Text categorization

Important task for both humans and machines

- object identification
- face recognition
- spoken word recognition

How does it apply to text?

Text categorization: example

Dear Prof. Sharon Goldwater:

My name is [XX]. I am an ambitious applicant for the Ph.D program of Electrical Engineering and Computer Science at your university. Especially being greatly attracted by your research projects and admiring for your achievements via the school website, I cannot wait to write a letter to express my aspiration to undertake the Ph.D program under your supervision.

I have completed the M.S. program in Information and Communication Engineering with a high GPA of 3.95/4.0 at [YY] University. In addition to throwing myself into the specialized courses in [...] I took part in the research projects, such as [...] I really enjoyed taking the challenges in the process of the researches and tests, and I spent two years on the research project [...]. We proved the effectiveness of the new method for [...] and published the result in [...].

Having read your biography, I found my academic background and research experiences indicated some possibility of my qualification to join your team. It is my conviction that the enlightening instruction, cutting-edge research projects and state-of-the-art facilities offered by your team will direct me to make breakthroughs in my career development in the arena of electrical engineering and computer science. Thus, I shall be deeply grateful if you could give me the opportunity to become your student. Please do not hesitate to contact me, should you need any further information about my scholastic and research experiences.

Yours sincerely, [XX].

Text categorization (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)
Text categorization (classification)

Or we might want to categorize the author of the text (authorship attribution):

- Native language identification (e.g., to tailor language tutoring)
- Diagnosis of disease (psychiatric or cognitive impairments)
- Identification of gender, dialect, educational background (e.g., in forensics [legal matters], advertising/marketing).

Formalizing the task

Given document $d$ and set of categories $C$, we want to assign $d$ to the most probable category $\hat{c}$:

$$\hat{c} = \arg\max_{c \in C} P(c|d)$$

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

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

Document model

Each document $d$ is represented by features $f_1, f_2, \ldots f_n$, e.g.:

- For topic classification: 2000 most frequent words, excluding stopwords like the, a, do, in.
- For sentiment classification: words from a sentiment lexicon

In fact, we only care about the feature counts, so this is a bag-of-words (unigram) model.

Task-specific features

Example words from a sentiment lexicon:

Positive:
- absolutely beaming calm
- adorable beautiful benefit
- believe bliss bountiful
- bountiful bouncy brave champion
- charm challenge cheerful
- choice classical clean
-... ...

Negative:
- abysmal adverse alarmed angry annoy
- anxious apathy appalling atrocious awful
- bad banal barbed belligerent bemoan
- beneath boring broken
-... ...

From http://www.enchantedlearning.com/wordlist/
Example documents

- Possible feature counts from training documents in a spam-detection task (where we did not exclude stopwords):

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>your</th>
<th>model</th>
<th>cash</th>
<th>Viagra</th>
<th>class</th>
<th>account</th>
<th>orderz</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc 1</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>doc 2</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>doc 3</td>
<td>25</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>doc 4</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>doc 5</td>
<td>17</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Document model, cont.

- Representing $d$ using its features gives us:

$$P(d|c) = P(f_1, f_2, \ldots f_n|c)$$

- But we can’t estimate this joint probability well (too sparse).

- So, make a Naive Bayes assumption: features are conditionally independent given class.

$$P(d|c) \approx P(f_1|c)P(f_2|c) \ldots P(f_n|c)$$

Full model

- Given document with features $f_1, f_2, \ldots f_n$ and set of categories $C$, choose

$$\hat{c} = \arg\max_{c\in C} P(c) \prod_{i=1}^{n} P(f_i|c)$$

- This is called a Naive Bayes classifier

  - see Basic Prob Theory reading Ex 5.5.3 for a non-text example

Generative process

- Naive Bayes classifier is another generative model.

