Probability
Lecture #7

Introduction to Natural Language Processing
CMPSCI 585, Fall 2007
University of Massachusetts Amherst

Andrew McCallum
Today’s Main Points

• Remember (or learn) about probability theory
  – samples, events, tables, counting
  – Bayes’ Rule, and its application
  – A little calculus?
  – random variables
  – Bernoulli and Multinomial distributions: the workhorses of Computational Linguistics.
  – Multinomial distributions from Shakespeare.
Probability Theory

- Probability theory deals with predicting how likely it is that something will happen.
  - Toss 3 coins, how likely is it that all come up heads?
  - See phrase “more lies ahead”, how likely is it that “lies” is noun?
  - See “Nigerian minister of defense” in email, how likely is it that the email is spam?
  - See “Le chien est noir”, how likely is it that the correct translation is “The dog is black”? 
Probability and CompLing

• Probability is the backbone of modern computational linguistics... because:
  – Language is ambiguous
  – Need to integrate evidence

• Simple example (which we will revisit later)
  – I see the first word of a news article: “glacier”
  – What is the probability the language is French? English?
  – Now I see the second word: “melange”.
  – Now what are the probabilities?
Experiments and Sample Spaces

• **Experiment** (or *trial*)
  - repeatable process by which observations are made
  - e.g. tossing 3 coins

• Observe *basic outcome* from *sample space*, \( \Omega \), (set of all possible basic outcomes), e.g.
  - one coin toss, *sample space* \( \Omega = \{ H, T \} \);
    *basic outcome* = H or T
  - three coin tosses, \( \Omega = \{ HHH, HHT, HTH, ..., TTT \} \)
  - Part-of-speech of a word, \( \Omega = \{ CC_1, CD_2, CT_3, ..., WRB_{36} \} \)
  - lottery tickets, \( |\Omega| = 10^7 \)
  - next word in Shakespeare play, \( |\Omega| = \) size of vocabulary
  - number of words in your Ph.D. thesis \( \Omega = \{ 0, 1, ..., \infty \} \) discrete, countably infinite
  - length of time of “a” sounds when I said “sample” continuous, uncountably infinite
Events and Event Spaces

- An *event*, $A$, is a set of basic outcomes, i.e., a subset of the sample space, $\Omega$.
  - Intuitively, a question you could ask about an outcome.
  - $\Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}$
  - e.g. basic outcome = THH
  - e.g. event = “has exactly 2 H’s”, $A=\{\text{THH, HHT, HTH}\}$
  - $A=\Omega$ is the certain event, $A=\emptyset$ is the impossible event.
  - For “not $A$”, we write $\bar{A}$

- A common *event space*, $F$, is the power set of the sample space, $\Omega$. (power set is written $2^\Omega$)
  - Intuitively: all possible questions you could ask about a basic outcome.
Probability

• A **probability** is a number between 0 and 1.
  – 0 indicates impossibility
  – 1 indicates certainty

• A **probability function**, $P$, (or **probability distribution**) assigns probability mass to events in the event space, $F$.
  – $P : F \rightarrow [0,1]$
  – $P(\Omega) = 1$
  – Countable additivity: For disjoint events $A_j$ in $F$
    \[ P(\bigcup_j A_j) = \sum_j P(A_j) \]

• We call $P(A)$ “the probability of event $A$”.

• Well-defined **probability space** consists of
  – sample space $\Omega$
  – event space $F$
  – probability function $P$
Probability (more intuitively)

• Repeat an experiment many, many times. (Let T = number of times.)
• Count the number of basic outcomes that are a member of event A. (Let C = this count.)
• The ratio C/T will approach (some unknown) but constant value.
• Call this constant “the probability of event A”; write it P(A).

