Data Mining
(EECS 6412)

Text Classification

Aijun An
Department of Electrical Enginering and Computer Science
York University

Text Mining

- It refers to data mining using text documents as data.
  - A text document could be an article, a web page, an xml file, an email message, a blog and so on.
- Tasks of text mining
  - Text classification
  - Text clustering
  - Text summarization
  - Topic detection
  - ...

Text Classification

- Learn a classification model from a set of pre-classified documents
- Classify new text documents using the learned model
- Applications
  - Classify articles into categories
  - Classify web pages into different categories
  - Classify emails into different categories
  - Spam email filtering
  - ...

3

Example Applications

- News topic classification (e.g., Google News)
  C={politics, sports, business, health, tech, …}
- "SafeSearch" filtering
  C={pornography, not pornography}
- Language classification
  C={English, Spanish, Chinese,…}
- Sentiment classification
  C={positive review, negative review}
- Email sorting
  C={spam, meeting reminders, invitations, …} – user-defined!

4

Text Representation

- Most classification learning programs require the examples to be represented as a tuple, which is a vector of attribute values.
- How to represent a document using a vector of attribute values?
- Attributes
  - "Bag of words" method: Use a set of words as attributes
- Attribute values
  - Method 1: use 0 or 1 as attribute value to indicate whether the word appears in the document.
  - Method 2: use the absolute or relative frequency of each word in the document as the attribute value.
  - Method 3: assign a weight to a word in a document using TF-IDF and use the weight as the attribute value

5

Text Representation (*Cont'd*)

Training data sets:

- Method 1:

	word$_1$	**word$_2$**	**...**	**word$_m$**	**Class**
document$_1$	0	1	…	1	C1
document$_2$	1	0	…	1	C2
…	…	…	…	…	…
document$_n$	1	0		0	C2

- Method 2 with absoluate term frequency:

	word$_1$	**word$_2$**	**...**	**word$_m$**	**Class**
document$_1$	0	3	…	1	C1
document$_2$	2	0	…	3	C2
…	…	…	…	…	…
document$_n$	5	0		0	C2

6

Method 3: TF-IDF Term Weighting

- TF: term frequency
  - Definition: TF = t_{ij}
    - frequency of term *i* in document *j*
  - Purpose: makes the frequent words *for the document* more important
- IDF: inverted document frequency
  - Definition: IDF = $\log(N/n_i)$
    - n_i : number of documents containing term *i*
    - N : total number of documents
  - Purpose: makes rare words *across documents* more important
- TF-IDF value of a term *i* in document *j*
  - Definition: TF×IDF = $t_{ij} \times \log(N/n_i)$

7

Example: TF-IDF Weighted Vectors

Assume there are three documents in the training set:

Document D1: "yes we got no bananas"

Document D2: "what you got"

Document D3: "yes I like what you got"

	yes	we	got	no	bananas	what	you	I	like
D1:	.18	0.48	0	0.48	0.48	0	0	0	0
D2:	0	0	0	0	0	0.18	0.18	0	0
D3:	0.18	0	0	0	0	0.18	0.18	0.48	.48

8

Text Processing for Selecting the Bag of Words

- Word (token) extraction
  - Extract all the words in a document
  - Convert them into lower cases
- Stop words removal
- Stemming
- Selecting words

9

Stop Words

- Many of the most frequently used words in English are worthless in text mining – these words are called *stop words*.
- Examples of stop words

  the, of, and, to, a, …
- Typically about 400 to 500 such words
- For an application, there may be additional domain-specific stop words
- These stop words are usually removed from the set of words for representing a document.

10

Stemming

- A technique used to find the root/stem of a word.
- For example:
  - discussed
  - discusses
  - discussing
  - discuss

  Stem: discuss
- Usefulness
  - Reduce the number of words
  - Improve effectiveness of text classification

11

Example Stemming Rules

- Remove ending
  - If a word ends with *s*, preceded by a consonant other than an *s*, then delete the *s*.
  - If a word ends with *ed*, preceded by a consonant, delete the *ed* unless this leaves only a single letter.
- Transform words
  - If a word ends with “ies” but not “eies” or “aies”, then “ies” is replaced with “y”.

