**Classification & Information Theory**

*Lecture #5*

**Introduction to Natural Language Processing**

*CMPSCI 585, Fall 2004*

*University of Massachusetts Amherst*

Aron Culotta

Slides courtesy of Andrew McCallum

---

**Recipe for Solving a NLP Task Statistically**

1) **Data**: Notation, representation
2) **Problem**: Write down the problem in notation
3) **Model**: Make some assumptions, define a parametric model
4) **Inference**: How to search through possible answers to find the best one
5) **Learning**: How to estimate parameters
6) **Implementation**: Engineering considerations for an efficient implementation

---

**A Probabilistic Approach to Classification: “Naïve Bayes”**

Pick the most probable class, given the evidence:

\[
    c^* = \arg \max_{c_j} \Pr(c_j \mid d)
\]

\[
    c_j \text{ is a class (like “Planning”)}
\]

\[
    d \text{ is a document (like “language intelligence proof...”)}
\]

Bayes Rule:

\[
    \Pr(c_j \mid d) = \frac{\Pr(c_j) \Pr(d \mid c_j)}{\Pr(d)} = \frac{\Pr(c_j) \prod_{i=1}^{n} \Pr(w_{ij} \mid c_j)}{\sum_{c_i} \Pr(c_i) \prod_{i=1}^{n} \Pr(w_{ij} \mid c_i)}
\]

\[
    w_{ij} \text{ is the } i\text{th word in } d \text{ (like “proof”)}
\]

---

**Document Classification by Machine Learning**

- “Temporal reasoning for planning has long been studied formally. We discuss the semantics of several planning tasks.”
- “Neural networks and other machine learning methods of classification...”
- “Garbage collection for strongly-typed languages...”
- “Multimedia streaming video for...”
- “User studies of GUI...”
- “Planning with temporal reasoning has been...”
- “Based on the semantics of program dependence...”

---

**Work out Naïve Bayes formulation interactively on the board**

**(Engineering) Components of a Naïve Bayes Document Classifier**

- Split documents into training and testing
- Cycle through all documents in each class
- Tokenize the character stream into words
- Count occurrences of each word in each class
- Estimate \( P(w_j | c) \) by a ratio of counts (+1 prior)
- For each test document, calculate \( P(c_j | d) \) for each class
- Record predicted (and true) class, and keep accuracy statistics
Parameter Estimation in Naïve Bayes

Estimate of $P(c)$

$$P(c_i) = \frac{1 + \sum \text{Count}(d \in c_i)}{1 + \sum \text{Count}(d \in c_i)}$$

Estimate of $P(w|c)$

$$\hat{P}(w_j | c_i) = \frac{1 + \sum \text{Count}(w_j, d_i)}{1 + \sum \text{Count}(w_j, d_i)}$$

Programming Assignment 2 Help

$$\Pr(c_j | d) \propto \Pr(c_j) \prod \Pr(w_j | c_j)$$

$$\log(\Pr(c_j | d)) \propto \log(\Pr(c_j)) + \sum_{i=1}^{k} \log(\Pr(w_j | c_j))$$

• To get back to $\Pr(c_j | d)$
  • Subtract a constant to make all positive
  • exp()

Common words in Tom Sawyer (71,370 words)

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq.</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>determiner (article)</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>conjunction</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>determiner</td>
</tr>
<tr>
<td>to</td>
<td>1725</td>
<td>preposition, verbal infinitive marker</td>
</tr>
<tr>
<td>of</td>
<td>1440</td>
<td>preposition</td>
</tr>
<tr>
<td>was</td>
<td>1161</td>
<td>auxiliary verb</td>
</tr>
<tr>
<td>it</td>
<td>1027</td>
<td>(personal/expletive) pronoun</td>
</tr>
<tr>
<td>in</td>
<td>906</td>
<td>preposition</td>
</tr>
<tr>
<td>that</td>
<td>877</td>
<td>complementizer, demonstrative</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>I</td>
<td>783</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>his</td>
<td>772</td>
<td>(possessive) pronoun</td>
</tr>
<tr>
<td>you</td>
<td>686</td>
<td>(personal) pronoun</td>
</tr>
<tr>
<td>Tom</td>
<td>679</td>
<td>proper noun</td>
</tr>
<tr>
<td>with</td>
<td>642</td>
<td>preposition</td>
</tr>
</tbody>
</table>

