Automated Text Categorization

Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

Introduction

- Purpose: classification of natural language texts into a set of predefined labels.

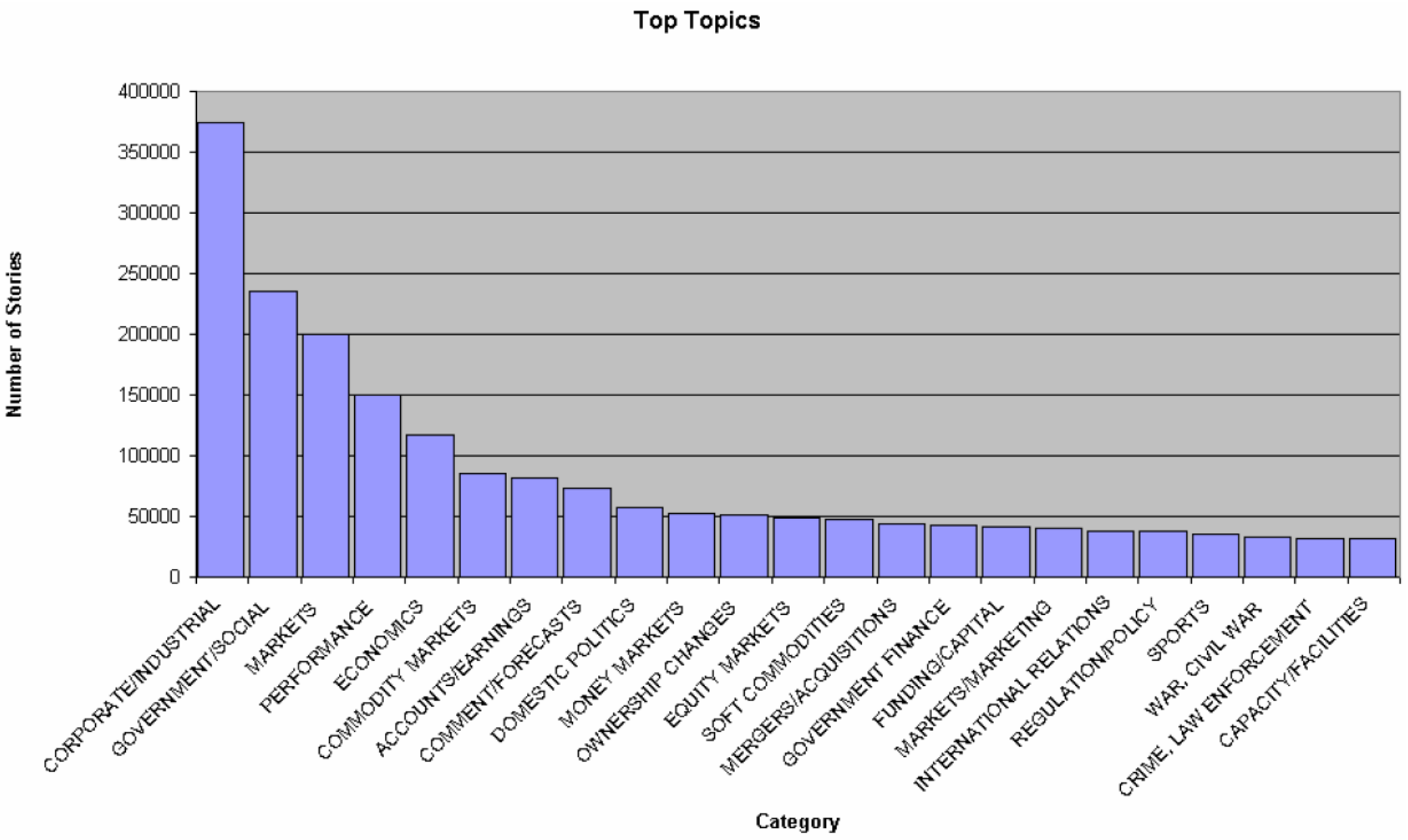
Save 15% on Supplements
Everyday!!
Join our Nutritional Supplement
Discount Program and take 15%
Off Supplement Shelf prices every
day of the year.
Quick, easy sign-up & start saving
immediately.

spam? Or legitimate?

