Overview

- So far: language models give $P(s)$
  - Help model fluency for various noisy-channel processes (MT, ASR, etc.)
  - N-gram models don’t represent any deep variables involved in language structure or meaning
  - Usually we want to know something about the input other than how likely it is (syntax, semantics, topic, etc)

- Next: Naïve-Bayes models
  - We introduce a single new global variable
  - Still a very simplistic model family
  - Lets us model hidden properties of text, but only very non-local ones...
  - In particular, we can only model properties which are largely invariant to word order (like topic)

Text Categorization

- Want to classify documents into broad semantic topics (e.g. politics, sports, etc.)
  - Democratic vice presidential candidate John Edwards on Sunday accused President Bush and Vice President Dick Cheney of misleading Americans by implying a link between deposed Iraqi President Saddam Hussein and the Sept. 11, 2001 terrorist attacks.
  - Which one is the politics document? (And how much deep processing did that decision take?)
  - One approach: bag-of-words and Naïve-Bayes models
  - Another approach later...
  - Usually begin with a labeled corpus containing examples of each class

Naïve-Bayes Models

- Idea: pick a topic, then generate a document using a language model for that topic.
- Naïve-Bayes assumption: all words are independent given the topic.
- Compare to a unigram language model:
  $$P(c, w_1, w_2, ..., w_n) = P(c) \prod_i P(w_i | c)$$
  - We have to smooth these!

Two NB Formulations

- Two NB models for text categorization
  - The class-conditional unigram model, a.k.a. multinomial model
    - One node per word in the document
    - Driven by words which are present
    - Multiple occurrences, multiple evidence
    - Better overall – plus, know how to smooth
  - The binomial (binary) model
    - One node for each word in the vocabulary
    - Incorporates explicit negative correlations
    - Know how to do feature selection (e.g. keep words with high mutual information with the class variable)

Using NB for Classification

- We have a joint model of topics and documents
  $$P(c, w_1, w_2, ..., w_n) = P(c) \prod_i P(w_i | c)$$
- Gives posterior likelihood of topic given a document
  $$P(c | w_1, w_2, ..., w_n) = \frac{P(c) \prod_i P(w_i | c)}{\sum_c P(c') \prod_i P(w_i | c')}$$
  - What about totally unknown words?
  - Can work shockingly well for textcat (especially in the wild)
  - How can unigram models be so terrible for language modeling, but class-conditional unigram models work for textcat?
  - Numerical / speed issues
  - How about NB for spam detection?
Example: Barometers

**Reality**
- Raining: \[ P(+, r) = 3/8 \]
- Sunny: \[ P(-, s) = 3/8 \]

**NB Model**
- Raining? \( M_1 \)
- Sunny? \( M_2 \)

**NB FACTORS:**
- \( P(s) = 1/2 \)
- \( P(+|s) = 1/4 \)
- \( P(+|r) = 3/4 \)

**PREDICTIONS:**
- \( P(r,+,+) = (1/2)(3/4)(3/4) \)
- \( P(s,+,+) = (1/2)(1/4)(1/4) \)
- \( P(r|+,+) = 9/10 \)
- \( P(s|+,+) = 1/10 \)

Overconfidence!

Example: Stoplights

**Reality**
- Lights Working: \[ P(g, r, w) = 3/7 \]
- Lights Broken: \[ P(r, r, b) = 1/7 \]

**NB Model**
- Working? \( NS \)
- Not Working? \( EW \)

**NB FACTORS:**
- \( P(w) = 6/7 \)
- \( P(r|w) = 1/2 \)
- \( P(g|w) = 1/2 \)
- \( P(b) = 1/7 \)
- \( P(r|b) = 1 \)
- \( P(g|b) = 0 \)

**PREDICTIONS:**
- \( P(b|r,r) = 4/10 \)

(Non-)Independence Issues

- **Mild Non-Independence**
  - Evidence all points in the right direction
  - Observations just not entirely independent
  - Results
    - Inflated Confidence
    - Deflated Priors
  - What to do? Boost priors or attenuate evidence
    \[ P(c, w_1, w_2, \ldots, w_n) = \prod \frac{P(c|w_i)}{P(c|\text{source})} \]

- **Severe Non-Independence**
  - Words viewed independently are misleading
  - Interactions have to be modeled
  - What to do? Change your model!

Language Identification

How can we tell what language a document is in?

