun may, from the spongey fields, be seen to loom by husbandman and ploughboy. Most of the shops lighted two hours before their time—as the gas seems to know, for it has a haggard and unwilling look. The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest near that leaden-headed old obstruction, appropriate ornament for the threshold of a leaden-headed old corporation, Temple Bar.
Dickens London Walks, Temple Bar, A Tale of Two Cities, Sweeney...
In Bleak House he described it as "that leaden-headed old obstruction, ... In 1868 the leaden-headed old obstruction was transferred to Theobald's park in ...
www.london-walks.co.uk/3D/dickens-london-walks-temp.shtml - Similar pages

Language Log: Step on a crack, break a grammar rule
The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest near that leaden-headed old obstruction, appropriate ornament ...
lite.cis.upenn.edu/~myl/languagelog/archives/002224.html - 19k - Cached - Similar pages

www.sussex.ac.uk/Users/lyw/Bodle%20metaphor%20in%20...
File Format: Microsoft Powerpoint - View as HTML
muddiest near that leaden-headed old obstruction, appropriate ornament for the threshold of a leaden-headed old corporation, Temple Bar. And hard by ...
Similar pages

Randomhouse | Books | Bleak House by Charles Dickens
The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest, near that leaden-headed old obstruction, appropriate ornament ...
www.randomhouse.com/catalog/display. pper?isbn=9780375760051&view=excerpt - 29k - Cached - Similar pages

cityofsound: Bleak House Without A Foggy Day in London Town
"The raw afternoon is rawest, and the dense fog is densest, and the muddy streets are muddiest near that leaden-headed old obstruction, appropriate ornament ...
www.cityofsound.com/blog/2006/01/bleak_house_wt.html - 44k - Cached - Similar pages
Author identification

- Jane Austen (1775-1817), Pride and Prejudice
- Charles Dickens (1812-70), Bleak House
Author identification

- Federalist papers
  - 77 short essays written in 1787-1788 by Hamilton, Jay and Madison to persuade NY to ratify the US Constitution; published under a pseudonym
  - The authorships of 12 papers was in dispute (disputed papers)
  - In 1964 Mosteller and Wallace* solved the problem
  - They identified 70 *function* words as good candidates for authorships analysis
  - Using statistical inference they concluded the author was Madison
<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>15</th>
<th>do</th>
<th>29</th>
<th>is</th>
<th>43</th>
<th>or</th>
<th>57</th>
<th>this</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>all</td>
<td>16</td>
<td>down</td>
<td>30</td>
<td>it</td>
<td>44</td>
<td>our</td>
<td>58</td>
<td>to</td>
</tr>
<tr>
<td>3</td>
<td>also</td>
<td>17</td>
<td>even</td>
<td>31</td>
<td>its</td>
<td>45</td>
<td>shall</td>
<td>59</td>
<td>up</td>
</tr>
<tr>
<td>4</td>
<td>an</td>
<td>18</td>
<td>every</td>
<td>32</td>
<td>may</td>
<td>46</td>
<td>should</td>
<td>60</td>
<td>upon</td>
</tr>
<tr>
<td>5</td>
<td>and</td>
<td>19</td>
<td>for</td>
<td>33</td>
<td>more</td>
<td>47</td>
<td>so</td>
<td>61</td>
<td>was</td>
</tr>
<tr>
<td>6</td>
<td>any</td>
<td>20</td>
<td>from</td>
<td>34</td>
<td>must</td>
<td>48</td>
<td>some</td>
<td>62</td>
<td>were</td>
</tr>
<tr>
<td>7</td>
<td>are</td>
<td>21</td>
<td>had</td>
<td>35</td>
<td>my</td>
<td>49</td>
<td>such</td>
<td>63</td>
<td>what</td>
</tr>
<tr>
<td>8</td>
<td>as</td>
<td>22</td>
<td>has</td>
<td>36</td>
<td>no</td>
<td>50</td>
<td>than</td>
<td>64</td>
<td>when</td>
</tr>
<tr>
<td>9</td>
<td>at</td>
<td>23</td>
<td>have</td>
<td>37</td>
<td>not</td>
<td>51</td>
<td>that</td>
<td>65</td>
<td>which</td>
</tr>
<tr>
<td>10</td>
<td>be</td>
<td>24</td>
<td>her</td>
<td>38</td>
<td>now</td>
<td>52</td>
<td>the</td>
<td>66</td>
<td>who</td>
</tr>
<tr>
<td>11</td>
<td>been</td>
<td>25</td>
<td>his</td>
<td>39</td>
<td>of</td>
<td>53</td>
<td>their</td>
<td>67</td>
<td>will</td>
</tr>
<tr>
<td>12</td>
<td>but</td>
<td>26</td>
<td>if</td>
<td>40</td>
<td>on</td>
<td>54</td>
<td>then</td>
<td>68</td>
<td>with</td>
</tr>
<tr>
<td>13</td>
<td>by</td>
<td>27</td>
<td>in</td>
<td>41</td>
<td>one</td>
<td>55</td>
<td>there</td>
<td>69</td>
<td>would</td>
</tr>
<tr>
<td>14</td>
<td>can</td>
<td>28</td>
<td>into</td>
<td>42</td>
<td>only</td>
<td>56</td>
<td>things</td>
<td>70</td>
<td>your</td>
</tr>
</tbody>
</table>