Why is the probability this ratio of counts? Stay tuned! Maximum likelihood estimation at end.
Example: Counting

• “A coin is tossed 3 times. What is the likelihood of 2 heads?”
  – Experiment: Toss a coin three times, \( \Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \)
  – Event: basic outcome has exactly 2 H’s
    \( A = \{THH, HTH, HHT\} \)
• Run experiment 1000 times (3000 coin tosses)
• Counted 373 outcomes with exactly 2 H’s
• Estimated \( P(A) = 373/1000 = 0.373 \)
Example: Uniform Distribution

• “A *fair* coin is tossed 3 times. What is the likelihood of 2 heads?”
  – Experiment: Toss a coin three times,
    \[ \Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \]
  – Event: basic outcome has exactly 2 H’s
    \[ A = \{THH, HTH, HHT\} \]
• **Assume a *uniform distribution* over outcomes**
  – Each basic outcome is equally likely
    – \( P(\{HHH\}) = P(\{HHT\}) = \ldots = P(\{TTT\}) \)
• \( P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8} = 0.375 \)
Probability (again)

• A *probability* is a number between 0 and 1.
  – 0 indicates impossibility
  – 1 indicates certainty

• A *probability function*, $P$, (or *probability distribution*) distributes probability mass of 1 throughout the event space, $F$.
  – $P : F \rightarrow [0,1]$
  – $P(\Omega) = 1$
  – Countable additivity: For disjoint events $A_j$ in $F$
    \[ P(\bigcup_j A_j) = \sum_j P(A_j) \]

• The above are *axioms of probability theory*

• Immediate consequences:
  – $P(\emptyset) = 0$, $\bar{P}(A) = 1 - P(A)$, $A \subseteq B \rightarrow P(A) \leq P(B)$,
  – $\sum_{a \in \Omega} P(a) = 1$, for $a = \text{basic outcome}$. 
Vocabulary Summary

- **Experiment** = a repeatable process
- **Sample** = a possible outcome
- **Sample space** = all samples for an experiment
- **Event** = a set of samples

- **Probability distribution** = assigns a probability to each sample
- **Uniform distribution** = all samples are equi-probable
Collaborative Exercise

• You roll a fair die, then roll it again. What is the probability that you get the same number from both rolls?

• Explain in terms of event spaces and basic outcomes.
Joint and Conditional Probability

• **Joint probability** of A and B:
P(A ∩ B) is usually written P(A,B)

• **Conditional probability** of A given B:
P(A|B) = \frac{P(A,B)}{P(B)}

Updated probability of an event given some evidence

P(A) = *prior probability* of A

P(A|B) = *posterior probability* of A given *evidence* B
Joint Probability Table

What does it look like “under the hood”? 

P(precipitation, temperature)

<table>
<thead>
<tr>
<th></th>
<th>sun</th>
<th>rain</th>
<th>sleet</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>20s</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>30s</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>40s</td>
<td>0.06</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>50s</td>
<td>0.06</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>60s</td>
<td>0.06</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>70s</td>
<td>0.07</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>80s</td>
<td>0.07</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>90s</td>
<td>0.08</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>100s</td>
<td>0.08</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

it takes 40 numbers
### Conditional Probability Table

What does it look like “under the hood”?

\[
P(\text{precipitation} \mid \text{temperature})
\]

<table>
<thead>
<tr>
<th></th>
<th>sun</th>
<th>rain</th>
<th>sleet</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td>0.9</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>20s</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.2</td>
</tr>
<tr>
<td>30s</td>
<td>0.5</td>
<td>0.1</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>40s</td>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
<td>0.0</td>
</tr>
<tr>
<td>50s</td>
<td>0.6</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>60s</td>
<td>0.6</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>70s</td>
<td>0.7</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>80s</td>
<td>0.7</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>90s</td>
<td>0.8</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>100s</td>
<td>0.8</td>
<td>0.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

It takes 40 numbers
Two Useful Rules

- **Multiplication Rule**
  \[ P(A,B) = P(A|B) \cdot P(B) \]
  (equivalent to conditional probability definition from previous slide)

