NLP
Introduction to NLP

Text Classification
Classification

• Assigning documents or sentences to predefined categories
  – topics, languages, users …

• Input:
  – a document (or sentence) $d$
  – a fixed set of classes $C = \{c_1, c_2, ..., c_j\}$

• Output: a predicted class $c \in C$
Variants of Problem Formulation

• Binary categorization: only two categories
  – Retrieval: \{relevant-doc, irrelevant-doc\}
  – Spam filtering: \{spam, non-spam\}
  – Opinion: \{positive, negative\}

• K-category categorization: more than two categories
  – Topic categorization: \{sports, science, travel, business,...\}
  – Word sense disambiguation: \{bar1, bar2, bar3, ...\}

• Hierarchical vs. flat

• Overlapping (soft) vs non-overlapping (hard)
Hierarchical Classification

Image from Schütze & Krisnawati
Borges’s Classification

- The list divides all animals into 14 categories:
- Those that belong to the emperor
- Embalmed ones
- Those that are trained
- Sucking pigs
- Mermaids (or Sirens)
- Fabulous ones
- Stray dogs
- Those that are included in this classification
- Those that tremble as if they were mad
- Innumerable ones
- Those drawn with a very fine camel hair brush
- Et cetera
- Those that have just broken the flower vase
- Those that, at a distance, resemble flies

*Celestial Emporium of Benevolent Knowledge*
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
Supervised Machine Learning

- A given set of classes C
- Given text data $x$, determine its class $y$ in C
DEAR SIR

Funds for Investments

This letter may come to you as a surprise since I had no previous correspondence with you.

I am the Chairman Tender Board of Independent National Electoral Commission INEC. I got your contact in the course of my search for a reliable person with whom to handle a very confidential transaction involving the transfer of fund valued at twenty-one million six hundred thousand United States dollars US$20M to a safe foreign account.
SpamAssassin

- http://spamassassin.apache.org/
- HowScoresAreAssigned
- http://spamassassin.apache.org/tests_3_3_x.html

Example features:
- **body** Incorporates a tracking ID number
- **body** HTML and text parts are different
- **header** Date: is 3 to 6 hours before Received: date
- **body** HTML font size is huge
- **header** Attempt to obfuscate words in Subject:
- **header** Subject =~ /^urgent(?:\s=W)*\W(dollar) | .{1,40}(?:alert| response| assistance| proposal| reply| warning| noti(?:\s=ce| fication)| greeting| matter))$/
Features for Classification

• Vector-based
  – Words: “cat”, “dog”, “great”, “horrible”, etc.
  – Meta features: document length, author name, etc.
  – Each document (or sentence) is represented as a vector in an $n$-dimensional space
    • Similar documents appear nearby in the vector space
  – (more later)
Introduction to NLP

Vector Space Classification
Vector Space Classification

The diagram illustrates a vector space classification with two axes, $x_1$ and $x_2$. The points are labeled with red circles for topic 1 and white stars for topic 2.
Decision surfaces

\[ x_1 \]

\[ x_2 \]

☆ topic2

〇 topic1
Decision trees

x1

x2

☆ topic2
○ topic1
Classification Using Centroids

• Centroid
  – the point most representative of a class

• Compute centroid by finding vector average of known class members

• Decision boundary is a line that is equidistant from two centroids.

• New document on one side of the goes in one class; new document on the other side goes in the other.
Linear boundary
Classification Using Centroids

\[ x_1 \]

\[ x_2 \]

- topic2
- topic1
- centroid2
- centroid1
Introduction to NLP

Linear Models
Linear Separators

• Two-dimensional line:
  \[ w_1 x_1 + w_2 x_2 = b \] is the linear separator
  \[ w_1 x_1 + w_2 x_2 > b \] for the positive class

• In n-dimensional spaces:
  \[ \vec{w}^T \vec{x} = b \]
Linear Models

\[ \mathbf{w}^T \mathbf{x} = \sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \]

• One can also add \( w_1 = 1 \), \( x_0 = b \) (constant)
• \( \mathbf{w} \) is the weight vector
• \( \mathbf{x} \) is the feature vector
Example

\[ \mathbf{w}^T \mathbf{x} = b \]

- Bias \( b = 0 \) (in this example)
- Sentence is “A D E H”
- Its score will be

\[
0.6\times1 + 0.4\times1 + 0.4\times1 + (-0.5) \times 1 = 0.9 > 0
\]
How to Find the Linear Boundary?

• Find the linear boundary = $\vec{w}$
  find

• Many methods
  – Perceptron
  – Stochastic Gradient Descent
  – Linear least squares
  – ...

• Problem:
  – There are infinite number of linear boundaries if the two classes are linearly separable!
General Training Idea

• Go through the training data
• Predict the class $y$ (1 or $-1$)
• If the prediction is wrong, update $w$
  
  \[ w_{(t+1)} = w_t + yx \]
• Used in the perceptron
Naïve Bayes (again)

- Multinomial NB is a linear model
- $x = [1, \#w_1, \#w_2, \ldots]$  
- $w = [\log P(y), \log P(w_1|y), \log P(w_2|y), \ldots]$  
- (see separate lecture)
NLP