Frequencies of frequencies in Tom Sawyer

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency of Frequency</th>
<th>Frequency of Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1</td>
<td>3993</td>
</tr>
<tr>
<td>and</td>
<td>2</td>
<td>1292</td>
</tr>
<tr>
<td>a</td>
<td>3</td>
<td>664</td>
</tr>
<tr>
<td>to</td>
<td>4</td>
<td>410</td>
</tr>
<tr>
<td>of</td>
<td>5</td>
<td>243</td>
</tr>
<tr>
<td>was</td>
<td>6</td>
<td>199</td>
</tr>
<tr>
<td>it</td>
<td>7</td>
<td>131</td>
</tr>
<tr>
<td>in</td>
<td>8</td>
<td>82</td>
</tr>
<tr>
<td>that</td>
<td>9</td>
<td>91</td>
</tr>
<tr>
<td>he</td>
<td>10</td>
<td>540</td>
</tr>
<tr>
<td>I</td>
<td>11-50</td>
<td>99</td>
</tr>
<tr>
<td>you</td>
<td>51-100</td>
<td>102</td>
</tr>
</tbody>
</table>

Ziph’s law Tom Sawyer

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq.</th>
<th>Rank</th>
<th>$f^r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>3332</td>
<td>1</td>
<td>3332</td>
</tr>
<tr>
<td>and</td>
<td>2972</td>
<td>2</td>
<td>5944</td>
</tr>
<tr>
<td>a</td>
<td>1775</td>
<td>3</td>
<td>5235</td>
</tr>
<tr>
<td>to</td>
<td>1725</td>
<td>4</td>
<td>6880</td>
</tr>
<tr>
<td>of</td>
<td>1440</td>
<td>5</td>
<td>8450</td>
</tr>
<tr>
<td>was</td>
<td>1161</td>
<td>6</td>
<td>8820</td>
</tr>
<tr>
<td>it</td>
<td>1027</td>
<td>7</td>
<td>8880</td>
</tr>
<tr>
<td>in</td>
<td>906</td>
<td>8</td>
<td>9000</td>
</tr>
<tr>
<td>that</td>
<td>877</td>
<td>9</td>
<td>9400</td>
</tr>
<tr>
<td>he</td>
<td>877</td>
<td>10</td>
<td>9480</td>
</tr>
<tr>
<td>I</td>
<td>783</td>
<td>11</td>
<td>9920</td>
</tr>
<tr>
<td>his</td>
<td>772</td>
<td>12</td>
<td>10440</td>
</tr>
<tr>
<td>you</td>
<td>686</td>
<td>13</td>
<td>10400</td>
</tr>
<tr>
<td>Tom</td>
<td>679</td>
<td>14</td>
<td>10400</td>
</tr>
<tr>
<td>with</td>
<td>642</td>
<td>15</td>
<td>10400</td>
</tr>
</tbody>
</table>

Ziph’s law Tom Sawyer

<table>
<thead>
<tr>
<th>Word</th>
<th>Freq.</th>
<th>Rank</th>
<th>$f^r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>turned</td>
<td>51</td>
<td>200</td>
<td>10200</td>
</tr>
<tr>
<td>you'll</td>
<td>30</td>
<td>300</td>
<td>9000</td>
</tr>
<tr>
<td>name</td>
<td>21</td>
<td>400</td>
<td>8400</td>
</tr>
<tr>
<td>comes</td>
<td>16</td>
<td>500</td>
<td>8000</td>
</tr>
<tr>
<td>group</td>
<td>13</td>
<td>600</td>
<td>7800</td>
</tr>
<tr>
<td>lead</td>
<td>11</td>
<td>700</td>
<td>7700</td>
</tr>
<tr>
<td>friends</td>
<td>10</td>
<td>800</td>
<td>8000</td>
</tr>
<tr>
<td>begin</td>
<td>9</td>
<td>900</td>
<td>8100</td>
</tr>
<tr>
<td>family</td>
<td>8</td>
<td>1000</td>
<td>8000</td>
</tr>
<tr>
<td>brushed</td>
<td>4</td>
<td>2000</td>
<td>8000</td>
</tr>
<tr>
<td>sirs</td>
<td>2</td>
<td>3000</td>
<td>6000</td>
</tr>
<tr>
<td>Could</td>
<td>2</td>
<td>4000</td>
<td>8000</td>
</tr>
<tr>
<td>Applausive</td>
<td>1</td>
<td>8000</td>
<td>8000</td>
</tr>
</tbody>
</table>
Zipf’s law

\[ f \propto \frac{1}{r} \]

In other words, there is a constant, \( k \), such that

\[ f \cdot r = k \]

**Information Theory**

- "The sun will come up tomorrow."
- "Greenspan was shot and killed this morning."