Main Uses

- Indexing (e.g. Libraries)
- Organization
  - News articles (Reuters, GoogleNews)
  - Classified (Craigslist)
  - Webpages (Yahoo Directory)
- Filtering
  - News Feed
  - Spam

Main Uses


Top Topics
Number of Stories
400000
350000
300000
250000
200000
150000
100000
50000
0
CORPORATE/INDUSTRIAL
GOVERNMENT/SOCIAL
MARKETS
PERFORMANCE
ECONOMICS
COMMODITY MARKETS
ACCOUNTS/EARNINGS
COMMENT/FORECASTS
DOMESTIC POLITICS
MONEY MARKETS
OWNERSHIP CHANGES
EQUITY MARKETS
SOFT COMMODITIES
MERGERS/ACQUISITIONS
GOVERNMENT FINANCE
FUNDING/CAPITAL
MARKETS/MARKETING
INTERNATIONAL RELATIONS
REGULATION/POLICY
SPORTS
WAR, CIVIL WAR
CRIME, LAW ENFORCEMENT
CAPACITY/FACILITIES
Category

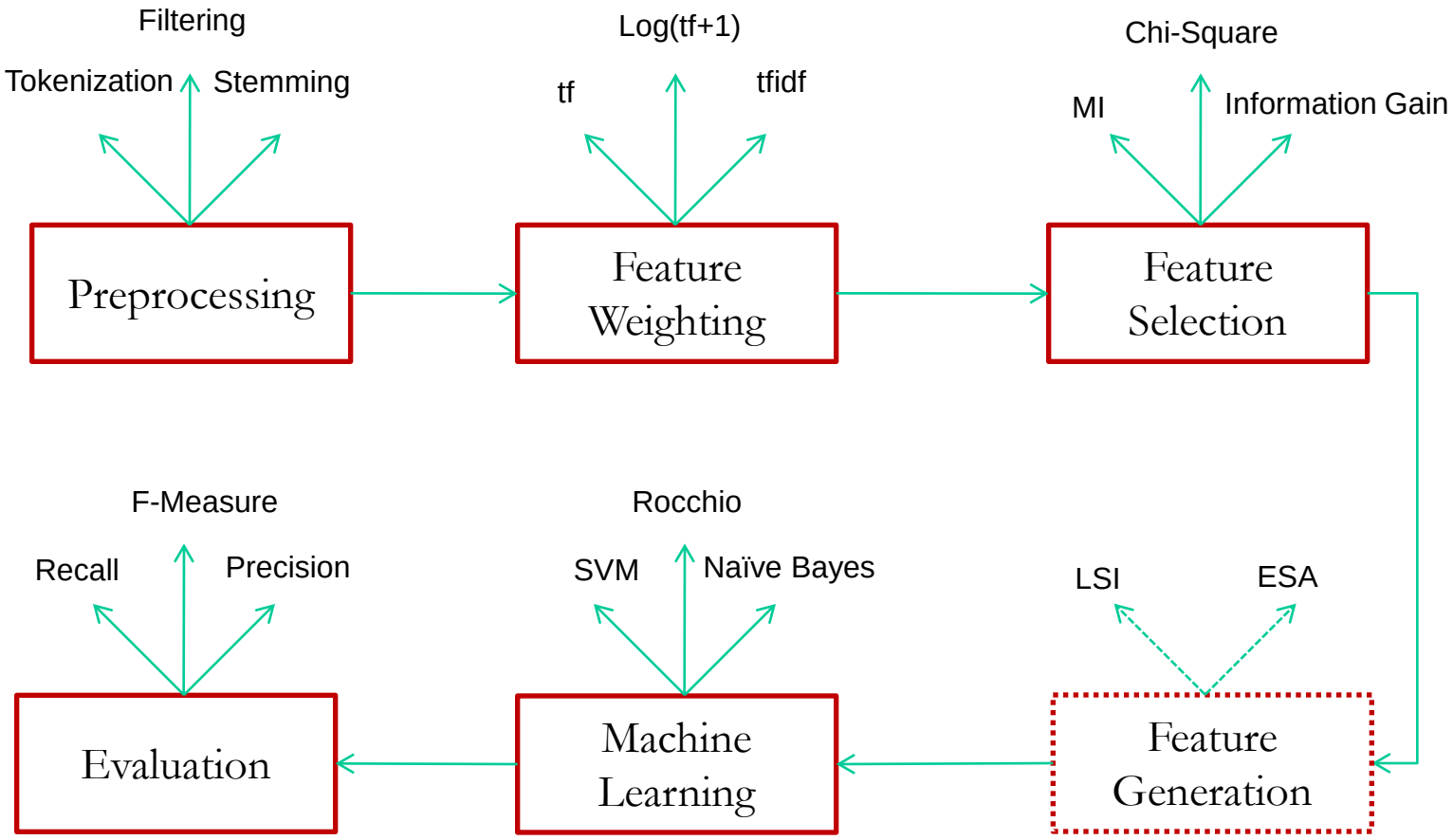
Other Uses

- Author Identification
- Genre Detection
- Language Identification
- Sentiment Classification
  - Market Analysis (Reuters)

Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

Anatomy of a Text Categorizer



Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