- Assumes the data (features in each doc) were generated as
  - For each document, choose its class $c$ with prob $P(c)$.
  - For each feature in each doc, choose the value of that feature with prob $P(f|c)$
Learning the class priors

- \( P(c) \) normally estimated with MLE:
  \[
  \hat{P}(c) = \frac{N_c}{N}
  \]
  - \( N_c \) = the number of training documents in class \( c \)
  - \( N \) = the total number of training documents

Learning the feature probabilities

- \( P(f_i|c) \) normally estimated with simple smoothing:
  \[
  \hat{P}(f_i|c) = \frac{\text{count}(f_i, c) + \alpha}{\sum_{f \in F}(\text{count}(f_i, c) + \alpha)}
  \]
  - \( \text{count}(f_i, c) \) = the number of times \( f_i \) occurs in class \( c \)
  - \( F \) = the set of possible features
  - \( \alpha \): the smoothing parameter, optimized on held-out data

Learning the class priors: example

- Given training documents with correct labels:

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>your</th>
<th>model</th>
<th>cash</th>
<th>Viagra</th>
<th>class</th>
<th>account</th>
<th>orderz</th>
<th>spam?</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc 1</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>doc 2</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>doc 3</td>
<td>25</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>doc 4</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>doc 5</td>
<td>17</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
</tbody>
</table>

- \( \hat{P}(\text{spam}) = 3/5 \)

Learning the feature probabilities: example

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>your</th>
<th>model</th>
<th>cash</th>
<th>Viagra</th>
<th>class</th>
<th>account</th>
<th>orderz</th>
<th>spam?</th>
</tr>
</thead>
<tbody>
<tr>
<td>doc 1</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>doc 2</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>+</td>
</tr>
<tr>
<td>doc 3</td>
<td>25</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>doc 4</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>doc 5</td>
<td>17</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>+</td>
</tr>
</tbody>
</table>
Learning the feature probabilities: example

| the your model cash Viagra class account orderz spam? |
|---|---|---|---|---|---|---|---|---|
| doc 1 | 12 | 3 | 1 | 0 | 0 | 2 | 0 | 0 | - |
| doc 2 | 10 | 4 | 0 | 4 | 0 | 0 | 2 | 0 | + |
| doc 3 | 25 | 4 | 0 | 0 | 0 | 1 | 1 | 0 | - |
| doc 4 | 14 | 2 | 0 | 1 | 3 | 0 | 1 | 1 | + |
| doc 5 | 17 | 5 | 0 | 2 | 0 | 1 | 1 | 1 | + |

\[ \hat{P}(\text{your}|+) = \frac{(4+2+5+\alpha)}{(\text{all words in + class})+\alpha F} = \frac{11 + \alpha}{(68 + \alpha F)} \]

\[ \hat{P}(\text{orderz}|+) = \frac{(2+\alpha)}{(\text{all words in + class})+\alpha F} = \frac{2 + \alpha}{(68 + \alpha F)} \]

Classifying a test document: example

- Test document \(d\):
  - get your cash and your orderz
- Suppose there are no other features besides those in previous table (so get and and are not counted). Then
  \[ P(+|d) \propto P(+) \prod_{i=1}^{n} P(f_i|+) \]
  \[ = P(+) \cdot \frac{11 + \alpha}{(68 + \alpha F)} \cdot \frac{7 + \alpha}{(68 + \alpha F)} \]
  \[ \cdot \frac{11 + \alpha}{(68 + \alpha F)} \cdot \frac{2 + \alpha}{(68 + \alpha F)} \]

Classifying a test document: example

- Test document \(d\):
  - get your cash and your orderz
- Do the same for \(P(−|d)\)
- Choose the one with the larger value
Alternative feature values and feature sets

- Use only **binary** values for $f_i$: did this word occur in $d$ or not?
- Use only a subset of the vocabulary for $F$
  - Ignore **stopwords** (function words and others with little content)
  - Choose a small task-relevant set (e.g., using a sentiment lexicon)
- Use more complex features (bigrams, syntactic features, morphological features, ...)

Task-specific features

- But: other words might be relevant for specific sentiment analysis tasks.
  - E.g., **quiet**, **memory** for product reviews.
- And for other tasks, stopwords might be very useful features
  - E.g., People with schizophrenia use more 2nd-person pronouns (Watson et al., 2012), those with depression use more 1st-person (Rude et al., 2004).
- Probably better to use too many irrelevant features than not enough relevant ones.

Advantages of **Naive Bayes**

- Very easy to implement
- Very fast to train and test
- Doesn’t require as much training data as some other methods
- Usually works reasonably well

This should be your baseline method for any classification task
Problems with Naive Bayes

- Naive Bayes assumption is naive!
- Consider categories Travel, Finance, Sport.
- Are the following features independent given the category?
  beach, sun, ski, snow, pitch, palm, football, relax, ocean

No! They might be closer if we defined finer-grained categories (beach vacations vs. ski vacations), but we don’t usually want to.