Table 1: Function Words and Their Code Numbers
Function words for Author Identification

Figure 1: Obtained Hyperplane in 3 dimensions
Text Categorization

- Speech Recognition
- Information Retrieval
- Computer Linguistics
- Everything else
Text categorization

• Topic categorization: classify the document into semantics topics

The U.S. swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan defeated Belarus's Max Mirnyi and Vladimir Voltchkov to give the Americans an unsurmountable 3-0 lead in the best-of-five semi-final tie.

One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as the plodding approach of Hurricane Jeanne prompted evacuation orders for hundreds of thousands of Floridians and high wind warnings that stretched 350 miles from the swamp towns south of Miami to the historic city of St. Augustine.
Israel targets Hamas leaders
The Standard - 1 hour ago
Israeli tanks and troops massed near Gaza for a threatened offensive against the Palestinians, and the Israeli government said it would target Hamas leaders if a captured soldier was not freed. Israeli tanks ...

The New York Times
National Review Online Blogs - 20 hours ago
The New York Times' decision to disclose the Terrorist Finance Tracking Program, a robust and classified effort to map terrorist networks through the use of financial data, was irresponsible and harmful to the security of Americans and freedom-loving ...

Buffett: Gates' charity 'sweeter way' to helping
TMonet - all 1,867 related

Intel Unveils Xeon 5100 Processors
Techtree.com - all 263 related

UNC throws away national title
NBC17.com - all 1,247 related

FW review: 'Superman' is only average
CNN International - all 252 related

Sexual orientation of men determined before birth
Reuters - all 411 related

Get recommended stories
Sign in to get recommended stories by using search history
Text categorization

- Reuters
  - Collection of (21,578) newswire documents.
  - For research purposes: a standard text collection to compare systems and algorithms
  - 135 valid topics categories
Top topics in Reuters

Top Topics

Number of Stories

Category

CORPORATE/INDUSTRIAL
GOVERNMENT/SOCIAL
MARKETS
PERFORMANCE
ACCOUNTING/MARKETS
COMMUNITY/SENIORS
DOMESTIC POLITICS
EXCHANGE/MARKETS
GOVERNMENT FINANCE
INTERNATIONAL RELATIONS
REGULATORY POLICY
WEAR, CIVIL WAR
CRIME, LAW ENFORCEMENT
CAPITAL/INFRASTRUCTURE
The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nation's pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter
Classification vs. Clustering
Classification vs. Clustering