Classification types

- Document Membership
  - Single Label
  - Multiple Labels
  - Binary
- Hard vs Ranking Classifiers
  - Hard = Decisive!
  - Ranking = Probabilistic

Supervised vs Unsupervised

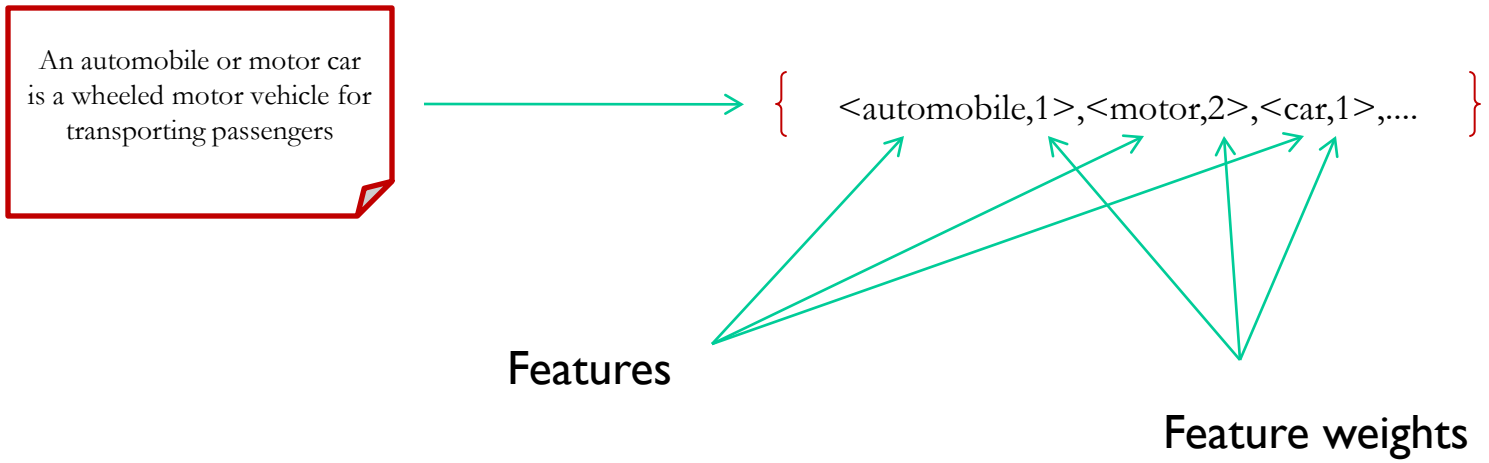
- Supervised Learning
  - Training classifier based on set of labeled documents
  - Training set vs Test set
- Unsupervised Learning
  - No labeled examples
  - The system tries to cluster documents based on some heuristics & distance measures

Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

Document Representation

- The idea is to process the natural language text in a document and transform into a vector



Document Representation

- Document Representation is a vector of term weights
- Each term represents specific information about the original document
- Terms are sometimes referred to as features
- Each term usually has an associated weight which represents its contribution to the document
- But..what is a term???