The 38th Parliament will meet on Monday, October 4, 2004, at 11:00 a.m.
The first item of business will be the election of the Speaker of the House of Commons. Her Excellency the Governor General will open the First Session of the 38th Parliament on October 5, 2004, with a Speech from the Throne.

How to tell the French from the English?
- Treat it as word-level textcat?
- Overkill, and requires a lot of training data
- You don't actually need to know about words!

Σύμφωνα σταθερότητας και ανάπτυξης

Patto di stabilità e di crescita

Class-Conditional LMs

- Can have a topic variable for other language models
  \[ P(c, w_1, w_2, \ldots, w_n) = P(c) \prod_i P(w_i | c) \]

- Could be characters instead of words, used for language ID (HW2)
- Could sum out the topic variable and use as a language model
- How might a class-conditional n-gram language model behave differently from a standard n-gram model?

Clustering / Pattern Detection

- Problem 1: There are many patterns in the data, most of which you don’t care about.
- Soccer team wins match
- Stocks close up 3%
- Investing in the stock market has …
- The first game of the world series …
Clustering vs. Classification

- **Classification**: we specify which pattern we want, features uncorrelated with that pattern are idle
  
  \[
  \begin{array}{cccc}
  \text{P(w|sports)} & \text{P(w|politics)} & \text{P(w|headline)} & \text{P(w|story)} \\
  0.1 & 0.1 & 0.05 & 0.1 \\
  \text{game} & 0.02 & \text{game} & 0.005 & \text{game} & 0.01 & \text{game} & 0.01 \\
  \text{win} & 0.02 & \text{win} & 0.01 & \text{win} & 0.01 & \text{win} & 0.01 \\
  \end{array}
  \]

- **Clustering**: the clustering procedure locks on to whichever pattern is most salient, statistically
  
  \[
  P(\text{content words} \mid \text{class}) \text{ will learn topics} \\
  P(\text{length, function words} \mid \text{class}) \text{ will learn style} \\
  P(\text{characters} \mid \text{class}) \text{ will learn "language"}
  \]

Model-Based Clustering

- **Clustering with probabilistic models**:

  \[
  \begin{array}{ccc}
  \text{Unobserved (Y)} & \text{Observed (X)} \\
  ?? & \text{LONDON – Soccer team wins match} \\
  ?? & \text{NEW YORK – Stocks close up 3%} \\
  ?? & \text{Investing in the stock market has ...} \\
  ?? & \text{The first game of the world series ...} \\
  \end{array}
  \]

  Build a model of the domain: \( P(x, y, \theta) \)

  Often: find \( \theta \) to maximize:
  
  \[
  P(x \mid y, \theta) = \sum_y P(x, y \mid \theta)
  \]

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

Learning Models with EM

- **Hard EM**: alternate between
  
  E-step: Find best "completions" \( Y \) for fixed \( \theta \)
  
  M-step: Find best parameters \( \theta \) for fixed \( Y \)

  **Example**: K-Means

  **Problem 3**: Data likelihood (usually) isn’t the objective you really care about

  **Problem 4**: You can’t find global maxima anyway

EM for Naïve-Bayes

- **First we calculate posteriors (completions)**:

  \[
  P(y \mid x) = \frac{P(y) \prod_i P(x_i \mid y)}{\sum_{y'} P(y') \prod_i P(x_i \mid y')}
  \]

- Then we re-estimate parameters \( P(y) \), \( P(x_i \mid y) \) from the (fractionally) labeled data:

  \[
  z(u, y) = \sum_{(x, y) \in D} P(y \mid x) \cdot [z(u \in x)]
  \]

  Can do this when some or none of the docs are labeled

EM in General

- We’ll use EM over and over again to fill in missing data

  **Convenience Scenario**: we want \( P(x) \), including \( y \) just makes the model simpler (e.g. mixing weights)

  **Induction Scenario**: we actually want to know \( y \) (e.g. clustering)

  NLP differs from much of machine learning in that we often want to interpret or use the induced variables (which is tricky at best)

  **General approach**: alternately update \( y \) and \( \theta \)

  E-step: make a guess at posteriors \( P(y \mid x, \theta) \)

  - This means scoring all completions with the current parameters
  
  M-step: fill \( \theta \) to these completions

  - This is usually the easy part – treat the completions as (fractional) complete data
  
  Formally, we are maximizing a lower bound on the observed data likelihood

  In general, we start with some noisy labelings and the noise adjusts into patterns based on the data and the model

  We’ll see lots of examples in this course

  EM is only locally optimal (why?)

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)