Learning Text Classification Rules (Intro)

Thorsten Joachims / Claire Cardie

Learning Text Classification Rules

- **Examples**
  - In-class exercise
- **Classification methods**
  - Supervised machine learning
- **Evaluating text classifiers**
- **Other issues**

Text Classification Example

Text Classification

- **Assign pieces of text to predefined categories based on content**
- **Types of text**
  - Documents (typical)
  - Paragraphs
  - Sentences
  - Websites
- **Different types of categories**
  - By topic
  - By function
  - By author
  - By style

Text Classification Examples

- **Assigning labels to documents or web-pages**
- **Labels are most often topics such as Yahoo-categories**
  - *"finance," "sports," "news>world>asia>business"*
- **Labels may be genres**
  - *"editorials" "movie-reviews" "news"*
- **Labels may be opinion on a person/product**
  - *"like", "hate", "neutral"*
- **Labels may be domain-specific**
  - *"interesting-to-me" : "not-interesting-to-me"*
  - *"contains adult language" : "doesn't"*
  - *language identification: English, French, Chinese, …*
  - *search vertical: about Linux versus not*
  - *"link spam" : "not link spam"*

Examples of Text Categorization?

Text categorization projects from previous years

- **Digital government**
  - Predict U.S. Supreme Court decisions based on court transcripts and amicus briefs
- **Movie reviews: + or -?**
  - Built upon existing techniques by also including information (features) about the director and actors.
- **Automatic categorization of iphone apps**
- **Build text categorization system**
  - Implement and evaluate a number of approaches
  - Extensive testing of many parameters of the system
  - Employed large, standard dataset (Reuters)

In-class Exercise

- **Reuters-21578 – the most famous text classification evaluation set**
  - **90 topics**
  - **~13,000 documents**
  - **Still widely used by lazy people (but now it's too small for realistic experiments – you should use Reuters RCV1)**

Example: Reuters Article (Multi-Label)

Categories: COFFEE, CRUDE

KENYAN ECONOMY FACES PROBLEMS, PRESIDENT SAYS

The Kenyan economy is heading for difficult times after a boom last year, and the country must tighten its belt to prevent the balance of payments swinging too far into deficit, President Daniel Arap Moi said.

In a speech at the state opening of parliament, Moi said high coffee prices and cheap oil in 1986 led to economic growth of five pct, compared with 4.1 pct in 1985. The same factors produced a two billion shilling balance of payments surplus and inflation fell to 5.6 pct from 10.7 pct in 1985, he added.

"But both these factors are no longer in our favour ... As a result, we cannot expect an increase in foreign exchange reserves during the year," he said.

…

acq	cotton-meal	groundnut	lumber	potato	soy-oil
alum	cotton-oil	groundnut-meal	lupin	propane	soybean
austdlr	cottonseed	groundnut-oil	meal-feed	rand	stg
austral	cpi	heat	mexpeso	rape-meal	strategic-metal
barley	cpu	hk	money-fx	rape-oil	sugar
bfr	crude	hog	money-supply	rapeseed	sun-meal
bop	cruzado	housing	naphtha	red-bean	sun-oil
can	dfl	income	nat-gas	reserves	sunseed
carcass	dkr	instal-debt	nickel	retail	tapioca
castor-meal	dlr	interest	nkr	rice	tea
castor-oil	dmk	inventories	nzdlr	ringgit	tin
castorseed	drachma	ipi	oat	rubber	trade
citruspulp	earn	iron-steel	oilseed	rupiah	tung
cocoa	escudo	jet	orange	rye	tung-oil
coconut	f-cattle	jobs	palladium	saudriyal	veg-oil
coconut-oil	ffr	l-cattle	palm-meal	sfr	wheat
coffee	fishmeal	lead	palm-oil	ship	wool
copper	flaxseed	lei	palmkernel	silk	wpi
copra-cake	fuel	lin-meal	peseta	silver	yen
corn	gas	lin-oil	pet-chem	singdlr	zinc
corn-oil	gnp	linseed	platinum	skr	
cornglutenfeed	gold	lit	plywood	sorghum	
cotton	grain	livestock	pork-belly	soy-meal	

Example: Ohsumed Abstract

Categories: Animal, Blood_Proteins/Metabolism, DNA/Drug_Effects, Mycotoxins/Toxicity, …

How aspartame prevents the toxicity of ochratoxin A.

