10.04 Text Classification

NATURAL LANGUAGE PROCESSING

NLP

Introduction to NLP

Text Classification

Classification

- Assigning documents to predefined categories
  - topics, languages, users
- A given set of classes C
  - Given x, determine its class in C
- Hierarchical vs. flat
- Overlapping (soft) vs non-overlapping (hard)

Classification

- Ideas: manual classification using rules
  - e.g., Columbia AND University → Education
    Columbia AND "South Carolina" → Geography
- Popular techniques
  - generative (k-nn, Naïve Bayes) vs. discriminative (SVM, regression)
- Generative
  - model joint prob p(x,y) and use Bayesian prediction to compute p(y|x)
- Discriminative
  - model p(y|x) directly.

Representations For Document Classification (And Clustering)

- Typically: vector-based
  - Words: “cat”, “dog”, etc.
  - Features: document length, author name, etc.
- Each document is represented as a vector in an *n*-dimensional space
- Similar documents appear nearby in the vector space (distance measures are needed)

Naïve Bayesian classifiers

- Naïve Bayesian classifier

$$P(d \in C \mid F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots F_k \mid d \in C) P(d \in C)}{P(F_1, F_2, \ldots F_k)}$$

- Assuming statistical independence

$$P(d \in C \mid F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j \mid d \in C) P(d \in C)}{\prod_{j=1}^{k} P(F_j)}$$

- Features = words (or phrases) typically

Issues with Naïve Bayes

- Where do we get the values $P(d \in C)$
  - use maximum likelihood estimation (N_i/N)
- Same for the conditionals
  - these are based on a multinomial generator and the MLE estimator is ($T_{ji}/\Sigma T_{ji}$)
- Smoothing is needed
  - why
  - Laplace smoothing ($(T_{ji}+1)/\Sigma(T_{ji}+1)$)
- Implementation
  - how to avoid floating point underflow

Spam Recognition

Return-Path: <ig_esq@rediffmail.com>
X-Sieve: CMU Sieve 2.2
From: "Ibrahim Galadima" <ig_esq@rediffmail.com>
Reply-To: galadima_esq@netpiper.com
To: webmaster@aclweb.org
Subject: Gooday

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/
- http://spamassassin.apache.org/tests_3_3_x.html
- Examples:
  - 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]*(dollar) | .{1,40}(?:alert| response| assistance| proposal| reply| warning| noti(?:ce| fication)| greeting| matter))/i

Feature Selection: The X^2 Test

- For a term *t:*

		I_t	
		0	1
C	0	k_{00}	k_{01}
	1	k_{10}	k_{11}

- C=class, i_t = feature
- Testing for independence: $P(C=0,I_t=0)$ should be equal to $P(C=0)\ P(I_t=0)$
  - $P(C=0) = (k_{00}+k_{01})/n$
  - $P(C=1) = 1-P(C=0) = (k_{10}+k_{11})/n$
  - $P(I_t=0) = (k_{00}+K_{10})/n$
  - $P(I_t=1) = 1-P(I_t=0) = (k_{01}+k_{11})/n$

Feature Selection: The X^2 Test

$$X^2 = \frac{n(k_{11}k_{00} - k_{10}k_{01})^2}{(k_{11} + k_{10})(k_{01} + k_{00})(k_{11} + k_{01})(k_{10} + k_{00})}$$

- High values of X^2 indicate lower belief in independence.
- In practice, compute X^2 for all words and pick the top k among them.

Feature Selection: Mutual Information

- No document length scaling is needed
- Documents are assumed to be generated according to the multinomial model
- Measures amount of information: if the distribution is the same as the background distribution, then MI=0
- X = word; Y = class

