CS 294-5: Statistical Natural Language Processing

Unsupervised Learning I

Dan Klein

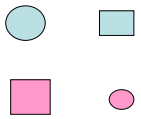
Supervised Learning

- Systems duplicate correct analyses from training data
- Every system you deploy in the real world requires:
  - A learning algorithm (but they all work about the same)
  - A labeled data set (requires a lot of work – more than you think)
  - Feature engineering (the key to good practical systems in NLP)
- Data sets are usually the bottleneck for most tasks

Unsupervised Learning

- Systems take raw data and automatically detect patterns
- Unsupervised systems need:
  - A pattern detection method (not all the same, not well understood)
  - An unlabeled data set (these are always around – if you have a task, you have a data set)
  - Feature engineering?
- Big drawback: unsupervised systems don't generally work as well (if at all!)

Clustering / Pattern Detection



LONDON -- Soccer team wins match

NEW YORK – Stocks close up 3%

Investing in the stock market has …

The first game of the world series …

- Problem 1: There are many patterns in the data, most of which you don't care about.

Model-Based Clustering

- Clustering with probabilistic models:

Unobserved (Y)	Observed (X)
c1	LONDON -- Soccer team wins match
c2	NEW YORK – Stocks close up 3%
c3	Investing in the stock market has …
c4	The first game of the world series …

Find Y and θ to maximize $P(X,Y|\theta)$

- Problem 2: The relationship between the structure of your model and the kinds of patterns it will detect is complex.

Clustering vs. Classification

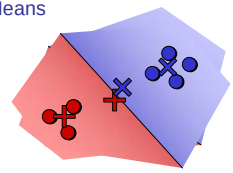
- Classification: we specify which pattern we want, features uncorrelated with that pattern are idle

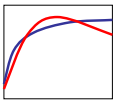
P(w|sports)	P(w|politics)	P(w|headline)	P(w|story)
the 0.1	the 0.1	the 0.05	the 0.1
game 0.02	game 0.005	game 0.01	game 0.01
win 0.02	win 0.01	win 0.01	win 0.01

- Clustering: the clustering procedure locks on to whichever pattern is most salient
  - P(content words | class) will learn topics
  - P(length, function words | class) will learn style
  - P(characters | class) will learn "language"

Learning Models with EM

- Alternate between
  E-step: Find Y to maximize $P(X,Y|\theta)$ for fixed θ
  M-step: Find θ to maximize $P(X,Y|\theta)$ for fixed Y
- Example: K-Means [Hard EM]





- Problem 3: Data likelihood (usually) isn't the objective you really care about
- Problem 4: You can't find global maxima anyway

Heuristic Clustering?

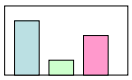
- Many methods of clustering have been developed
  - Most start with a pairwise distance function
  - Most can be interpreted probabilistically (with some effort)
  - Axes: flat / hierarchical, agglomerative / divisive, incremental / iterative, probabilistic / graph theoretic / linear algebraic
- Examples:
  - Single-link agglomerative clustering
  - Complete-link agglomerative clustering
  - Ward's method
  - Hybrid divisive / agglomerative schemes

Document Clustering

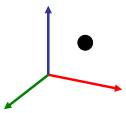
- Typically want to cluster documents by topic
- Bag of words models usually do detect topic
  - It's detecting deeper structure, syntax, etc. where it gets really tricky!
- All kinds of games to focus the clustering
  - Stopword lists
  - Term weighting schemes (from IR, more later)
  - Dimensionality reduction (more later)

Two Views of Documents

- Probabilistic
  - A document is a collection of words sampled from some distribution, an empirical distribution
  - Correlations between words flows through hidden model structure
  - Distance: divergences

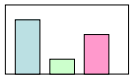


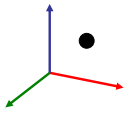
- Vector Space
  - A document is a point in a high-dimensional vector space
  - Correlations between words reflects low rank of valid document subspace
  - Distance: Euclidean / cosine



High Dimensional Data

- Both of these pictures are absolutely misleading!
  - Documents are zero in almost all axes
  - Most document pairs are very far apart (i.e. not strictly orthogonal, but only share very common words and a few scattered others)
  - In classification terms: virtually all document sets are separable, for most any classification





Dimensionality Reduction

- Most document clustering posits that some small number of axes / variables account for all that variation
- Probabilistic statement:
- Vector-space statement:
- Low rank representations used for IR, more later.

Semi-Supervised Learning

- A middle ground: Semi-supervised methods
  - Use a small labeled training set and a large unlabeled extension set
  - Use labeled data to lock onto the desired patterns
  - Use unlabeled data to flesh out model parameters
- Broad approaches
  - Constrained clustering
  - Self-training
  - Adaptation / anchoring
- Also: active learning

Incorporating Supervision

What's Next?

- Next class:
  - Learning parts of speech
  - Interaction between models and patterns
- Section on Wednesday
  - Q+A on HW4 techniques, if you need it
- Readings:
  - M+S Chapter 14