ITCS 4111/5111: Introduction to NLP

Text Classification using Naïve Bayes

Razvan C. Bunescu
Department of Computer Science @ CCI
rbunescu@uncc.edu
Text Classification: Sentiment Analysis

Movie reviews:

Positive: This was a great movie, which I thoroughly enjoyed.

Negative: I was very disappointed in this movie, it was a waste of time.

• Lexical features, e.g. presence of words such as great or disappointed, can be used to determine the sentiment orientation.
  – Can you think of examples where the same word may be used for both types of sentiment? How would you fix that?

• Represent each review as a bag-of-words feature vector:
  – High dimensional, sparse feature vector => use sparse representations that map features to indices.
  – Feature value is 1 if word is present, 0 otherwise:
    • Can use more sophisticated word weighting schemes from IR, such as tf.idf.
    • Can use stems instead of tokens.
Why sentiment analysis?

- **Movies:**
  - Is this review positive or negative?
    - Predict box office performance from sentiment in initial reviews, …
- **Products:**
  - What do people think about the new iPhone?
    - Predict market share, value of manufacturer company stock, …
- **Public sentiment:**
  - How is consumer confidence?
    - Predict debt, mortgage lending, credit card use, …
- **Politics:**
  - What do people think about this candidate or issue?
    - Predict election outcomes, ballot vote outcomes, …
Scherer Typology of Affective States

- **Emotion**: brief organically synchronized ... evaluation of a major event
  - angry, sad, joyful, fearful, ashamed, proud, elated
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - cheerful, gloomy, irritable, listless, depressed, buoyant
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous
- **Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - liking, loving, hating, valuing, desiring
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - nervous, anxious, reckless, morose, hostile, jealous
Sentiment Analysis:
Is the attitude of the text positive or negative?

- **Emotion**: brief organically synchronized … evaluation of a major event
  - angry, sad, joyful, fearful, ashamed, proud, elated
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - cheerful, gloomy, irritable, listless, depressed, buoyant
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - friendly, flirtatious, distant, cold, warm, supportive, contemptuous
- **Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - liking, loving, hating, valuing, desiring
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - nervous, anxious, reckless, morose, hostile, jealous
Text Classification

• Sentiment analysis.

• Spam detection.

• Authorship identification.

• Language identification.

• Assigning subject categories, topics, or genres.
Text classification: Spam detection

- Email filtering:
  - Provide emails labeled as \{Spam, Ham\}.
  - Train Naïve Bayes model to discriminate between the two.
    - [Sahami, Dumais & Heckerman, AAAI’98]
Is this spam?

Subject: Important notice!
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients;;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.


Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

© Stanford University. All Rights Reserved.

• Adversarial setting:
  – Text encapsulated in images.
  – Misspelled words, …
Text classification: Authorship identification

• Who wrote which Federalist papers?
  – 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
    • Authors tried to conceal their identities. => authorship of 12 of the letters in dispute.
    • 1963: solved by Mosteller and Wallace using Bayesian methods.

• Who is the author of Shakespeare’s plays?
  – Disclaimer: It is widely accepted that Shakespeare is the author of Shakespeare.
  – Theory: Francis Bacon wrote the plays.
    • Bacon’s official rise might have been impacted if he were known as the authors of plays for the public stage.
      – The plays were credited to Shakespeare, a front for shielding Bacon.
Text classification: Mapping Medline article to MeSH categories

• MeSH Subject Category Hierarchy:
  – Antagonists and Inhibitors
  – Blood Supply
  – Chemistry
  – Drug Therapy
  – Embryology
  – Epidemiology
  – …
Text Classification: Definition

- **Input:**
  - a document $d \in D$
  - a fixed set of classes $C = \{C_1, C_2, \ldots, C_K\}$

- **Output:**
  - a predicted class $C_k \in C$. 
Rule-based vs. Machine learning

• Hand-coded Rules based on combinations of words or other features:
  – Spam filtering: black-list-address OR (“dollars” AND “you have been selected”).
  – Sentiment: use affective lexicons.
    • Accuracy can be high:
      – If rules carefully refined by expert.
      – But building and maintaining these rules is expensive.

• Supervised learning:
  – Input:
    • A fixed set of classes $C = \{C_1, C_2, \ldots, C_K\}$
    • A training set of $N$ hand-labeled documents $(d_1, t_1), (d_2, t_2), \ldots, (d_N, t_N)$, where $t_n \in C$
  – Output:
    • A learned classifier $h: D \rightarrow C$
Classification Algorithms

• Train a classification algorithm on the labeled feature vectors, i.e. training examples.
  – Use trained model to determine the sentiment orientation of new, unseen reviews.

