Document Clustering using Information Bottleneck with Annealing

Lisa Miller
Machine Learning
Dr. S. Still
Spring 2010
Overview

• Project description
  – Information Bottleneck Method
  – Document Set

• Preprocessing
  – Parsing
    – Dimensionality
  – Word contribution to MI
    – Dimensionality reductions

• Progress on clustering
Project

- Want to classify a large set of unlabeled documents for my astrobiology work
  - No labels => must do unsupervised clustering
- Investigate Information Bottleneck method
  - Implement Dr. Still's IB with annealing method for large data set
  - Choose a standard dataset in order to evaluate performance
  - Compare output with other popular methods
Information Bottleneck Method

- Implemented as an E-M algorithm by Dr. Still
  - E-Step
    - $p(c|x) \sim e^{-\frac{1}{T}(E(x,c) - f(x)) - \beta D_{KL}[p(y|x) \| p(y|c)]}$
  - M-Step
    - $p(y|c) = \sum_x p(y|x) p(c|x) \frac{p(x)}{p(c)}$

- Until reach convergence criteria
- Adjust beta and split clusters
- Run EM again until all N clusters filled and some beta limit reached or assignments are deterministic
Data Set

- Reuters' 21578 Modified Apte split
  - Developed for supervised learning
  - Training set (9,603 docs)
  - Test set (3,299 docs)
  - 12,902 documents total
  - 135 categories with economic topics
    - Earnings
    - Grain
    - Crude
    - Copper
    - Etc.
Preprocessing

- Documents are in 22 files
  - 1000 docs each
  - SGML markup format
- Removed all punctuation and extra spaces
- Not using numbers
Dimensionality

- Full data set (12,902 documents)
  - 32,012 unique words
- Implemented per-word contribution to mutual information to determine how to reduce dimensionality
  
  \[ I(y) = p(y) \sum_{x \in X} p(x|y) \log \left( \frac{p(x|y)}{p(x)} \right) \]
  
  - Calculating requires calculation of \( p(y|x), p(x;y), p(x|y) \) on the full set.
    
    - Data is sparse
    - Hash tables work great!
Word Contribution to Mutual Information

- Cumulative Information vs word count

\[ I(S) = \sum_{y \in S} i(y) = \sum_{y \in S} p(y) \sum_{x \in X} p(x|y) \log \left( \frac{p(x|y)}{p(x)} \right) \]

90% Info

< 75% Info

2000 words

7,941 Words

Slonim & Tishby
Dimensionality Reduction

- Had hoped curve would be less smooth.
- Will doing Taylor expansion to estimate sampling error give an intuitive cutoff?
  - Dimensionality and high word occurrences
    - 1,564,726 total words in set
  - Sampling error estimated to be negligible
- Can keep 90% mutual information in ~25% of the words
- Dimensionality is still high with 7000+ words
New questions to answer

- Can we run the algorithm in a reasonable time with 7000 words?
- Can we get the same results keeping less?
  - Slonim & Tishby's results were quite good with 2000
Progress and work to do

- Rewrite algorithm code using hash tables for the two largest matrices
  - p(y|x)
  - p(x,y)
  - Almost finished
- Attempt to run with 90% information kept
- If success: compare results with keeping various percentages of information
- Compare results with other methods.
- Try other data sets, esp. 20 newsgroups
- ...
References

