CS 401R:
Natural Language Processing

Lecture #12: Text Classification

Thanks to Dan Klein of UC Berkeley for many of the materials used in this lecture.

Announcements

- Project #1, Part 1
  - Due: Today
- Reading Report #6
  - M&S 7
  - Due: Wednesday
- Project #1, Part 2
  - Early: Wednesday
  - Due: Friday

LM Big Picture

$\underline{w} = w_1, w_2, \dots, w_n$

REAL: $P(\underline{w}) = P(w_1, w_2, \dots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, \dots, w_{i-1})$

EST. $\hat{P}(\underline{w}) = \prod_{i=1}^{n} P(w_i \mid w_{i-k}, \dots, w_{i-1})$

$= \prod_{i=1}^{n} \frac{c(w_{i-k}, \dots, w_{i-1}, w_i)}{c(w_{i-k}, \dots, w_{i-1})}$

RE-EST. $P^*(\underline{w}) = \prod_{i=1}^{n} \frac{c^*(\underline{w}_k, w_i)}{c^*(\underline{w}_k)}$

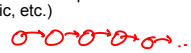
$P^{**}(\underline{w}) = \lambda P_1^*(\underline{w}) + (1-\lambda) P_2^*(\underline{w})$

Objectives

- Introduce the problem of text classification
- Understand the two event representations for Naïve Bayes
- Understand why they work and where they break down

Overview

- So far: language models give P(w)
  - Help model fluency for various noisy-channel processes (MT, ASR, etc.)
  - N-gram models don't represent any deep variables in natural 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.

Text Categorization

- Want to classify documents into broad semantic classes (e.g. sports, entertainment, technology, politics, etc.)

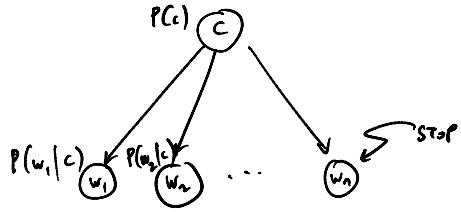
Democratic vice presidential candidate John Edwards on Sunday accused President Bush and Vice President Dick Cheney of misleading Americans by implying a link between deposed Iraqi President Saddam Hussein and the Sept. 11, 2001 terrorist attacks.

While No. 1 Southern California and No. 2 Oklahoma had no problems holding on to the top two spots with lopsided wins, four teams fell out of the rankings — Kansas State and Missouri from the Big 12 and Clemson from the Atlantic Coast Conference and Oregon from the Pac-10.

- 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 in an upcoming lecture …

Naïve-Bayes Models

- Idea: pick a class, then generate a document using a language model given that class.



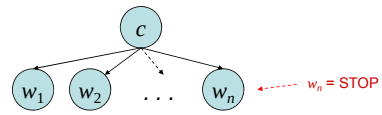
Naïve-Bayes Models

- Naïve-Bayes assumption: all words are conditionally independent of one another *given the class*.

$= P(c)P(w_1|c)P(w_2|w_1,c)\ldots P(w_n|w_1,\ldots,w_{n-1},c)$

$$P(c, w_1, w_2, \ldots w_n) = P(c)\prod_i \boxed{P(w_i \mid c)}$$

We have to smooth these!



- Compare to a unigram language model:

$$P(w_1, w_2, \ldots w_n) = \prod_i P(w_i)$$

Using NB for Classification

- We consider a set of classes:
  $C = \{c_1, c_2, \ldots, c_m\}$
- We have a joint model of classes and documents:

$$P(c,d) = P(c, w_1, w_2, \ldots w_n) = P(c)\prod_i P(w_i \mid c)$$

- We can easily compute the posterior probability of a class given a document:

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)} = \frac{P(c)P(d \mid c)}{\sum_{c' \in C} P(c', d)} = \frac{P(c)P(d \mid c)}{\sum_{c' \in C}[P(c')P(d \mid c')]}$$

$$= P(c \mid w_1, w_2, \ldots w_n) = \frac{P(c)\prod_i P(w_i \mid c)}{\sum_{c'}\left[P(c')\prod_i P(w_i \mid c')\right]}$$