Creppy EE, Baudrimont I, Anne-Marie

Toxicology Department, University of Bordeaux, France.

The ubiquitous mycotoxin ochratoxin A (OTA) is found as a frequent contaminant of a large variety of food and feed and beverage such as beer, coffee and win. It is produced as a secondary metabolite of moulds from Aspergillus and Penicillium genera. Ochratoxin A has been shown experimentally to inhibit protein synthesis by competition with phenylalanine its structural analogue and also to enhance oxygen reactive radicals production. The combination of these basic mechanisms with the unusual long plasma half-life time (35 days in non-human primates and in humans), the metabolisation of OTA into still active derivatives and glutathione conjugate both potentially reactive with cellular macromolecules including DNA could explain the multiple toxic effects, cytotoxicity, teratogenicity, genotoxicity, mutagenicity and carcinogenicity. A relation was first recognised between exposure to OTA in the Balkan geographical

Learning Text Classification Rules

- **Examples**
  - In-class exercise
- **Classification methods**
  - Supervised machine learning
- **Evaluating text classifiers**
- **Other issues**

Classification Methods (1)

- **Manual classification**
  - Used by the original Yahoo! Directory
  - Looksmart, about.com, PubMed
  - Very accurate when job is done by experts
  - Consistent when the problem size and team is small
  - Difficult and expensive to scale
    - Means we need automatic classification methods for big problems

Classification Methods (2)

- **Automatic document classification**
  - Hand-coded rule-based systems
    - One technique used by CS dept's spam filter, Reuters, CIA, etc.
    - It's what Google Alerts is doing
      - Widely deployed in government and enterprise
    - Companies provide "IDE" for writing such rules
    - E.g., assign category if document contains a given (boolean) combination of words
    - Standing queries: Commercial systems have complex query languages (everything in IR query languages +score accumulators)
    - Accuracy is often very high if a rule has been carefully refined over time by a subject expert
    - Building and maintaining these rules is expensive

A Verity topic
A complex classification rule

```
comment line      # Beginning of art topic definition
top-level topic   art ACCRUE
                      /author = "fsmith"
topic definition modifiers
                      /date   = "30-Dec-01"
                      /annotation = "Topic created
                                     by fsmith"
subtopic topic    * 0.70 performing-arts ACCRUE
  evidence topic  ** 0.50 WORD
    topic definition modifier
                      /wordtext = ballet
  evidence topic  ** 0.50 STEM
    topic definition modifier
                      /wordtext = dance
  evidence topic  ** 0.50 WORD
    topic definition modifier
                      /wordtext = opera
  evidence topic  ** 0.30 WORD
    topic definition modifier
                      /wordtext = symphony
subtopic          * 0.70 visual-arts ACCRUE
                  ** 0.50 WORD
                      /wordtext = painting
                  ** 0.50 WORD
                      /wordtext = sculpture
subtopic          * 0.70 film ACCRUE
                  ** 0.50 STEM
                      /wordtext = film
subtopic          ** 0.50 motion-picture PHRASE
                  *** 1.00 WORD
                      /wordtext = motion
                  *** 1.00 WORD
                      /wordtext = picture
                  ** 0.50 STEM
                      /wordtext = movie
subtopic          * 0.50 video ACCRUE
                  ** 0.50 STEM
                      /wordtext = video
                  ** 0.50 STEM
                      /wordtext = vcr
                  # End of art topic
```

Note:

