Of text analysis.

Tik-122.101 Special Course in Information Technology: Modelling and mining the Web
Modeling the Internet and the Web: Probabilistic Methods and Algorithms.
Chapter 4.
Wednesday 17.9.2003

Nikolaj Tatti
ntatti@cc.hut.fi

Extracting raw text from a document.

- A document is usually in some format, e.g. HTML, PDF or PS, so it need to be converted into raw text.
- HTML $\Longrightarrow$ TEXT is quite easy, though the parser should be robust.
- http://search.cpan.org/dist/HTML-Parser
- PDF $\Longrightarrow$ TEXT or PS $\Longrightarrow$ TEXT is painful because PDF/PS are more like programming languages.

Stemming

- Similar words should be grouped together: e.g. fishing, fishes, fisher.
- Stemming is required: Words are changed/trunked. For example, $\{fishing, fishes, fisher\} \Longrightarrow fish$.
- Porter stemmer (developed in 1980) available http://www.tartarus.org/ martin/PorterStemmer/.

Bag-of-words

- Document is represented as a vector in R^k such that each item in this vector represents some (a priori chosen) word.
- The word order of the original document is destroyed (so what).
- The simplest representation is that 1 means that the word is present in a document and 0 that the word is absent.

Feature selection

- The dimension of the vectors representing documents are typically very high.
- One way to reduce dimension is to use information theory.
- Assume that we have classified documents into K classes.
- Let $P(c)$ be the (prior) probability of the class.
- Let $P(w_j)$ be the probability for some word.
- Let $P(c, w_j)$ be the mutual probability for some word and some class.

..continues

- Calculate information gain

$$\sum_{c=1}^{K} P(c, w_j = 1) \log \frac{P(c, w_j = 1)}{P(c)P(w_j = 1)} + \sum_{c=1}^{K} P(c, w_j = 0) \log \frac{P(c, w_j = 0)}{P(c)P(w_j = 0)}$$

- Choose n words which have the highest gain.

More of feature selection

- Previous technique examines items individually.
- However, items that seem to have little information may have high information gain when considered together (XOR-problem).
- Framework proposed by Koller and Sahami:
- Let x be the original vector and let x_G be the vector which is projected into some subvocalbulary G.
- Calculate $\Delta_G = \sum_x P(x)P(c \mid x) \log \frac{P(c|x)}{P(c|x_G)}$. The smaller the difference, the better.

Text categorisation - kNN

- Let X be the labelled data and let x be vector to be labelled.
- Pick k vectors $N \subset X$ and assign x to the label which has majority in these vectors.
- Alternatively, assign score for each class

$$s_c(x) = \sum_{y \in N, y \text{ belongs to } c} \cos(x, y).$$

- Pick a class with the highest score.

Text categorisation - SVM

- Assume that there is labelled data set with two classes.
- Original data points are mapped to some high-dimensional space.
- A hyper-plane is drawn in this space such that the data is divided correctly into two classes. It is assumed that such plane exists.
- Extra condition: The minimum distance between the data points and the hyper-plane (a.k.a. margin) should as large as possible.
- If the mapping is nice, that is, if we can calculate the inner product in high-dimension space fast, then the learning algorithm is fast.

Text categorisation - Naive Bayes classifier

- Assume that given some class c the word probabilities are mutually independent $P(w_1 \cdots w_k \mid c) = \prod P(w_i \mid c)$.
- We can easily estimate $P(w_i \mid c)$ from the training data.
- Given some document d calculate

$$P(c \mid d) \propto P(c)P(d \mid c) = P(c) \prod P(w_i \mid c).$$

- Use this as a score to determine the class of the document.

Measuring the performance

- Assume n categories. A confusion matrix C of size nxn is defined C_{ij} = ”number of those documents that were classified to category i, but actually belong to category j”.
- Precision for category i: $\pi = C_{ii} / \sum_j C_{ij}$.
- Recall for category i: $\rho = C_{ii} / \sum_j C_{ji}$.
- Usually a trade-off can be made between these two measures using some algorithm-specific parameter.
- A breakeven point is the best configuration of the algorithm such that $\pi = \rho$.

Information extraction

- We want to extract automatically information from raw text.
- For example, authors and books of some online bookstore.
- Hand-built models that are based on state machines.
- Sliding window: pick a text chunk with k words in some position and try to categorise it.

..continues

- We can use HMM techniques.
- Different states represent different fields in document e.g. title, author, email...
- The idea is that, for example, the address of the author is near to the name of the author.

Applications - Reuters

- Reuters-21578 is perhaps most widely used benchmark for text categorisation.
- 21 578 manually labelled articles. 672 categories. 148 of them has more than 20 documents.

Prediction model	Performance
Naive Bayes	73.4%
kNN	82.0%
Support vector (RBF)	86.3%

Applications - Email&News filtering

- In spam word proximity is important, so bag-of-words representation may lose important information.
- Better features can be obtained if small chunks of text are used. For example, 'confidential message'.
- Sahami et al. manually identified these features.
- NewsWeeder is personalised filter for Usenet articles.
- The training data set (19 997 articles, 20 groups) used is available at http://www.ai.mit.edu/people/jrennie/20Newsgroups/.