Classifying with Naïve Bayes

- Given document *d*,

$$P(c \mid d) = P(c \mid w_1, w_2, \ldots w_n) = \frac{P(c)\prod_i P(w_i \mid c)}{\sum_{c'}\left[P(c')\prod_i P(w_i \mid c')\right]}$$

$$\hat{c} = \arg\max_c P(c \mid d) = \arg\max_c \frac{P(c)\prod_i P(w_i \mid c)}{\sum_{c'}\left[P(c')\prod_i P(w_i \mid c')\right]} \quad = P(d)$$

$$= \arg\max_c P(c)\prod_i P(w_i \mid c)$$

Log-space Calculations

$$P(c \mid d) = \frac{P(c)\prod_i P(w_i \mid c)}{\sum_{c'}\left[P(c')\prod_i P(w_i \mid c')\right]}$$

$$\log P(c|d) = \log\left[P(c)\prod_i P(w_i|c)\right] - \log\left[\sum_{c'} P(c')\prod_i P(w_i|c')\right]$$

$$= \log P(c) + \sum_i \log P(w_i|c)$$

$$- \log \sum_{c'}\left[\log P(c') + \sum_i \log P(w_i|c')\right]$$

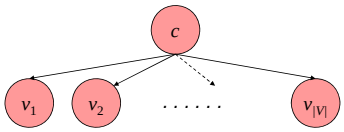
$$\hat{c} = \arg\max_c \log P(c|d) = \arg\max_c \log P(c) + \sum_i \log P(w_i|c)$$

Using NB for Classification

- What about totally unknown words?
- Can work shockingly well for text categorization (esp. in the wild)
- How can unigram models be so terrible for language modeling, but class-conditional unigram models work for text cat.?
- Numerical / speed issues
- How about NB for spam detection?

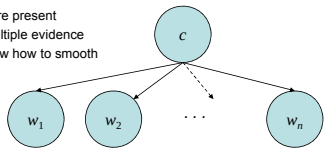
Two NB Formulations

- The binary model: "Multivariate Bernoulli"
  - 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



Two NB Formulations

- "Multinomial model" (using McCallum & Nigam's convention)
  - One node per seen word *type* in the document representing a count of word occurrence
  - Involves multinomial coefficient (see the paper)
- Easier to conceptualize as: Categorical model; i.e., The class-conditional unigram model
  - One node per word *token* in the document
  - Equivalent to the multinomial model in the end
- Both:
  - Driven by words which are present
  - Multiple occurrences, multiple evidence
  - Better overall – plus, know how to smooth



Performance on 20 Newsgroups

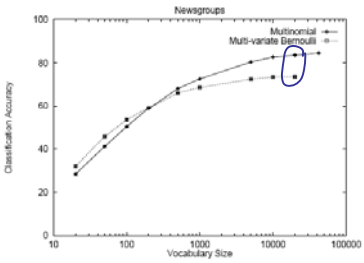
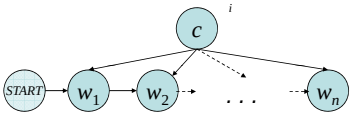


Figure 3 from McCallum & Nigam (1998)

Class-Conditional LMs

- Can have a topic variable for other language model types

$$P(c, w_1, w_2, \ldots w_n) = P(c) \prod_i P(w_i \mid w_{i-1}, c)$$



- Could be characters instead of words
  - as in PNP classification -- in Project #2 coming up!
- Could sum out the topic variable and use as a language model

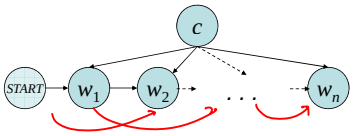
$$P(w_1, w_2, \ldots w_n) = \sum_{c \in C} P(c, w_1, w_2, \ldots w_n)$$

- How might a class-conditional n-gram language model behave differently from a standard n-gram model?

Class-Conditional LMs

- For a trigram:

$$P(c, w_1, w_2, \ldots w_n) = P(c) \prod_i P(w_i \mid w_{i-1}, w_{i-2}, c)$$



-

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.

La 38e législature se réunira à 11 heures le lundi 4 octobre 2004, et la première affaire à l'ordre du jour sera l'élection du président de la Chambre des communes. Son Excellence la Gouverneure générale ouvrira la première session de la 38e législature avec un discours du Trône le mardi 5 octobre 2004.

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

      Σύμφωνο σταθερότητας και ανάπτυξης
      Patto di stabilità e di crescita

  - Option: build a character-level language model

Later this Semester

- Problem: What if your data doesn't have labels?
- Solution: Text clustering
- Technique: Expectation Maximization

Next

- Word Sense Disambiguation!