- **Total Probability Rule (Sum Rule)**
  \[ P(A) = P(A,B) + P(A,B^c) \]
  or more generally, if B can take on n values
  \[ P(A) = \sum_{i=1}^{n} P(A,B_i) \]
  (from additivity axiom)
Bayes Rule

- $P(A,B) = P(B,A)$, since $P(A \cap B) = P(B \cap A)$
- Therefore $P(A|B) \ P(B) = P(B|A) \ P(A)$, and thus...
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

“Normalizing constant”

Bayes Rule lets you swap the order of the dependence between events...

calculate $P(A|B)$ in terms of $P(B|A)$. 
Reverend Thomas Bayes

• Rumored to have been tutored by De Moivre.

• Was elected a Fellow of the Royal Society in 1742 despite the fact that at that time he had no published works on mathematics!


1702 - 1761

Same year Mozart wrote his symphony #1 in E-flat.
Independence

• Can we compute $P(A,B)$ from $P(A)$ and $P(B)$?
• Recall:
  \[ P(A,B) = P(B|A) P(A) \quad \text{(multiplication rule)} \]
• We are almost there: How does $P(B|A)$ relate to $P(B)$?
  $P(B|A) = P(B)$ iff $B$ and $A$ are independent!

• Examples:
  – Two coin tosses
  – Color shirt I’m wearing today, what Bill Clinton had for breakfast.
• Two events $A, B$ are independent from each other if
  $P(A,B) = P(A) P(B)$  \quad \text{Equivalent to } P(B) = P(B|A) \ (\text{if } P(A) \neq 0)$
• Otherwise they are dependent.
Joint Probability with Independence

Independence means we need far fewer numbers!

\[
P(\text{precipitation, temperature})
\]

\[
P(\text{precipitation}) P(\text{temperature})
\]

<table>
<thead>
<tr>
<th></th>
<th>sun</th>
<th>rain</th>
<th>sleet</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>20s</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>30s</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>40s</td>
<td>0.06</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>50s</td>
<td>0.06</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>60s</td>
<td>0.06</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>70s</td>
<td>0.07</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>80s</td>
<td>0.07</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>90s</td>
<td>0.08</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>100s</td>
<td>0.08</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>sun</th>
<th>rain</th>
<th>sleet</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>10s</td>
<td>0.5</td>
<td>0.3</td>
<td>0.05</td>
<td>0.15</td>
</tr>
<tr>
<td>20s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>30s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>40s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>50s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>60s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>70s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>80s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>90s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>100s</td>
<td>0.1</td>
<td>0.1</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

it takes 40 numbers  

it takes 14 numbers
Chain Rule

\[ P(A_1, A_2, A_3, A_4, \ldots A_n) = \]

\[ P(A_1|A_2, A_3, A_4, \ldots A_n) \]
\[ \times P(A_2, A_3, A_4, \ldots A_n) \]

Analogous to \( P(A,B) = P(A|B) \cdot P(B) \).
Chain Rule

\[ P(A_1, A_2, A_3, A_4, \ldots A_n) = \]
\[ P(A_1 | A_2, A_3, A_4, \ldots A_n) \]
\[ P(A_2 | A_3, A_4, \ldots A_n) \]
\[ P(A_3, A_4, \ldots A_n) \]
Chain Rule

\[
P(A_1, A_2, A_3, A_4, \ldots A_n) = \\
P(A_1 | A_2, A_3, A_4, \ldots A_n) \\
P(A_2 | A_3, A_4, \ldots A_n) \\
P(A_3 | A_4, \ldots A_n) \\
\ldots \\
P(A_n)
\]

Furthermore, if \( A_1 \ldots A_n \) are all independent from each other…
Chain Rule

If $A_1 \ldots A_n$ are all independent from each other

$$P(A_1, A_2, A_3, A_4, \ldots A_n) = P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot \ldots \cdot P(A_n)$$
Example: Two ways, same answer