• Machine learning models:
  – Perceptron
  – Support Vector Machines
  – **Logistic Regression**
  – Naïve Bayes
  – Neural networks
  – k-Nearest Neighbors
  – …
Naïve Bayes model

• Simple ("naive") classification method based on:
  – Bayes rule.
  – Simple class-conditional independence between words.
• Relies on very simple representation of document:
  – Bag of Words (BoW).
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!
The bag of words representation

\[ h(\text{seen} = 2, \text{sweet} = 1, \text{whimsical} = 1, \text{recommend} = 1, \text{happy} = 1, \ldots) = C_k \]
Movie reviews:

**Positive:** This was a great movie, which I thoroughly enjoyed.

**Positive:** Despite the bad reviews I read online, I liked this move.

**Negative:** The movie was not as good as I expected.

- It appears that the bag-of-words approach is not sufficient.
- Can try to address negation:
  - Use bigram NOT_X for all words X following the negation [Pang et al. EMNLP’02].
- Model sentiment compositionality:
  - Train recursive deep models over sentiment treebanks [Socher et al., EMNLP’13]
- Apply more sophisticated classifiers:
  - Convolutional Neural Networks (CNNs) [Kim, 2014]
More examples showing the limitations of *bag-of-words* models [Eisenstein, 2019]:

a. That’s not bad for the first day.
b. This is not the worst thing that can happen.
c. It would be nice if you acted like you understood.
d. There is no reason at all to believe that the polluters are suddenly going to become reasonable. (Wilson et al., 2005)
e. This film should be brilliant. The actors are first grade. Stallone plays a happy, wonderful man. His sweet wife is beautiful and adores him. He has a fascinating gift for living life fully. It sounds like a great plot, however, the film is a failure. (Pang et al., 2002)
Bayes’ Rule Applied to Documents and Classes

- For a document $d$ and a class $c$:

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

- **Inference** $\equiv$ find $c_{MAP}$ to minimize misclassification rate:

$$c_{MAP} = \arg\max_{c \in C} P(c \mid d) = \arg\max_{c \in C} \frac{P(d \mid c)P(c)}{P(d)} = \arg\max_{c \in C} P(d \mid c)P(c)$$
Naive Bayes Classifier

- **Inference** (at test time): find maximum a posteriori (MAP) class:

\[
\begin{align*}
    c_{MAP} &= \arg\max_{c \in C} P(d \mid c)P(c) \\
    &= \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n \mid c)P(c)
\end{align*}
\]

- If each feature \( x_j \in X \) and class \( c \in C \), then \(|X|^{n \times |C|}\) params \( P(x_1, x_2, \ldots, x_n \mid c) \):
  - \( x_j \) could be word at position \( j \) in document \( d \), \( X \) could be entire vocabulary.
    - Number of params is **exponential** in the size of vocabulary!
      - Could only be estimated if a very, very large number of training examples was available.
        » Unfeasible in practice.
The Naïve Bayes Model

• **NB assumption**: features are conditionally independent given the target class.

\[
P(d|c) = P(x_1, \ldots, x_n|c) = P(x_1|c) P(x_2|c) \cdots P(x_n|c)
\]

• **BoW assumption**: assume position doesn’t matter.

\[
P(d|c) = P(x_1|c) P(x_2|c) \cdots P(x_n|c) = \prod_{x \in d} P(x|c)
\]

• Assuming binary features, i.e. word \(w\) appears (or not) at position \(j\):

\Rightarrow \text{need to estimate only } |X| \times |C| \text{ parameters, a lot less than } |X|^{n \times |C|}
MAP inference at test time, using Naïve Bayes model:

\[ c_{MAP} = \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n | c) P(c) \]

\[ c_{MAP} = c_{NB} = \arg\max_{c \in C} P(c) \prod_{x \in d} P(x | c) \]

• Use probabilities over all word positions in the document \( d \):

\[ c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j) \]

Multiplying lots of probabilities can result in floating-point underflow!
Naïve Bayes: Use log-space to avoid underflow

• Instead of this:

\[ c_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j) \]

\[ \log(ab) = \log(a) + \log(b) \]

• Work in log-space:

\[ c_{NB} = \arg\max_{c_j \in C} \left[ \log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i | c_j) \right] \]

• This is ok since \( \log \) doesn't change the ranking of the classes:
  – class with highest prob still has highest log prob.

• Model is now just max of sum of weights:
  – A **linear** function of the inputs. So naive bayes is a **linear classifier**
Learning the Multinomial Naive Bayes Model

- **Maximum Likelihood** estimates:
  - Use the frequencies of features in the data.

\[
\hat{P}(c_j) = \frac{\text{doccount}(C = c_j)}{N_{\text{doc}}} \quad \hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}
\]

- Create mega-document for topic \( j \) by concatenating all docs in this topic:
  - Use frequency of \( w \) in mega-document.

\[
\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}
\]

fraction of times word \( w_i \) appears among all words in documents of topic \( c_j \)
Problem with Maximum Likelihood

- What if we have seen no training documents with the word *fantastic* and classified in the topic *positive (thumbs-up)*?

\[
\hat{P}("fantastic" \mid \text{positive}) = \frac{\text{count}("fantastic", \text{positive})}{\sum_{w \in V} \text{count}(w, \text{positive})} = 0
\]

- Zero probabilities cannot be conditioned away, no matter the other evidence!

\[
c_{\text{MAP}} = \arg\max_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)
\]
Laplace Smoothing for Naïve Bayes

- **Laplace (add-1) smoothing:**
  - $|V|$ “hallucinated” examples spread evenly over all $|V|$ values of $w_i$.

\[
\hat{P}(w_i | c) = \frac{\text{count}(w_i, c)}{\sum_{w \in V} (\text{count}(w, c))} = \frac{\text{count}(w_i, c) + 1}{\left( \sum_{w \in V} \text{count}(w, c) \right) + |V|}
\]