$$MI(X,Y) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

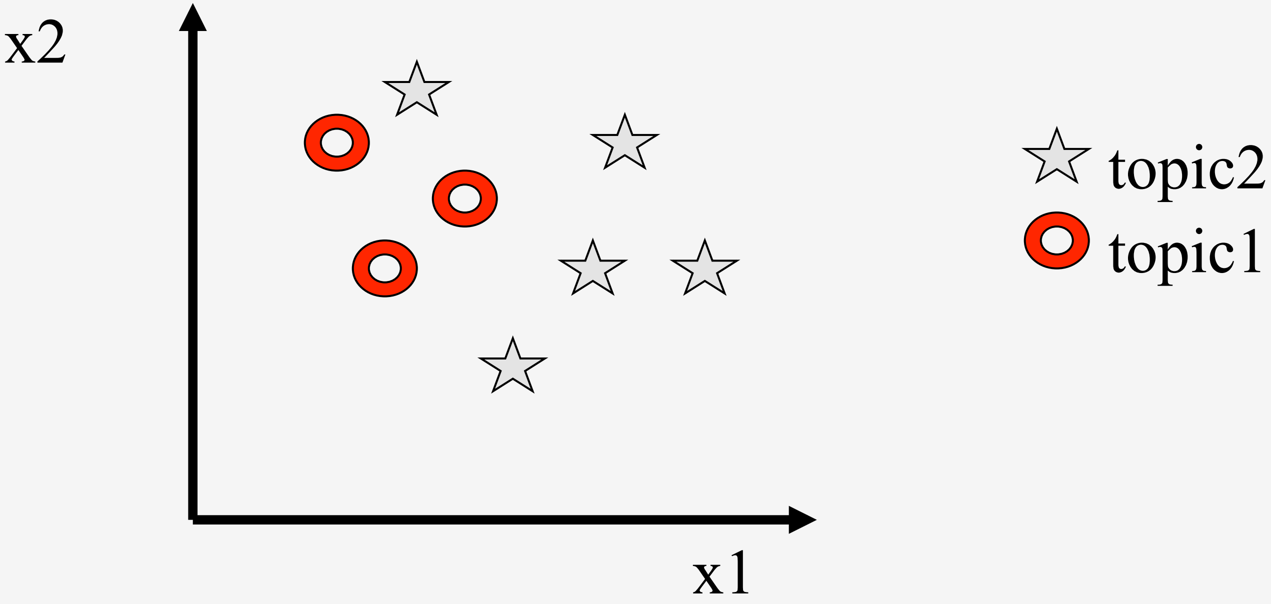
Well-known Datasets

- 20 newsgroups
  - http://qwone.com/~jason/20Newsgroups/
- Reuters-21578
  - http://www.daviddlewis.com/resources/testcollections/reuters21578/
  - Cats: grain, acquisitions, corn, crude, wheat, trade…
- WebKB
  - http://www-2.cs.cmu.edu/~webkb/
  - course, student, faculty, staff, project, dept, other
- RCV1
  - http://www.daviddlewis.com/resources/testcollections/rcv1/
  - Larger Reuters corpus