• “A fair coin is tossed 3 times. What is the likelihood of 3 heads?”
  – Experiment: Toss a coin three times, 
    \( \Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\} \)
  – Event: basic outcome has exactly 3 H’s
    \( A = \{HHH\} \)

• Chain rule
  \[
  P(HHH) = P(H) P(H|H) P(H|HH) \\
  = P(H) P(H) P(H) = (1/2)^3 = 1/8
  \]

• Size of event spaces
  \[
  P(HHH) = \frac{|A|}{|\Omega|} = \frac{1}{8}
  \]
Collaborative Exercise

• Suppose one is interested in a rare syntactic construction, parasitic gaps, which occur on average once in 100,000 sentences. Peggy Linguist has developed a complicated pattern matcher that attempts to identify sentences with parasitic gaps. It's pretty good, but it's not perfect: if a sentence has a parasitic gap, it will say so with probability 0.95, if it doesn't it will wrongly say so with probability 0.005.

• Suppose the test says that a sentence contains a parasitic gap. What is the probability that this is true?
Finding most likely posterior event

• $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$  
  (for example, $P(\text{"lies"} = \text{Noun}|\text{"more lies ahead"})$

• Want to find most likely A given B, but $P(B)$ is sometimes a pain to calculate…

• $\arg \max_A P(B|A)P(A) = \arg \max_A \frac{P(B|A)P(A)}{P(B)}$

  because B is constant while changing A
Random Variables

• A *random variable* is a function $X : \Omega \rightarrow Q$
  – in general $Q=\mathbb{R}^n$, but more generally simply $Q=\mathbb{R}$
  – makes it easier to talk about numerical values related to event space

• Random variable is *discrete* if $Q$ is countable.
• Example: coin $Q=\{0,1\}$, die $Q=\{1,6\}$
• Called an *indicator variable* or *Bernoulli trial* if $Q \subseteq \{0,1\}$

• Example:
  – Suppose event space comes from tossing two dice.
  – We can define a random variable $X$ that is the sum of their faces
  – $X : \Omega \rightarrow \{2,\ldots,12\}$

Because a random variable has a numeric range, we can often do math more easily by working with values of the random variable than directly with events.
Probability Mass Function

- \( p(X=x) = P(A_x) \) where \( A_x = \{a \in \Omega : X(a)=x\} \)
- Often written just \( p(x) \), when \( X \) is clear from context. Write \( X \sim p(x) \) for “\( X \) is distributed according to \( p(x) \)”.
- In English:
  - Probability mass function, \( p \)…
  - maps some value \( x \) (of random variable \( X \)) to…
  - the probability random variable \( X \) taking value \( x \)
  - equal to the probability of the event \( A_x \)
  - this event is the set of all basic outcomes, \( a \), for which the random variable \( X(a) \) is equal to \( x \).
- Example, again:
  - Event space = roll of two dice; e.g. \( a=<2,5>, |\Omega|=36 \)
  - Random variable \( X \) is the sum of the two faces
  - \( p(X=4) = P(A_4), A_4 = \{<1,3>, <2,2>, <3,1>\}, P(A_4) = 3/36 \)

Random variables will be used throughout the *Introduction to Information Theory*, coming next class.
Expected Value

• ... is a weighted average, or mean, of a random variable
  \[ E[X] = \sum_{x \in X(\Omega)} x \cdot p(x) \]

• Example:
  – \( X \) = value of one roll of a fair six-sided die:
    \[ E[X] = (1+2+3+4+5+6)/6 = 3.5 \]
  – \( X = \) sum of two rolls...
    \[ E[X] = 7 \]

• If \( Y \sim p(Y=y) \) is a random variable, then any function \( g(Y) \)
  defines a new random variable, with expected value
  \[ E[g(Y)] = \sum_{y \in Y(\Omega)} g(y) \cdot p(y) \]