- **Classification** assumes labeled data: we know how many classes there are and we have examples for each class (labeled data).
- Classification is supervised
- In **Clustering** we don’t have labeled data; we just assume that there is a natural division in the data and we may not know how many divisions (clusters) there are
- Clustering is unsupervised
Classification

Class1

Class2
Classification

Class1

Class2
Classification

Class 1

Class 2
Classification

- Green smiley faces: Class1
- Red smiley faces: Class2

[Diagram showing a linear boundary separating Class1 and Class2]
Clustering
Clustering
Clustering
Clustering
Clustering
Binary vs. multi-way classification

- Binary classification: two classes
- Multi-way classification: more than two classes
- Sometime it can be convenient to treat a multi-way problem like a binary one: one class versus all the others, for all classes
Flat vs. Hierarchical classification

- Flat classification: relations between the classes undetermined
- Hierarchical classification: hierarchy where each node is the sub-class of its parent’s node
Single- vs. multi-category classification

- In single-category text classification each text belongs to exactly one category.

- In multi-category text classification, each text can have zero or more categories.
Getting Features for Text Categorization
Feature terminology

• Feature: An aspect of the text that is relevant to the task
• Feature value: the realization of the feature in the text
  • Words present in text: Clinton, Schumacher, China...
  • Frequency of word: Clinton(10), Schumacher(1)...
  • Are there dates? Yes/no
  • Are there PERSONS? Yes/no
  • Are there ORGANIZATIONS? Yes/no
  • WordNet: Holonyms (China is part of Asia), Synonyms (China, People's Republic of China, mainland China)
Feature Types

- **Boolean (or Binary) Features**
- Features that generate boolean (binary) values.
- Boolean features are the simplest and the most common type of feature.

\[
\begin{align*}
  f_1(\text{text}) &= 1 \text{ if text contain “Clinton”} \\
  &= 0 \text{ otherwise} \\
  f_2(\text{text}) &= 1 \text{ if text contain PERSON} \\
  &= 0 \text{ otherwise}
\end{align*}
\]
Feature Types

- **Integer Features**
- Features that generate integer values.
- Integer features can be used to give classifiers access to more precise information about the text.

  - \( f_1(\text{text}) = \text{Number of times text contains “Clinton”} \)
  - \( f_2(\text{text}) = \text{Number of times text contains PERSON} \)
When Do We Need Feature Selection?

- If the algorithm cannot handle all possible features
  - e.g. language identification for 100 languages using all words
  - text classification using $n$-grams

- Good features can result in higher accuracy
  - But! Why feature selection?
  - What if we just keep all features?
    - Even the unreliable features can be helpful.
    - But we need to weight them:
      - In the extreme case, the bad features can have a weight of 0 (or very close), which is... a form of feature selection!
Why Feature Selection?

- Not all features are equally good!
  - Bad features: best to remove
    - Infrequent
      - unlikely to be met again
      - co-occurrence with a class can be due to chance
    - Too frequent
      - mostly function words
    - Uniform across all categories
  - Good features: should be kept
    - Co-occur with a particular category
    - Do not co-occur with other categories
  - The rest: good to keep
Types Of Feature Selection?

- Feature selection reduces the number of features
  - Usually:
    - Eliminating features
    - Weighting features
    - Normalizing features
  - Sometimes by transforming parameters
    - e.g. Latent Semantic Indexing using Singular Value Decomposition
- Method may depend on problem type
  - For classification and filtering, may use information from example documents to guide selection
Feature Selection

- Task independent methods
  - Document Frequency (DF)
  - Term Strength (TS)

- Task-dependent methods
  - Information Gain (IG)
  - Mutual Information (MI)
  - $\chi^2$ statistic (CHI)