Terms

- Simplest approach: a term is a word
  - Bag of Words
- Preprocessing
  - Stopword removal (“a”,”the”,”of”,”and”)
  - Stemming (“stemming”,”stemmed”,”stemmer”)
- Ignore word order

Terms

- Sophisticated Approaches
- Higher Order statistics
- Phrases (how to define?)
  - Syntactically – according to grammer (Noun phrases)
  - Statistically – strongly occurring patterns of words

Weights

- Term frequency $tf(t_k)$

$$tf(t_k) = \sum_i \frac{n_k}{n_i}$$

- tf.idf

$$tf.idf(t_k) = tf(t_k).idf(t_k)$$

  - The more often the term appears in a document, the more the representative is it of the document.

$$idf(t_k) = \log(\frac{N_i}{N_k})$$

  - The more documents the term appears in the less discriminating it is.

- Normalized tfidf
  - Normalize the tf.idf values to the range[0,1]

$$w(t_k) = \frac{tf.idf(t_k)}{\sqrt{\sum_{s=1} (tf.idf(t_s))^2}}$$

Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

High Dimensionality

- There are many terms
- Many learning algorithms don't deal with extremely hight dimensions
- Over fitting problem
- Not all terms are equally effective
- Solution? Eliminate unwanted terms

Dimensionality Reduction

- Also known as feature Selection
- Idea: find a more efficient document representation, with much fewer dimensions, with a minimal loss of effectiveness (accuracy).

- Local vs Global Policies
  - Local Policy: For each category, find the best terms.
  - Global Policy: Given all the categories find the best terms.

Term Filtering

- A simple filtering can be done by ignoring rare terms
- Remove terms that occur in less that n documents
  - Experiments has shown a good performance
    - Dimensionality reduction factor of 10 without loss in accuracy.
    - Dimensionality reduction factor of 100 with small loss in accuracy.

Term Selection

- Out of original set of terms, t, find a much smaller subset, t', that yields high-test effectiveness(accuracy).
- Examples
  - Chi Square
  - Mutual Information
  - Information Gain
  - Information Ratio
  - Odd Ratio

Mutual Information

- Measure the association between to objects
  - It is a ratio of how many times the objects observed together normalized by the product of the occurrence of each object.

$$MI(A,B) = \frac{P(AB)}{P(A)P(B)}$$

Chi-Square

- The key idea of the chi-square test is a comparison of observed and expected values.

$$\chi^2 = \sum_{ij} \frac{(Observed_{ij} - Expected_{ij})^2}{Expected_{ij}}$$

Feature Generation

- Term Clustering
  - Unsupervised
  - Supervised
  - Distributional clustering

- Latent Semantic Indexing
- Explicit Semantic Indexing

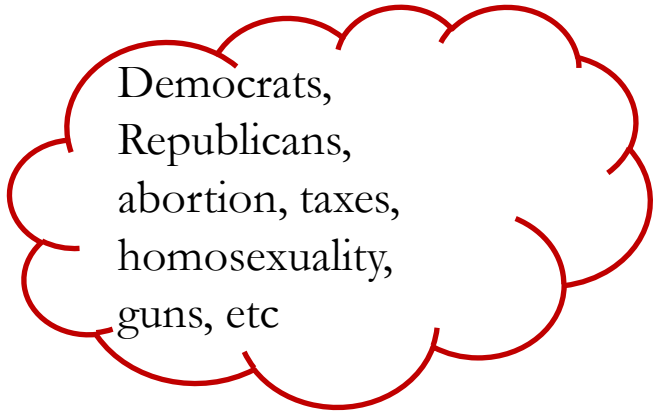
Latent Semantic Indexing (LSI)

- Words by themselves are not a good measure.
  - Synonyms (car, automobile)
  - Polysemous (Apple, Jaguar)
- LSI: a method for inferring the contextual similarity of terms
  - Finds the best m uncorrelated terms that best describe the original n terms.
  - Uncover latent information (synonyms)