Non-independent features

- Features are not usually independent given the class
- Adding multiple feature types (e.g., words and morphemes) often leads to even stronger correlations between features
- Accuracy of classifier can sometimes still be ok, but it will be highly overconfident in its decisions.
  - Ex: NB sees 5 features that all point to class 1, treats them as five independent sources of evidence.
  - Like asking 5 friends for an opinion when some got theirs from each other.

How to evaluate performance?

- Important question for any NLP task
- Intrinsic evaluation: design a measure inherent to the task
  - Language modeling: perplexity
  - POS tagging: accuracy (% of tags correct)
  - Categorization: F-score (coming up next)


### How to evaluate performance?

- Important question for any NLP task

- **Intrinsic** evaluation: design a measure inherent to the task
  - Language modeling: perplexity
  - POS tagging: accuracy (% of tags correct)
  - Categorization: F-score (coming up next)

- **Extrinsic** evaluation: measure effects on a downstream task
  - Language modeling: does it improve my ASR/MT system?
  - POS tagging: does it improve my parser/IR system?
  - Categorization: does it reduce user search time in an IR setting?

### Intrinsic evaluation for categorization

- Categorization as detection: document about sport or not?

- Classes may be very unbalanced.

- Can get 95% accuracy by always choosing “not”; but this isn’t useful.

- Need a better measure.

### Two measures

- Assume we have a gold standard: correct labels for test set

- We also have a system for detecting the items of interest (docs about sport)

\[
\text{Precision} = \frac{\# \text{ items detected and was right}}{\# \text{ items system detected}}
\]

\[
\text{Recall} = \frac{\# \text{ items detected and was right}}{\# \text{ items system should have detected}}
\]

### Example of precision and recall

<table>
<thead>
<tr>
<th>Gold standard</th>
<th>System output</th>
<th>Doc about sports?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y Y N N Y N N N Y N N</td>
<td>Y Y N N Y N N N Y N N</td>
<td>N Y N Y N N N Y N N</td>
</tr>
</tbody>
</table>

- # ‘Y’ we got right = 2
- # ‘Y’ we guessed = 3
- # ‘Y’ in GS = 4
- Precision = 2/3
- Recall = 2/4
Why use both measures?

Systems often have (implicit or explicit) tuning thresholds on how many answers to return.

- e.g., Return as `Y` all docs where system thinks $P(C=\text{sport})$ is greater than $t$.
- Raise $t$: higher precision, lower recall.
- Lower $t$: lower precision, higher recall.

<table>
<thead>
<tr>
<th>Doc</th>
<th>Sys prob</th>
<th>Gold</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>0.99</td>
<td>Y</td>
</tr>
<tr>
<td>12</td>
<td>0.98</td>
<td>Y</td>
</tr>
<tr>
<td>45</td>
<td>0.93</td>
<td>Y</td>
</tr>
<tr>
<td>01</td>
<td>0.93</td>
<td>Y</td>
</tr>
<tr>
<td>37</td>
<td>0.89</td>
<td>N</td>
</tr>
<tr>
<td>24</td>
<td>0.84</td>
<td>Y</td>
</tr>
<tr>
<td>16</td>
<td>0.78</td>
<td>Y</td>
</tr>
<tr>
<td>18</td>
<td>0.75</td>
<td>N</td>
</tr>
<tr>
<td>20</td>
<td>0.72</td>
<td>Y</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>38</td>
<td>0.03</td>
<td>N</td>
</tr>
<tr>
<td>19</td>
<td>0.03</td>
<td>N</td>
</tr>
</tbody>
</table>

**F-measure**

- Can also combine precision and recall into single **F-measure**:
  
  $$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2P + R}$$

- Normally we just set $\beta = 1$ to get $F_1$:
  
  $$F_1 = \frac{2PR}{P + R}$$

- $F_1$ is the harmonic mean of $P$ and $R$: similar to arithmetic mean when $P$ and $R$ are close, but penalizes large differences between $P$ and $R$.

**References**