Empirically compared by Yang & Pedersen (1997)
• Compared feature selection methods for text categorization
  - 5 feature selection methods:
    • DF, MI, CHI, IG, TS
    • Features were just words
  - 2 classifiers:
    • kNN: k-Nearest Neighbor (to be covered next week)
    • LLSF: Linear Least Squares Fit
  - 2 data collections:
    • Reuters-22173
    • OHSUMED: subset of MEDLINE (1990&1991 used)
Document Frequency (DF)

*DF: number of documents a term appears in*

- Based on Zipf’s Law
- Remove the **rare** terms: (met 1-2 times)
  - Non-informative
  - Unreliable – can be just noise
  - Not influential in the final decision
  - Unlikely to appear in new documents
- Plus
  - Easy to compute
  - Task independent: do not need to know the classes
- Minus
  - Ad hoc criterion
  - Rare terms can be good discriminators (e.g., in IR)
Examples of Frequent Words:
Most Frequent Words in Brown Corpus

<table>
<thead>
<tr>
<th>Word</th>
<th>Instances % Frequency</th>
<th>Instances % Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>69970 6.8872</td>
<td>at</td>
</tr>
<tr>
<td>of</td>
<td>36410 3.5839</td>
<td>by</td>
</tr>
<tr>
<td>and</td>
<td>28854 2.8401</td>
<td>I</td>
</tr>
<tr>
<td>to</td>
<td>26154 2.5744</td>
<td>this</td>
</tr>
<tr>
<td>a</td>
<td>23363 2.2996</td>
<td>had</td>
</tr>
<tr>
<td>in</td>
<td>21345 2.1010</td>
<td>not</td>
</tr>
<tr>
<td>that</td>
<td>10594 1.0428</td>
<td>are</td>
</tr>
<tr>
<td>is</td>
<td>10102 0.9943</td>
<td>but</td>
</tr>
<tr>
<td>was</td>
<td>9815 0.9661</td>
<td>from</td>
</tr>
<tr>
<td>He</td>
<td>9542 0.9392</td>
<td>or</td>
</tr>
<tr>
<td>for</td>
<td>9489 0.9340</td>
<td>have</td>
</tr>
<tr>
<td>it</td>
<td>8760 0.8623</td>
<td>an</td>
</tr>
<tr>
<td>with</td>
<td>7290 0.7176</td>
<td>they</td>
</tr>
<tr>
<td>as</td>
<td>7251 0.7137</td>
<td>which</td>
</tr>
<tr>
<td>his</td>
<td>6996 0.6886</td>
<td>one</td>
</tr>
<tr>
<td>on</td>
<td>6742 0.6636</td>
<td>you</td>
</tr>
<tr>
<td>be</td>
<td>6376 0.6276</td>
<td>were</td>
</tr>
</tbody>
</table>
Stop Word Removal

• Common words from a predefined list
  • Mostly from closed-class categories:
    • unlikely to have a new word added
    • include: auxiliaries, conjunctions, determiners, prepositions, pronouns, articles
  • But also some open-class words like numerals

• Bad discriminators
  • uniformly spread across all classes
  • can be safely removed from the vocabulary
    • *Is this always a good idea?* (e.g. *author identification*)
Information Gain

- A measure of importance of the feature for predicting the presence of the class.
- Defined as:
  - The number of “bits of information” gained by knowing the term is present or absent
  - Based on Information Theory
- Plus:
  - sound information theory justification
- Minus:
  - computationally expensive
Information Gain (IG)

IG: number of bits of information gained by knowing the term is present or absent

\[ G(t) = - \sum_{i=1}^{m} P(c_i) \log P(c_i) \]

\[ + P(t) \sum_{i=1}^{m} P(c_i | t) \log P(c_i | t) \]

\[ + P(\bar{t}) \sum_{i=1}^{m} P(c_i | \bar{t}) \log P(c_i | \bar{t}) \]

\[ t \text{ is the term being scored,} \]
\[ c_i \text{ is a class variable} \]
Mutual Information (MI)