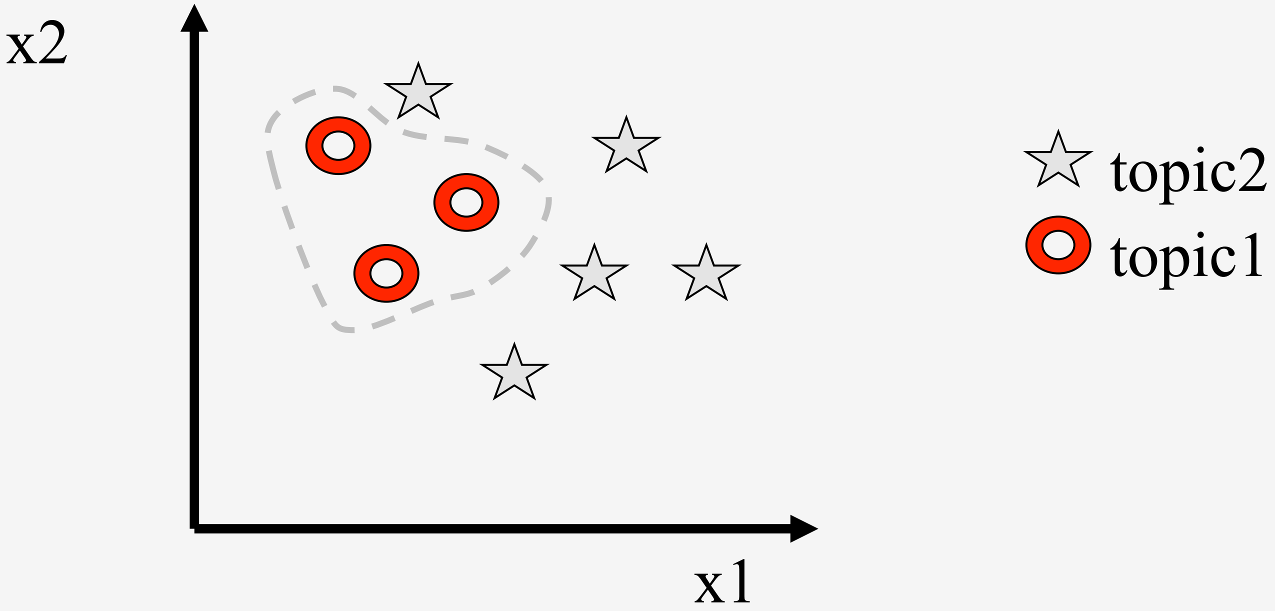
Evaluation Of Text Classification

- Microaveraging
  - average over classes
- Macroaveraging
  - uses pooled table

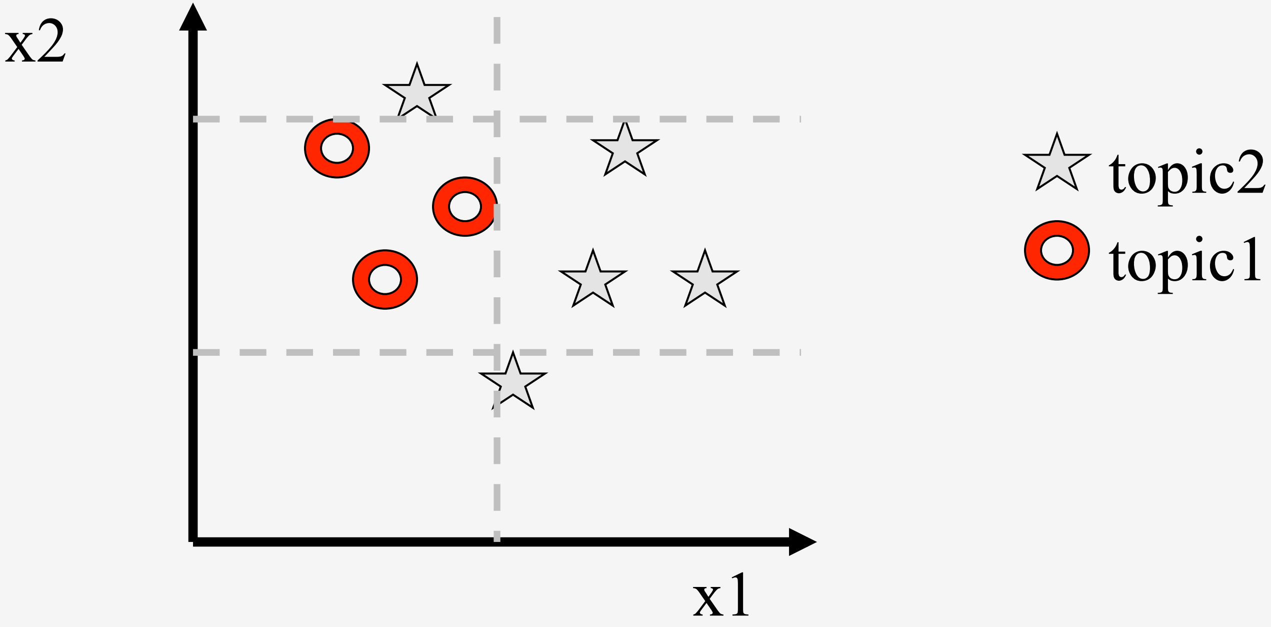
Vector Space Classification



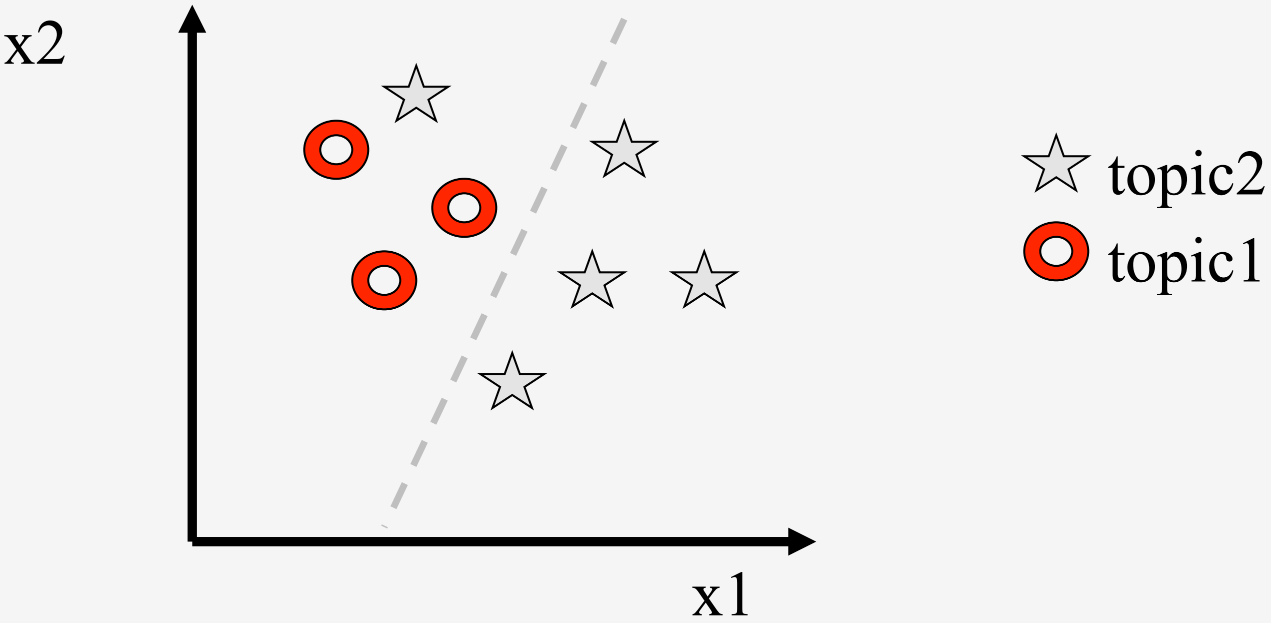
Decision Surfaces



Decision Trees



Linear Boundary



Vector Space Classifiers

- Using centroids
- Boundary
  - line that is equidistant from two centroids

Linear Separators

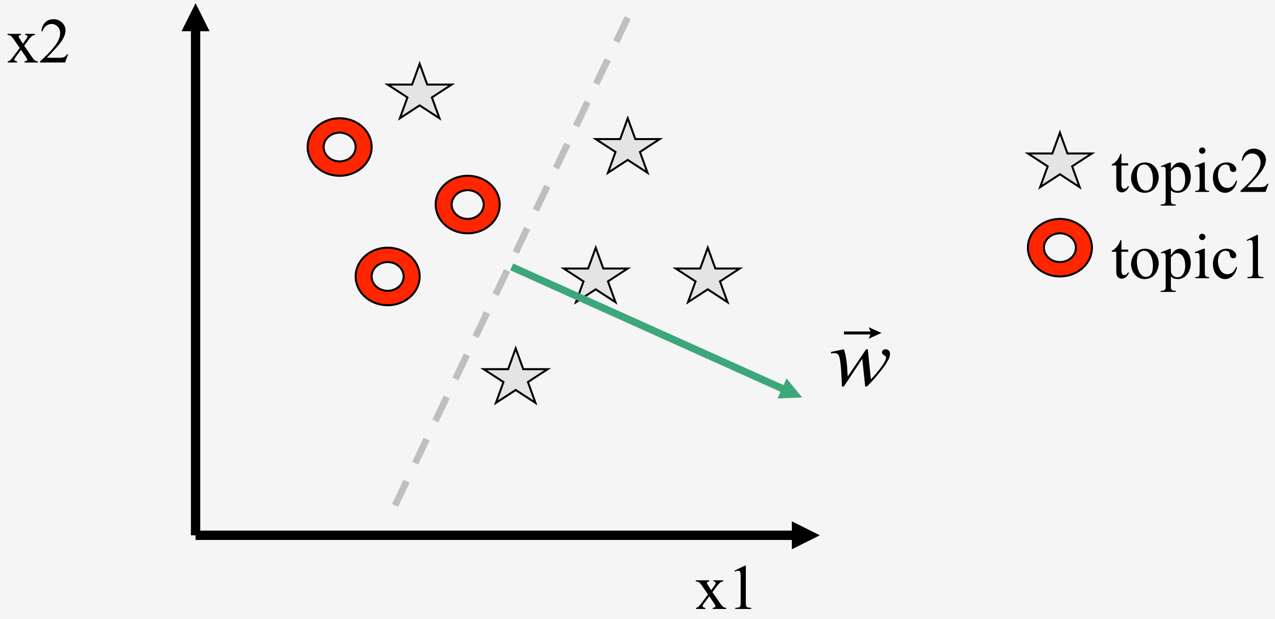
- Two-dimensional line:

  $w_1x_1+w_2x_2=b$ is the linear separator

  $w_1x_1+w_2x_2>b$ for the positive class
- In n-dimensional spaces:

$$\vec{w}^T\vec{x} = b$$

Decision Boundary



Example

- Bias b=0
- Document is “A D E H”
- Its score will be

  0.6*1+0.4*1+0.4*1+(–0.5)*1 = 0.9>0

w_i	x_i	w_i	x_i
0.6	A	-0.7	G
0.5	B	-0.5	H
0.5	C	-0.3	I
0.4	D	-0.2	J
0.4	E	-0.2	K
0.3	F	-0.2	L

Perceptron Algorithm

Input: $S = ((\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)), \vec{x}_1 \in \Re^N, y_i \in \{-1,1\}$

