CS5740: Natural Language Processing

Text Classification

Instructor: Yoav Artzi

Slides adapted from Dan Klein, Dan Jurafsky, Chris Manning, Michael Collins, Luke Zettlemoyer, Yejin Choi, and Slav Petrov
Overview

- Classification Problems
  - Spam vs. Non-spam, Text Genre, Word Sense, etc.

- Supervised Learning
  - Naïve Bayes
  - Log-linear models (Maximum Entropy Models)
  - Weighted linear models and the Perceptron
  - Neural networks
Supervised Learning: Data

• Learning from annotated data
• Often the biggest problem
• Why?
  – Annotation requires specific expertise
  – Annotation is expensive
  – Data is private and not accessible
  – Often difficult to define and be consistent
• Always think about the data, and how much of it your model needs (even better: think of the data first, model second)
Held-out Data

• Important tool for estimating generalization:
  – Train on one set, and evaluate during development on another
  – Test data: only use once!
Classification

• Automatically make a decision about inputs
  – Example: document $\rightarrow$ category
  – Example: image of digit $\rightarrow$ digit
  – Example: image of object $\rightarrow$ object type
  – Example: query + webpage $\rightarrow$ best match
  – Example: symptoms $\rightarrow$ diagnosis
  – ...

• Three main ideas:
  – Representation as feature vectors
  – Scoring by linear functions
  – Learning by optimization
Example: Spam Filter

- Input: email
- Output: spam/ham
- Setup:
  - Get a large collection of example emails, each labeled “spam” or “ham”
  - Note: someone has to hand label all this data!
  - Goal: learn to predict labels of new, future emails
- Features: The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: SenderInContacts
  - ...

---

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.
More Examples

• Variants of spam
  – Comments, email
• Abuse and harassment
• Fake reviews
• Simple intent (”siri, what is the time?”)
• Topic (for news) and domain (for papers)
• Medical
  – Disease, surgical procedure, insurance code
Probabilistic Classifiers

- Two broad approaches to predicting a class $y^*$

- Joint / Generative (e.g., Naïve Bayes)
  - Work with a \textit{joint} probabilistic model of the data
  - Assume functional form for $P(X|y)$, $P(y)$
  - Estimate probabilities from data
  - Use Bayes’ rule to calculate $P(y|X)$
  - E.g., represent $p(y,X)$ as Naïve Bayes model, compute $y^* = \arg\max_y p(y,X) = \arg\max_y p(y)p(X|y)$
  - \textbf{Advantages:} learning weights is easy and well understood

- Conditional / Discriminative (e.g., Logistic Regression)
  - Work with \textit{conditional} probability $p(y|X)$
  - We can then directly compute $y^* = \arg\max_y p(y|X)$
  - Estimate parameters from data
  - \textbf{Advantages:} Don’t have to model $p(X)$! Can develop \textit{feature rich} models for $p(y|X)$
Text Categorization

- Goal: classify documents into broad semantic topics

Obama is hoping to rally support for his $825 billion stimulus package on the eve of a crucial House vote. Republicans have expressed reservations about the proposal, calling for more tax cuts and less spending. GOP representatives seemed doubtful that any deals would be made.

California will open the 2009 season at home against Maryland Sept. 5 and will play a total of six games in Memorial Stadium in the final football schedule announced by the Pacific-10 Conference Friday. The original schedule called for 12 games over 12 weekends.

- Which one is the politics document? (And how much deep processing did that decision take?)
- First approach: bag-of-words and Naïve-Bayes models
- More approaches later…
- Usually begin with a labeled corpus containing examples of each class
General Text Classification

• Input:
  – Document $X$ of length $|X|$ is a sequence of tokens:
    $$ X = \langle x_1, \ldots, x_{|X|} \rangle $$

• Output:
  – One of $k$ labels $y$
Naïve-Bayes Models

- Generative model: pick a topic, then generate a document.
- Naïve-Bayes assumption:
  - All words are independent given the topic.

\[
p(y, X) = q(y) \prod_{i=1}^{\left|X\right|} q(x_i | y)
\]
Using NB for Classification

• We have a joint model of topics and documents

\[ p(y, X) = q(y) \prod_{i=1}^{|X|} q(x_i \mid y) \]

• To assign a label \( y^* \) to a new document \( \langle x_1, x_2, ..., x_n \rangle \):

\[ y^* = \arg \max_y p(y, X) = \arg \max_y q(y) \prod_{i=1}^{|X|} q(x_i \mid y) \]