Logarithmic version of correlation to term \( t \) with category \( c \)

\[
I(t, c) = \log \left( \frac{P(t, c)}{P(t)P(c)} \right)
\]

\[
= \log \left( \frac{P(t \mid c)}{P(t)} \right)
\]

\[
= \log \left( \frac{P(c \mid t)}{P(c)} \right)
\]
Using Mutual Information

- Compute MI for each category and then combine
  - If we want to discriminate well across all categories, then we need to take the expected value of MI:

    \[ I_{\text{avg}}(t) = \sum_{i=1}^{m} P(c_i) I(t, c_i) \]

- To discriminate well for a single category, then we take the maximum:

    \[ I_{\text{max}}(t) = \max_{i=1\ldots m} I(t, c_i) \]
Mutual Information

- **Plus**
  - \( I(t, c) \) is 0, when \( t \) and \( c \) are independent
  - Sound information-theoretic interpretation

- **Minus**
  - Small numbers produce unreliable results
  - No weighting with frequency of a pair \((t,c)\)
\( \chi^2 \) statistic

- The most commonly used method of comparing proportions.

**Example:** Let us measure the dependency between a term \( t \) and a category \( c \).
  - the groups would be:
    - 1) the documents from a category \( c_i \)
    - 2) all other documents
  - the characteristic would be:
    - “document contains term \( t \)”
**Is “jaguar” a good predictor for the “auto” class?**

<table>
<thead>
<tr>
<th></th>
<th>Term = jaguar</th>
<th>Term ≠ jaguar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class = auto</td>
<td>2</td>
<td>500</td>
</tr>
<tr>
<td>Class ≠ auto</td>
<td>3</td>
<td>9500</td>
</tr>
</tbody>
</table>

We want to compare:
- the observed distribution above; and
- null hypothesis: that *jaguar* and *auto* are independent
\( \chi^2 \) statistic

Under the null hypothesis: (jaguar and auto – independent): How many co-occurrences of jaguar and auto do we expect?

- We would have: \( P(j,a) = P(j) \times P(a) \)
- \( P(j) = (2+3)/N; P(a) = (2+500)/N; N=2+3+500+9500 \)
- Num. co-occur. :
  - \( N \times P(j,a) = N \times P(j) \times P(a) \)
  - \( N \times (5/N) \times (502/N) = 2510/N = 2510/10005 \approx 0.25 \)

<table>
<thead>
<tr>
<th>Class = auto</th>
<th>Term = jaguar</th>
<th>Term ≠ jaguar</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 (0.25)</td>
<td>500</td>
</tr>
<tr>
<td>Class ≠ auto</td>
<td>3</td>
<td>9500</td>
</tr>
</tbody>
</table>
### $\chi^2$ statistic

<table>
<thead>
<tr>
<th></th>
<th>Term $=$ jaguar</th>
<th>Term $\neq$ jaguar</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class $=$ auto</strong></td>
<td>2 (0.25)</td>
<td>500 (502)</td>
</tr>
<tr>
<td><strong>Class $\neq$ auto</strong></td>
<td>3 (4.75)</td>
<td>9500 (9498)</td>
</tr>
</tbody>
</table>
\[ \chi^2\text{ statistic} \]

\[ \chi^2 \text{ is interested in } (f_o - f_e)^2/f_e \text{ summed over all table entries:} \]

\[ \chi^2(j,a) = \sum (O - E)^2 / E = (2 - .25)^2 / .25 + (3 - 4.75)^2 / 4.75 \]
\[ + (500 - 502)^2 / 502 + (9500 - 9498)^2 / 9498 = 12.9 \]