• For example,
  – let \( g(Y) = aY+b \), then \( E[g(Y)] = a \cdot E[Y] + b \)
  – \( E[X+Y] = E[X] + E[Y] \)
  – if \( X \) and \( Y \) are independent, \( E[XY] = E[X] \cdot E[Y] \)
Variance

- **Variance**, written $\sigma^2$
- Measures how consistent the value is over multiple trials
  - “How much on average the variable’s value differs from its mean.”
- $\text{Var}[X] = \mathbb{E}[(X-E[X])^2]$
Joint and Conditional Probabilities with Random Variables

• Joint and Conditional Probability Rules
  – Analogous to probability of events!
• Joint probability
  \[ p(x,y) = P(X=x, Y=y) \]
• **Marginal distribution** \( p(x) \) obtained from the joint \( p(x,y) \)
  \[ p(x) = \sum_{y} p(x,y) \] (by the total probability rule)
• Bayes Rule
  \[ p(x|y) = \frac{p(y|x) p(x)}{p(y)} \]
• Chain Rule
  \[ p(w,x,y,z) = p(z) p(y|z) p(x|y,z) p(w|x,y,z) \]
Parameterized Distributions

• Common probability mass functions with same mathematical form…
• …just with different constants employed.
• A family of functions, called a distribution.
• Different numbers that result in different members of the distribution, called parameters.
• $p(a;b)$
Binomial Distribution

- A discrete distribution with two outcomes \( \Omega = \{0, 1\} \) (hence bi-nomial)
- Make \( n \) experiments.
- “Toss a coin \( n \) times.”

- Interested in the probability that \( r \) of the \( n \) experiments yield 1.
- Careful! It’s not a uniform distribution. (\( q = \text{prob of H} \))

\[
p(R = r \mid n, q) = \binom{n}{r} q^r (1 - q)^{n-r}
\]

where
\[
\binom{n}{r} = \frac{n!}{(n-r)!r!}
\]
Pictures of Binomial Distribution

binomial (n,q):

b(10,0.1)

b(10,0.3)

b(10,0.5)

b(10,0.7)

b(10,0.9)

b(10,0.99)
Multinomial Distribution

• A discrete distribution with \( m \) outcomes \( \Omega = \{0, 1, 2, \ldots, m\} \)
• Make \( n \) experiments.
• Examples: “Roll a \( m \)-sided die \( n \) times.”
  “Assuming each word is independent from the next, generate an \( n \)-word sentence from a vocabulary of size \( m \).”

• Interested in the probability of obtaining counts \( c = c_1, c_2, \ldots, c_m \) from the \( n \) experiments.

\[
p(c \mid n, q) = \left( \frac{n!}{c_1!c_2!\ldots c_m!} \right) \prod_{i=1 \ldots m} (q_i)^{c_i}
\]

Unigram language model
Parameter Estimation

• We have been assuming that P is given, but most of the time it is unknown.
• So we assume a parametric family of distributions and estimate its parameters…

• …by finding parameter values most likely to have generated the observed data (evidence).
• …treating the parameter value as a random variable!

Not the only way of doing parameter estimation. This is maximum likelihood parameter estimation.
Maximum Likelihood Parameter Estimation
Example: Binomial

• Toss a coin 100 times, observe $r$ heads
• Assume a binomial distribution
  – Order doesn’t matter, successive flips are independent
  – One parameter is $q$ (probability of flipping a head)
  – Binomial gives $p(r|n,q)$. We know $r$ and $n$.
  – Find $\arg \max_q p(r|n, q)$
Maximum Likelihood Parameter Estimation

Example: Binomial

• Toss a coin 100 times, observe \( r \) heads
• Assume a binomial distribution
  – Order doesn’t matter, successive flips are independent
  – One parameter is \( q \) (probability of flipping a head)
  – Binomial gives \( p(r|n,q) \). We know \( r \) and \( n \).
  – Find \( \arg \max_q p(r|n, q) \)