12

Stemming Algorithms

- Porter stemming algorithm
  - The most widely used stemming algorithm
  - Developed by Martin Porter at the University of Cambridge in 1980
  - http://www.tartarus.org/~martin/PorterStemmer/ contains source codes in a few languages
- Other stemming algorithms
  - http://www.comp.lancs.ac.uk/computing/research/stemming/general/

13

Text Processing for Selecting the Bag of Words

- Word (token) extraction
  - Extract all the words in a document
  - Convert them into lower cases
- Stop words removal
- Stemming
- *Selecting words*

14

Feature Selection

- Selecting the "bag of words" to represent documents
- Why do we need to select?
  - The number of unique words in a set of documents can be too many.
  - Leaning program may not be able to handle all possible features
  - Good features can result in higher accuracy

15

What are Good and Bad Features?

- Good features: (should be kept)
  - Co-occur with a particular category
  - Do not co-occur with other categories
- Bad features: (best to remove)
  - Uniform across all categories
  - Very infrequent (appear 1 or 2 times in the whole training set of documents)
    - unlikely to be met again
    - can be noise
    - co-occurrence with a class can be due to chance

16

Feature Selection Methods

- Class independent methods (Unsupervised)
  - Document Frequency (DF)
  - Term Strength (TS)
- Class-dependent methods (Supervised)
  - Information Gain (IG)
  - Mutual Information (MI)
  - χ^2 statistic (CHI)

17

Document Frequency (DF)

- Document frequency of a word *w:*

  DF (*w*) = number of documents containing *w*

- Rank the words according to their document frequency
- Select the first *m* words with high DF values

18

Document Frequency (*Cont'd*)

- Advantages
  - Easy to compute
  - Can remove rare words (hence noise)
- Disadvantages
  - Class independent:
    - If the word appears frequently in many classes, it cannot distinguish the classes well
  - Some infrequent terms can be good discriminators, which cannot be selected by this method.

19

Information Gain

- A measure of importance of the feature for predicting the classes of documents
- Defined as:
  - The number of "bits of information" gained by knowing the word is present or absent

$$Gain(w) = -\sum_{i=1}^{k} P(C_i)\log P(C_i) + P(w)\sum_{i=1}^{k} P(C_i \mid w)\log P(C_i \mid w) + P(\overline{w})\sum_{i=1}^{k} P(C_i \mid \overline{w})\log P(C_i \mid \overline{w})$$

where w is a word and $C_1, C_2, \ldots, C_k$ are classes.

- Rank the words according to their information gain value
- Select the first m words with high gain values

20

Information Gain (*Cont'd*)

- Advantage:
  - Consider the classes
- Disadvantage:
  - Computationally expensive (compared to using DF)
  - Noisy words occurring only once in the document collection have high IG
- Solution
  - Remove rare words (appears 1 or 2 times) first. This can
    - reduce the amount of computation, and
    - remove noisy words that have by-chance correlations with the classes.

21

What Do People Do In Practice?

- Rare term removal
  - rare across the whole collection (i.e. DF is very low)
  - met in a single document
- Most frequent term removal (i.e. removing stop words)
- Stemming. (*often*)
- Use a class-dependent method (e.g., the information gain method) to select features.

22

Beyond Words

- Bag of words representation
  - does not consider the position or order of words in a document
  - does not consider the context a word is in.
- It would be great to include multi-word features like "New York", rather than just "New" and "York"
- Bigram document representation (or n-gram in general)
  - a pair of consecutive words in the document
- But: including all pairs of words, or all consecutive pairs of words, as features creates WAY too many features to deal with, and many are very sparse.

23

Summary

- Text classification has many applications
- The most important issue is how to represent documents
  - Word extraction
  - Stop word removal
  - Stemming
  - Feature selection
  - Represent document with values of the selected features (e.g., the frequency of the word in the document).

24

References

- Feature Selection
  - Yang Y., J. Pedersen. A comparative study on feature selection in text categorization. In J. D. H. Fisher, editor, The Fourteenth International Conference on Machine Learning (ICML'97), pages 412-420. Morgan Kaufmann, 1997.
- Term Weighting
  - Salton G., C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing and Management: an International Journal, v.24 n.5, p.513-523, 1988.
  - Salton, G. 1989. Automatic text processing. Chapter 9.

25