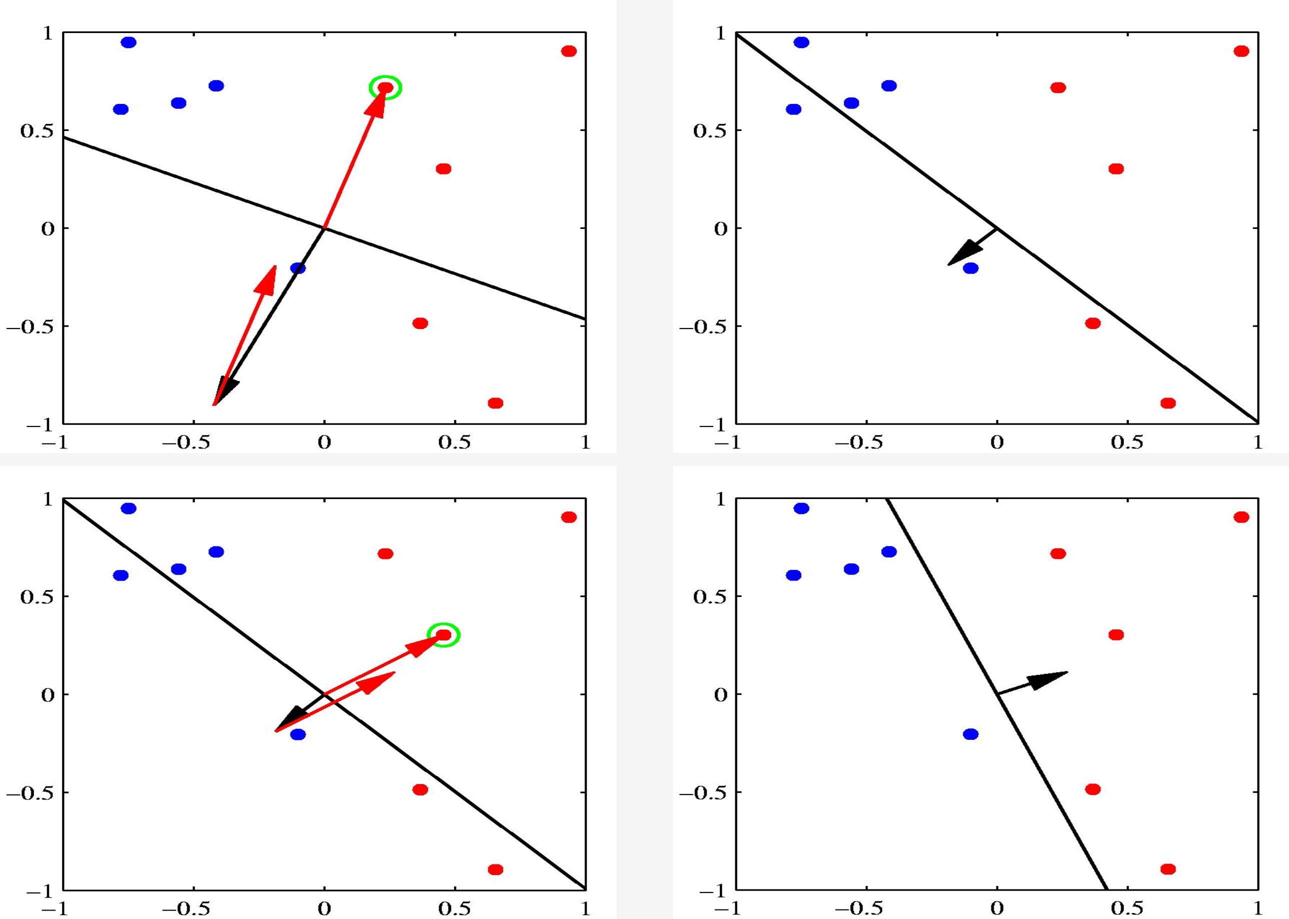
$\eta \in \Re$

Algorithm:

```
$\vec{w}_0 = \vec{0}, k = 0$
FOR $i = 1$ TO $n$
    IF $y_i(\vec{w}_k \bullet \vec{x}_i) \leq 0$
        $\vec{w}_{k+1} = \vec{w}_k + \eta y_i \vec{x}_i$
        $k = k + 1$
    END
END
```

Output: $\vec{w}_k$

10.04 Text Classification



[Example from Chris Bishop]

Generative Models: knn

- Assign each element to the closest cluster
- K-nearest neighbors

$$score(c,d_q) = b_c + \sum_{d \in kNN(d_q)} s(d_q,d)$$

- Very easy to program
- Issues:
  - choosing k, b?
- Demo:
  - http://www-2.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html

NLP