- maintenance issues (author, etc.)
- hand-weighting of terms

Why <u>learn</u> text classifiers

Classifying documents by hand is costly and does not scale well

- e.g. browse all WWW pages to filter out those about job announcements

People are not really all that good at constructing text classification rules

- It is hard to write good queries

Sometimes there is no expert available

- e.g. rules for routing email

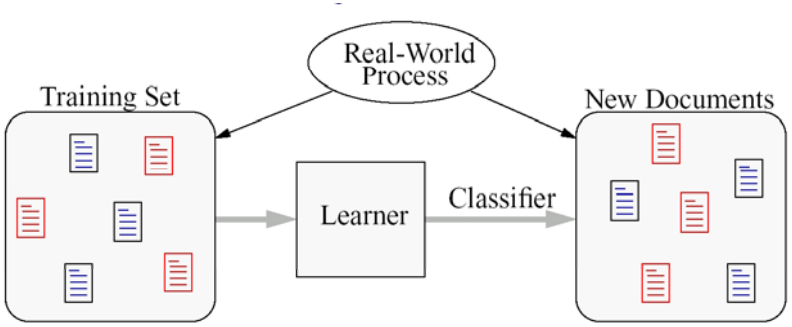
Often training data is cheap and plenty

- e.g. clickthrough from users

Classification Methods (3)

- **Supervised inductive learning of a document-label assignment function**
  - Many systems partly rely on machine learning (Autonomy, Microsoft, Yahoo!, …)
    - k-Nearest Neighbors (simple, powerful)
    - Naive Bayes (simple, common method)
    - Support-vector machines (new, more powerful)
    - … plus many other methods
    - No free lunch: requires hand-classified training data
    - But data can be built up (and refined) by amateurs
- **Many commercial systems use a mixture of methods**

Learning Setting



Goal:

- Learner uses training set to find classifier with low prediction error.

Project Info: Text Categorization

- **Select the task**
- **Map the task onto the text categorization definition**
  - How many classes/categories?
- **ML approach**
  - Find (or create --- ugh!) the appropriate data set / test collection
  - Text representation
  - Choose appropriate learning algorithm
  - Determine the evaluation/performance measures
  - Determine reasonable baselines to compare to

Categorization/Classification

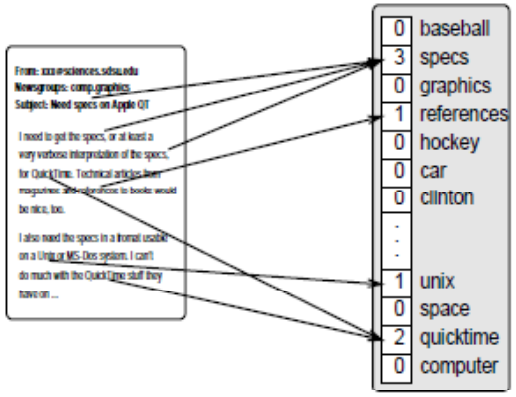
- **Given:**
  - A description of an instance, $d \in X$
    - X is the *instance language* or *instance space*.
      - Issue: how to represent text documents.
      - Usually some type of high-dimensional space
  - A fixed set of classes:
    $C = \{c_1, c_2, \ldots, c_J\}$
- **Determine:**
  - The category of d: $h(d) \in C$, where $h(d)$ is a *classification function* whose domain is X and whose range is C.
    - We want to know how to build classification functions ("classifiers").

Supervised Inductive Classification

- **Given:**
  - A description of an instance, $d \in X$
    - X is the *instance language* or *instance space*.
  - A fixed set of classes:
    $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set D of labeled documents with each labeled document $\langle d,c\rangle \in X \times C$
- **Determine:**
  - A learning method or algorithm which will enable us to learn a classifier $h: X \rightarrow C$
  - For a test document d, we assign it the class $h(d) \in C$

Representing Text for ML Approaches

- Attribute-value representation



Representing Text for ML Approaches

- Attribute-value representation
  - Each distinct word w_i corresponds to a feature (attribute)
  - Value is $TF(w_i, x)$
    - # of times word w_i occurs in document x
- Scale the values of the feature vector with their *inverse document frequency* $IDF(\mathbf{w_i})$
  - $DF(w_i)$ is the # of documents w_i appears in
  - $IDF(w_i) = log\ (n / DF(w_i))$
  - n is the total # of documents
- Each document *vector* is normalized to unit length

Learning Setting

Process:
- Generator: Generate documents/snippets according to distribution $P(X)$.
- Teacher: Assigns a value to each document based on $P(Y|X)$.

Training Examples $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \sim P(X, Y)$

Goal:
- Find a classification rule h with low prediction error on new examples from distribution $P(X,Y)$

$$Err_P(h) = P(h(\vec{x}) \neq y) = \int \Delta(h(\vec{x}), y) P(\vec{x}, y) dx dy$$

Test Collections

- **Reuters-21578**
  - Reuters newswire articles classified by topic
  - 90 categories (multi-label)
  - 9603 training documents / 3299 test documents (ModApte)
  - ~27,000 features
- **WebKB Collection**
  - WWW pages classified by function (e.g. personal HP, project HP)
  - 4 categories (multi-class)
  - 4183 training documents / 226 test documents
  - ~38,000 features
- **Ohsumed MeSH**
  - Medical abstracts classified by subject heading
  - 20 categories from "disease" subtree (multi-label)
  - 10,000 training documents/ 10,000 test documents
  - ~38,000 features

Assumptions of Naïve Bayes

- **Words occur independently given the class according to one multinomial distribution per class**
- **Each document is in exactly one class**
- **Word probabilities do not depend on the document length**

Pros and Cons for Naïve Bayes

- **Pros:**
  - Explicit theoretical foundation
  - Relatively effective
  - Very simple
  - Fast in learning and classification
- **Cons:**
  - Multinomial model / independence assumption clearly wrong for text
  - On some datasets it really fails badly

Learning Text Classification Rules

- **Examples**
  - In-class exercise
- **Classification methods**
  - Supervised machine learning
- **Evaluating text classifiers**
- **Other issues**

ML no-no's

- **Train on dataset A, test on dataset A**
- **Estimating any parameters based on the test set**
- **Evaluating w.r.t. (or looking at) the test set repeatedly while developing your approach**
- **Evaluation methodology is key**

Evaluating Categorization

- **Evaluation must be done on test data that are independent of the training data (usually a disjoint set of instances).**
  - Sometimes use cross-validation (averaging results over multiple training and test splits of the overall data)
- **It's easy to get good performance on a test set that was available to the learner during training (e.g., just memorize the test set).**

Performance Measures

- **Precision/Recall /F/Break-Even Point**
  - Intersection of PR-curve with the identity line
- **Accuracy**
  - Adequate if one class per document
  - Otherwise F measure for each class
- **Macro-averaging**
  - First compute the measure, then compute average
  - Means average over tasks
- **Micro-averaging**
  - Compute the measure for each decision, then compute average
  - Means average over each individual classification decision

Go to Sebastiani results

Learning Text Classification Rules

- **Examples**
  - In-class exercise
- **Classification methods**
  - Supervised machine learning
- **Evaluating text classifiers**
- **Other issues**

Multi-Class / Multi-Label

- **Cannot learn multi-label rules directly**
  - Most classifiers assume that each document is in exactly one class
  - Many classifiers can only learn binary classification rules
- **Most common solution: Multi-Class**
  - Learn one binary classifier for each label
  - Put example into the class with the highest probability (or some approximation thereof)
- **Most common solution: Multi-Label**
  - Learn one binary classifier for each label
  - Attach all labels, for which some classifier says positive

Feature Selection: Why?

- **Text collections have a large number of features**
  - 10,000 – 1,000,000 unique words … and more
- **May make using a particular classifier feasible**
  - Some classifiers can't deal with 100,000 of features
- **Reduces training time**
  - Training time for some methods is quadratic or worse in the number of features
- **Can improve generalization (performance)**
  - Eliminates noise features
  - Avoids overfitting

Feature selection: How?

- **Forward/backward selection**
- **Hypothesis testing statistics:**
  - Are we confident that the value of one categorical variable is associated with the value of another
    - Chi-square test (χ^2)
    - Statistical foundation
    - May select very slightly informative frequent terms that are not very useful for classification
- **Use another learning algorithm**
  - E.g. decision tree induction (Cardie, ICML 1994)
- **Just use the commonest terms?**
  - No particular foundation
  - In practice, this is often 90% as good