<table>
<thead>
<tr>
<th>Class = auto</th>
<th>Term = jaguar</th>
<th>Term ≠ jaguar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class ≠ auto</td>
<td>3 (4.75)</td>
<td>9500 (9498)</td>
</tr>
<tr>
<td>Class = auto</td>
<td>2 (0.25)</td>
<td>500 (502)</td>
</tr>
</tbody>
</table>

\[ \chi^2 = \sum (O - E)^2 / E \]

\[ = (2 - .25)^2 / .25 + (3 - 4.75)^2 / 4.75 \]
\[ + (500 - 502)^2 / 502 + (9500 - 9498)^2 / 9498 = 12.9 \]

\[ \chi^2(j,a) = \sum (O - E)^2 / E \]

\[ = (2 - .25)^2 / .25 + (3 - 4.75)^2 / 4.75 \]
\[ + (500 - 502)^2 / 502 + (9500 - 9498)^2 / 9498 = 12.9 \]
$\chi^2$ statistic

Alternatives:

- Look up value for $\chi^2$ in a table
- Calculate it from

\[
f(x, k) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}
\]

- Look it up in the internet
The null hypothesis is rejected with confidence 0.9997
\(\chi^2\) statistic

Collect all the terms to calculate \(\chi^2\) directly from contingency table

\[
\chi^2(t, c) = \frac{N(AD - CB)^2}{(A + B)(A + C)(B + D)(C + D)}
\]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>#(t,c)</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>#(t,\neg c)</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>#(\neg t,c)</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>#(\neg t, \neg c)</td>
<td></td>
</tr>
</tbody>
</table>

\[
N = A + B + C + D
\]
Using Mutual Information

- Compute MI for each category and then combine
  - If we want to discriminate well across all categories, then we need to take the expected value of MI:

\[
I_{avg}(t) = \sum_{i=1}^{m} P(c_i) I(t, c_i)
\]

- To discriminate well for a single category, then we take the maximum:

\[
I_{avg}(t) = \max_{i=1\ldots m} I(t, c_i)
\]
$\chi^2$ statistic

How to use $\chi^2$ for multiple categories?

Compute $\chi^2$ for each category and then combine:

- we can require to discriminate well across all categories, then we need to take the expected value of $\chi^2$:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P(c_i) \chi^2(t, c_i)$$

- or to discriminate well for a single category, then we take the maximum:

$$\chi^2_{\text{max}}(t) = \max_{i=1\ldots m} \chi^2(t, c_i)$$
\( \chi^2 \) statistic

- **Plus**
  - normalized and thus comparable across terms
  - \( \chi^2(t, c) \) is 0, when \( t \) and \( c \) are independent
  - sound theoretical background

- **Minus**
  - unreliable for low frequency terms
  - computationally expensive
Term strength

Term strength:

\[ s(t) = p(t \in y \mid t \in x) \]

\(x, y\): topically related document
  (e.g. from a clustering algorithm)

• measures co-occurrence of terms (unlike idf)
• For more details see:
  Wilbur and Sorotkin
  **The automatic identification of stop words**
Comparison on Reuters

Figure 1. Average precision of kNN vs. unique word count

- IG
- DF
- TS
- CHF.max
- MI.max

Number of features (unique words)

11-pt average precision
Correlation of feature selection criteria
Correlation of feature selection criteria

Figure 4. Correlation between DF and CHI values of words in Reuters

DF-CHI plots
Feature Selection Summary (From Yang and Pedersen)

<table>
<thead>
<tr>
<th>Method</th>
<th>DF</th>
<th>IG</th>
<th>CHI</th>
<th>MI</th>
<th>TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>favoring common terms</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y/N</td>
</tr>
<tr>
<td>using categories</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>using term absence</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>performance in kNN/LLSF</td>
<td>excellent</td>
<td>excellent</td>
<td>excellent</td>
<td>poor</td>
<td>ok</td>
</tr>
</tbody>
</table>
Classification Algorithms
Overview

- There is a large zoo of classification algorithms
  - Decision Trees
  - Naïve Bayes
    - Maximum Entropy methods
  - k nearest neighbor classifiers
  - Neural networks
  - Support vector machines
- Many of them have been covered in other lectures
Decision Tree for Reuter classification

**Figure 16.1** A decision tree. This tree determines whether a document is part of the topic category “earnings” or not. $P(c|n_i)$ is the probability of a document at node $n_i$ to belong to the “earnings” category $c$. 