**Efficient Encoding**

- I have a 8-sided die. How many bits do I need to tell you what face I just rolled?
- My 8-sided die is unfair
  - \( P(1)=0.5, P(2)=0.125, P(3)=...=P(8)=0.0625 \)

**Entropy (of a Random Variable)**

- Average length of message needed to transmit the outcome of the random variable.
- First used in:
  - Data compression
  - Transmission rates over noisy channel
"Coding" Interpretation of Entropy

- Given some distribution over events P(X)...
- What is the average number of bits needed to encode a message (a event, string, sequence)
- = Entropy of P(X):
  \[ H(p(X)) = - \sum_{x \in X} p(x) \log_2(p(x)) \]

Notation: \( H(X) = H(p) = H_X(p) = H(p_X) \)

What is the entropy of a fair coin? A fair 32-sided die?
What is the entropy of an unfair coin that always comes up heads?
What is the entropy of an unfair 6-sided die that always \( \{1,2\} \)
Upper and lower bound? (Prove lower bound?)

Entropy and Expectation

- Recall \( E[X] = \sum_{x \in X} x \cdot p(x) \)
- Then \( E[-\log_2(p(x))] = \sum_{x \in X} -\log_2(p(x)) \cdot p(x) = H(X) \)

Entropy, intuitively

- High entropy ~ "chaos", fuzziness, opposite of order
- Comes from physics:
  - Entropy does not go down unless energy is used
- Measure of uncertainty
  - High entropy: a lot of uncertainty about the outcome, uniform distribution over outcomes
  - Low entropy: high certainty about the outcome

Claude Shannon

- Claude Shannon 1916 - 2001
  Creator of Information Theory
- Lays the foundation for implementing logic in digital circuits as part of his Masters Thesis (1939)
- "A Mathematical Theory of Communication" (1948)

Joint Entropy and Conditional Entropy

- Two random variables: X (space \( W_X \)), Y (Y)
- Joint entropy
  - no big deal: \( (X,Y) \) considered a single event:
    \[ H(X,Y) = - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x,y) \]
- Conditional entropy:
  \[ H(X|Y) = - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 p(x|y) \]
  - recall that \( H(X) = E[-\log_2(p(x))] \)
  (weighted average, and weights are not conditional)
  - How much extra information you need to supply to transmit X given that the other person knows Y.
Conditional Entropy (another way)

\[ H(Y | X) = \sum_x p(x)H(Y | X = x) \]
\[ = \sum_x p(x)(-\sum_y p(y | x)\log_2(p(y | x))) \]
\[ = -\sum_x \sum_y p(x)p(y | x)\log_2(p(y | x)) \]
\[ = -\sum_x \sum_y p(x,y)\log_2(p(y | x)) \]

Chain Rule for Entropy

- Since, like random variables, entropy is based on an expectation..
- \[ H(X, Y) = H(X|Y) + H(Y) \]
- \[ H(X, Y) = H(Y|X) + H(X) \]

Cross Entropy

- What happens when you use a code that is sub-optimal for your event distribution?
  - I created my code to be efficient for a fair 8-sided die.
- How many bits on average for the optimal code?
- How many bits on average for the sub-optimal code?

\[ H(p,q) = -\sum_{x \in X} p(x)\log_2(q(x)) + \sum_{x \in X} p(x)\log_2(p(x)) \]

KL Divergence

- What are the average number of bits that are wasted by encoding events from distribution \( p \) using distribution \( q \)?

\[ D(p \| q) = H(p,q) - H(p) \]
\[ = -\sum_{x \in X} p(x)\log_2(q(x)) + \sum_{x \in X} p(x)\log_2(p(x)) \]
\[ = \sum_{x \in X} p(x)\log_2\left(\frac{p(x)}{q(x)}\right) \]

A sort of “distance” between distributions \( p \) and \( q \), but it is not symmetric! It does not satisfy the triangle inequality!

Mutual Information

- Recall: \( H(X) \) = average # bits for me to tell you which event occurred from distribution \( P(X) \).
- Now, first I tell you event \( y \in Y \), \( H(X|Y) \) = average # bits necessary to tell you which event occurred from distribution \( P(X|Y) \)?
- By how many bits does knowledge of \( Y \) lower the entropy of \( X \)?

\[ I(X;Y) = H(X) - H(X|Y) \]
\[ = H(X) - H(Y) - H(X,Y) \]
\[ = \sum_x p(x)\log_2\left(\frac{1}{p(x)}\right) + \sum_y p(y)\log_2\left(\frac{1}{p(y)}\right) - \sum_{x,y} p(x,y)\log_2\left(\frac{p(x)}{p(x)p(y)}\right) \]
\[ = \sum_{x,y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \]

Mutual Information

- Symmetric, non-negative.
- Measure of independence.
  - \( I(X;Y) = 0 \) when \( X \) and \( Y \) are independent
  - \( I(X;Y) \) grows both with degree of dependence and entropy of the variables.
- Sometimes also called "information gain"
Pointwise Mutual Information

- Previously measuring mutual information between two random variables.
- Could also measure mutual information between two events

\[ I(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \]