Today’s topic

- Automatic document classification
  - Rule-based classification
  - Supervised learning

Classification

- Given one or more topics, decide which one(s) a given document belongs to.
- Applications
  - Classification into a topic taxonomy
  - Intelligence analysts
  - Routing email to help desks/customer service

Step back

- Manual classification
  - accurate when done by subject experts
  - consistent when done by a small team
  - difficult to scale
  - used by Yahoo!, Looksmart, about.com, ODP
    - hundreds of subject editors maintain thousands of topics
    - (topics organized in a tree-like navigation structure)
Supervised vs. unsupervised learning

- Unsupervised learning:
  - Given corpus, infer structure implicit in the docs, without prior training.
- Supervised learning:
  - Train system to recognize docs of a certain type (e.g., docs in Italian, or docs about religion)
  - Decide whether or not new docs belong to the class(es) trained on

Challenges

- Must teach machine a model of each topic
- Given new doc, must measure fit to model(s)
- Evaluation: how well does the system perform?
- Threshold of pain: how confident is the system’s assessment?
  - Sometimes better to give up.

Teaching the system models

- Through an explicit query
- Through exemplary docs
- Combination

Explicit queries

- For the topic “Performing Arts”, query accrues evidence from stemmed and non-stemmed words.
- Query built by a subject expert.
- New doc scored against query:
  - if accrued evidence exceeds some threshold, declare it to belong to this topic.
Explicit queries

Topic queries can be built up from other topic queries.

Large scale applications

- Document routing
- Customer service
- Profiled newsfeeds
- Spam/porn filtering

Typical example

- Dow Jones
  - Over 100,000 standing profiles
  - A profile can have >100 atomic terms
  - Common sub-expressions shared by different topics
  - Optimizing this sharing is a hard problem.

Example of sharing
Example of sharing

Costa Rica Eco-Tourism

AND

Costa Rica Eco-Tourism Hotels

Hotel

AND

Lake Tahoe area Hotels

room

OR

suite

OR

Tahoe

Donner

Boreal

Measuring classification

- Figures of merit include:
  - Accuracy of classification (more below)
  - Speed of classification (docs/hour)
  - Effort in training system (human hours/topic)

Factors affecting measures

- Documents
  - size, length
  - quality/style of authorship
    - uniformity of vocabulary
- Accuracy measurement
  - need definitive judgement on which topic(s) a doc belongs to
    - usually human

Accuracy measurement

- Confusion matrix

This $(i,j)$ entry means 53 of the docs actually in topic $i$ were put in topic $j$ by the classifier.
Confusion matrix

- Function of classifier, topics and test docs.
- For a perfect classifier, all off-diagonal entries should be zero.

Confusion measures

- Fraction of docs in topic $i$ classified correctly:

\[ \frac{c_{ii}}{\sum_j c_{ij}} \]

- Fraction of docs assigned topic $i$ that are actually about topic $i$:

\[ \frac{c_{ii}}{\sum_i c_{ii}} \]

- Fraction of docs classified correctly:

\[ \frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}} \]

Classification by exemplary docs

- Feed system exemplary docs on topic *(training)*
- Positive as well as negative examples
- System builds its model of topic
- Subsequent *test* docs evaluated against model
  - decides whether test is a member of the topic

More generally, set of topics

- Exemplary docs for each
- Build model for each topic
  - differential models
- Given test doc, decide which topic(s) it belongs to
Recall doc as vector

- Each doc is a vector, one component for each term.
- Normalize to unit length.
- Have a vector space
  - terms are axes
  - docs live in this space
  - even with stemming, may have 10000+ dimensions

Classification using vector spaces

- Each training doc a point (vector) labeled by its topic
- Hypothesis: docs of the same topic form a contiguous region of space
- Define surfaces to delineate topics in space

Topics in a vector space

Given a test doc

- Figure out which region it lies in
- Assign corresponding topic
Issues

- How do we define (and find) the separating surfaces?
- How do we compose separating surfaces into regions?
- How do we test which region a test doc is in?

Separation by hyperplanes

- Assume linear separability for now:
  - in 2 dimensions, can separate by a line
  - in higher dimensions, need hyperplanes.
- Can find separating hyperplane by linear programming:
  - separator can be expressed as $ax + by = c$;
Linear programming

Find $a, b, c$, such that

$$ax + by \geq c$$
for red points

$$ax + by \leq c$$
for green points.

Which hyperplane?

In general, lots of possible solutions for $a, b, c$.

Support Vector Machine (SVM)

• Quadratic programming problem
• The decision function is fully specified by training samples which lie on two parallel hyper-planes

Support vectors

Maximize margin

Building an SVM classifier

• Now we know how to build a separator for two linearly separable topics
• What about topics whose exemplary docs are not linearly separable?
• What about >2 topics?
**Not linearly separable**

Find a line that penalizes points on “the wrong side”.

**Exercise**

- Suppose you have $n$ points in $d$ dimensions, labeled red or green. How big need $n$ be (as a function of $d$) in order to create an example with the red and green points not linearly separable?
- E.g., for $d=2$, $n \geq 4$.

**Penalizing bad points**

Define distance for each point with respect to separator $ax + by = c$:
- $(ax + by) - c$ for red points
- $c - (ax + by)$ for green points.

**Solve quadratic program**

- Solution gives “separator” between two topics: choice of $a, b$.
- Given a new point $(x, y)$, can score its proximity to each class:
  - evaluate $ax+by$.
  - Set confidence threshold.
**Category: Interest**

- Example SVM features
  - $w_i \cdot t_i$
    - 0.70 prime
    - 0.67 rate
    - 0.63 interest
    - 0.60 rates
    - 0.46 discount
    - 0.43 bundesbank
    - 0.43 baker
  - $w_i \cdot t_i$
    - -0.71 dhrs
    - -0.35 world
    - -0.33 sees
    - -0.25 year
    - -0.24 group
    - -0.24 dlr
    - -0.24 january

**Separating multiple topics**

- Build a separator between each topic and its complementary set (docs from all other topics).
- Given test doc, evaluate it for membership in each topic.
- Declare membership in topics above threshold.

**Negative examples**

- Formulate as above, except negative examples for a topic are added to its complementary set.

**Recall explicit queries**

- Can be viewed as defining a region in vector space.
- No longer linear separators.
**Simple example**

- Query = *Costa Rica AND hotel*

**Challenge**

- Combining rule-based and machine learning based classifiers.
  - Nonlinear decision surfaces vs. linear.
  - User interface and expressibility issues.

**UI issues**

- Can specify rule-based query in the interface.
- Can exemplify docs.
- What is the representation of the combination?

**Classification - closing remarks**

- Can also use Bayesian nets to formulate classification
  - Compute probability doc belongs to a class, conditioned on its contents
- Many fancy schemes exist for term weighting in vectors, beyond simple $tf\times idf$. 
Resources