From Manning & Schütze
Decision Boundaries for Decision Trees
1-Nearest Neighbor
1-Nearest Neighbor
3-Nearest Neighbor
3-Nearest Neighbor

Assign the category of the majority of the neighbors

But this is closer.. We can weight neighbors according to their similarity
Decision Boundaries for $k$ Nearest Neighbor (schematic)
Bayes Decision Rule

\[
\omega_k = \arg \max_{\omega_k} [P(x | \omega_k)P(\omega_k)]
\]

\( \omega_k \): class label

\( x \): features
Naïve Bayes

- $x$ is not a single feature, but a bag of features
e.g. different key-words for your spam-mail detection system
- Assume statistical independence of features

$$P(\{x_1...x_N\} | \omega_k) \approx \prod_{i=1}^{N} P(x_i | \omega_k)$$
Maximum Entropy Methods

- A way to estimate probabilities
- Features are taken into account as constraints for the probabilities
- Otherwise as “unbiased” probability estimate as possible
Linear binary classification using a Perceptron (Simplest Neural Network)

- **Data:** \( \{(x_i, y_i)\}_{i=1}^{n} \)
  - \( x \) in \( \mathbb{R}^d \) (\( x \) is a vector in \( d \)-dimensional space)
    - feature vector
  - \( y \) in \( \{-1, +1\} \)
    - label (class, category)

- **Question:**
  - Design a linear decision boundary: \( wx + b \) (equation of hyperplane) such that the classification rule associated with it has minimal probability of error

- **Classification rule:**
  - \( y = \text{sign}(wx + b) \) which means:
    - if \( wx + b > 0 \) then \( y = +1 \)
    - if \( wx + b < 0 \) then \( y = -1 \)
Linear binary classification

- Find a good hyperplane \((w, b) \in \mathbb{R}^{d+1}\) that correctly classifies data points as much as possible.
- In online fashion: one data point at the time, update weights as necessary.

Classification Rule: \(y = \text{sign}(wx + b)\)
Perceptron algorithm

- **Initialize**: $w_1 = 0$
- **Updating rule** For each data point $x$
  - If $\text{class}(x) \neq \text{decision}(x, w)$
  - then
  $$w_{k+1} = w_k + y_i x_i$$
  $$k = k + 1$$
  - else
  $$w_{k+1} = w_k$$

- **Function** $\text{decision}(x, w)$
  - If $wx + b > 0$ return $+1$
  - Else return $-1$
Perceptron algorithm

- **Online**: can adjust to changing target, over time
- **Advantages**
  - Simple and computationally efficient
  - Guaranteed to learn a linearly separable problem (convergence, global optimum)
- **Limitations**
  - Only linear separations
  - Only converges for linearly separable data
  - Not really “efficient with many features”
Another family of linear algorithms

Intuition (Vapnik, 1965)

If the classes are linearly separable:
  • Separate the data
  • Place hyper-plane “far” from the data: large margin
  • Statistical results guarantee good generalization

Large margin classifier
Large margin classifier

- **Intuition** (Vapnik, 1965) if linearly separable:
  - Separate the data
  - Place hyperplane “far” from the data: **large margin**
  - Statistical results guarantee **good generalization**
Large margin classifier

If not linearly separable

- **Allow** some errors
- Still, try to place hyperplane “far” from each class
Large Margin Classifiers

- **Advantages**
  - Theoretically better (better error bounds)
- **Limitations**
  - Computationally more expensive, large quadratic programming
Support Vector Machine (SVM)

- Large Margin Classifier
- Linearly separable case
- Goal: find the hyperplane that maximizes the margin

\[ w^T x + b = 0 \]

\[ w^T x_a + b = 1 \]

\[ w^T x_b + b = -1 \]
Summary

- Types of text classification
- Features and feature selection
- Classification algorithms