Explicit Semantic Analysis

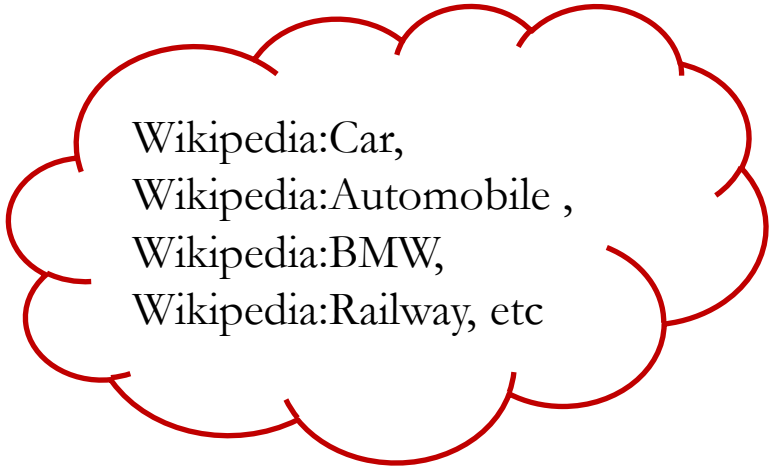
- Expand the terms using concept space (e.g. Wikipedia)
  - BOW

{ American politics } ⟶



  - ESA

{ Car } ⟶



Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

Types of Classifiers

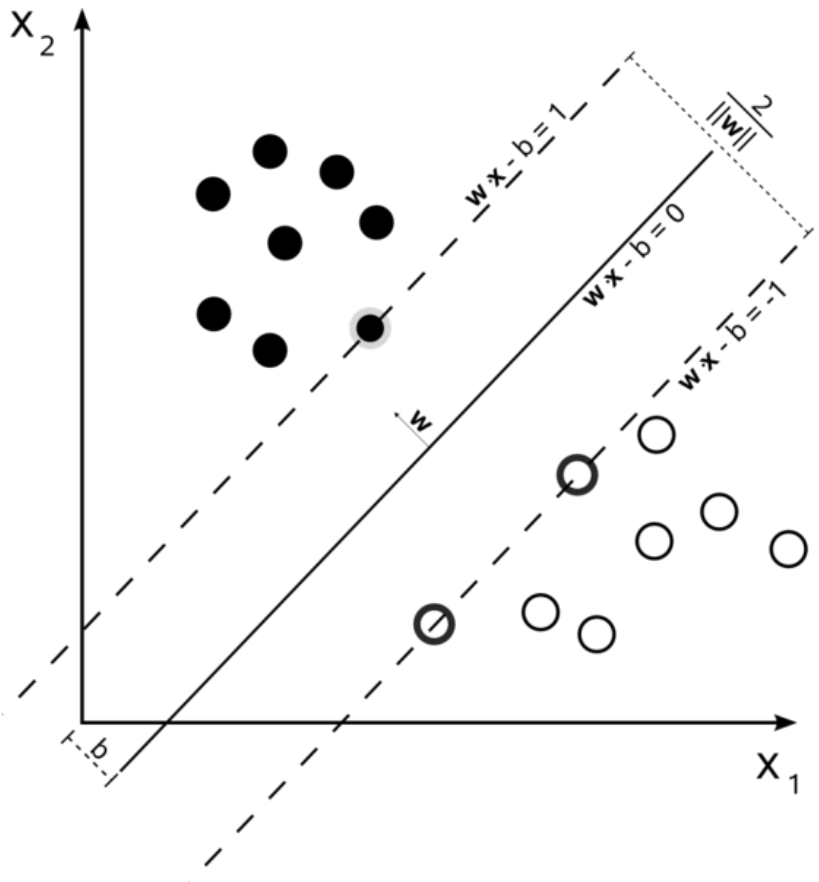
- Naïve Bayes
  - Calculate the $P(C_k \mid D)$, the probability that document D belong to the class C_k
  - By Bayes' theorem

$$P(C_k \mid D) = \frac{P(C_k)P(D \mid C_k)}{P(D)}$$

$$P(D \mid C_k) = \prod_i P(w_i \mid C_k)$$

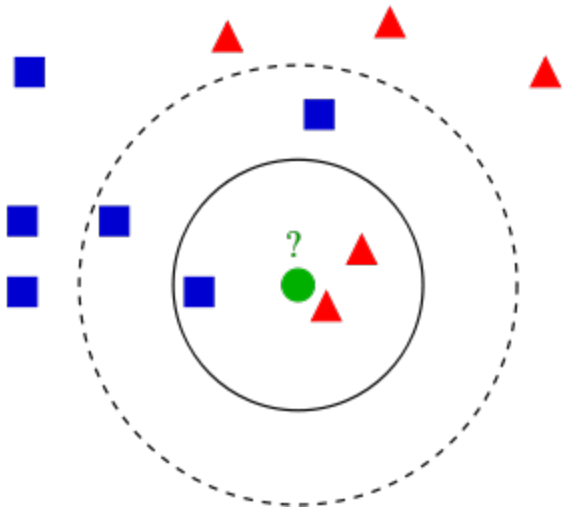
SVM

- Find the best hyper plan that separates the data points of two classes which a maximum separation (margin)



K Nearest Neighbor(K-NN)

- Document is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors
- To measure the distance two vectors
  - Euclidian Distance
  - Cosine Angle



Types of Classifiers

- Decision Trees
- Decision Rules
- Linear Least Square Fit
- Neural Networks
- Genetic Algorithms
- Committee/Ensembles

Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

Results & Evaluation

- How to measure the effectiveness of your classification?
- Precision
- Recall
- F-Measure
- Accuracy
- Micro/Macro Averaging
- Breakeven

Results & Evaluation

	Correct=Y	Correct=N
Assigned=Y	a	b
Assigned=N	c	d

- Accuracy = (a+d)/(a+b+c+d)
- Precision = a/(a+b)
- Recall= a/(a+c)
- F-Measure = 2*Precision*Recall/(Precision+Recall)
- Micro/Macro Averaging
- Breakeven (When Precision=Recall)

Outline

- Introduction
- Anatomy of a Text Categorizer
- Variant of the problem
- Document Representation
- Dimensionality Reduction
- Types of Classifiers
- Results & Evaluations
- Example

Example

Cats	Dogs
Cat ate your tongue?	My dog name is rocky.
I have a tuxedo cat.	Dogs are very affectionate.
Cats are relatives of tigers.	Dogs are descendents of wolves.

Preprocessing

Cats	Dogs
cat eat your tongue?	my dog name be rocky.
i have a tuxedo cat.	dog be very affectionate.
cat be relative of tiger.	dog be descendant of wolf.

Example

Cats	Dogs
cat eat your tongue?	my dog name is rocky.
i have a tuxedo cat.	dog be very affectionate.
cat be relative of tiger.	dog be descendant of wolf.

Cats	
Term	freq
cat	3
tuxedo	1
relative	1
tiger	1

Dogs	
Term	freq
dog	3
affectionate	1
rocky	1
wolf	1

Example

Cats		
Term	freq	P
cat	3	4/12
tuxedo	1	2/12
relative	1	2/12
tiger	1	2/12

Dogs		
Term	freq	P
dog	3	4/11
affectionate	1	2/11
rocky	1	2/11
wolf	1	2/11

The Toyger is an exciting new breed of domestic cats.

the toyger is an exciting new breed of domestic cat.

$$P(D \mid C_k) = \prod_i P(w_i \mid C_k)$$

$$P(C_k \mid D) \ \alpha \ P(C_k)P(D \mid C_k)$$

$$P(D \mid Cats) = P(toyger \mid Cats) * P(exciting \mid Cats) * P(breed \mid Cats) * P(cat \mid Cats)$$

$$P(D \mid Cats) = \frac{1}{12} * \frac{1}{12} * \frac{1}{12} * \frac{4}{12} = 1.9 * ^{-4}$$

$$P(Cats \mid D) \ \alpha \ P(Cats)P(D \mid Cats) = \frac{3}{6} * (1.9 * 10^{-4}) = 9.5 * 10^{-5}$$

$$P(D \mid Dogs) = P(toyger \mid Dogs) * P(exciting \mid Dogs) * P(breed \mid Dogs) * P(cat \mid Dogs)$$

$$P(D \mid Dogs) = \frac{1}{11} * \frac{1}{11} * \frac{1}{11} * \frac{1}{11} = 6.8 * 10^{-5}$$

$$P(Dogs \mid D) \ \alpha \ P(Dogs)P(D \mid Dogs) = \frac{3}{6} * (6.8 * 10^{-5}) = 3.4 * 10^{-5}$$

Questions