(Notes for board)

\[
\text{likelihood} = p(R = r \mid n,q) = \binom{n}{r} q^r (1-q)^{n-r}
\]

\[
\log \text{ – likelihood} = L = \log(p(r \mid n,q)) \propto \log(q^r (1-q)^{n-r}) = r \log(q) + (n - r) \log(1-q)
\]

\[
\frac{\partial L}{\partial q} = \frac{r}{q} - \frac{n-r}{1-q} \Rightarrow r(1-q) = (n-r)q \Rightarrow q = \frac{r}{n}
\]

Our familiar ratio-of-counts is the maximum likelihood estimate!
Binomial Parameter Estimation Examples

- Make 1000 coin flips, observe 300 Heads
  - $P(\text{Heads}) = \frac{300}{1000}$
- Make 3 coin flips, observe 2 Heads
  - $P(\text{Heads}) = \frac{2}{3}$ ??
- Make 1 coin flips, observe 1 Tail
  - $P(\text{Heads}) = 0$ ???
- Make 0 coin flips
  - $P(\text{Heads}) =$ ???

- We have some “prior” belief about $P(\text{Heads})$ before we see any data.
- After seeing some data, we have a “posterior” belief.
Maximum A Posteriori Parameter Estimation

- We’ve been finding the parameters that maximize
  - $p(\text{data}|\text{parameters})$,
not the parameters that maximize
  - $p(\text{parameters}|\text{data})$ (parameters are random variables!)

- $p(q|n,r) = \frac{p(r|n,q) \ p(q|n)}{p(r|n)} = \frac{p(r|n,q) \ p(q)}{\text{constant}}$

- And let $p(q) = 2 \ q(1-q)$
Maximum A Posteriori Parameter Estimation

Example: Binomial

\[ \text{posterior} = p(r | n, q) p(q) = \binom{n}{r} q^r (1 - q)^{n-r} (2q(1-q)) \]

\[ \log - \text{posterior} = L \propto \log(q^{r+1}(1-q)^{n-r+1}) = (r + 1) \log(q) + (n - r + 1) \log(1 - q) \]

\[ \frac{\partial L}{\partial q} = \frac{(r + 1)}{q} - \frac{(n - r + 1)}{1 - q} \Rightarrow (r + 1)(1 - q) = (n - r + 1)q \Rightarrow q = \frac{r + 1}{n + 2} \]
Bayesian Decision Theory

• We can use such techniques for choosing among models:
  – Which among several models best explains the data?

• Likelihood Ratio

\[
\frac{P(\text{model1} \mid \text{data})}{P(\text{model2} \mid \text{data})} = \frac{P(\text{data} \mid \text{model1}) P(\text{model1})}{P(\text{data} \mid \text{model2}) P(\text{model2})}
\]
...back to our example: French vs English

- $p(\text{French} \mid \text{glacier, melange})$ versus $p(\text{English} \mid \text{glacier, melange})$?

- We have real data for
  - Shakespeare’s Hamlet
  - Charles Dickens’ Oliver Twist

- $p(\text{Hamlet} \mid \text{“hand”, “death”})$
  $p(\text{Oliver} \mid \text{“hand”, “death”})$
Continuing Homework Assignment?

• EC for whatever you hand in by tonight.

• For next week:
  – More time to create your own grammar
  – Modify parser to keep trace, and print parse trees
  – Try an additional grammar of your own creation, and investigate ambiguities
  – Work in small teams!
Training
data:

“Neural networks and other machine learning methods of classification…”

“Planning with temporal reasoning has been…”

“…based on the semantics of program dependence”

“Garbage collection for strongly-typed languages…”

“Multimedia streaming video for…”

“User studies of GUI…”

Testing
Document:

“Temporal reasoning for planning has long been studied formally. We discuss the semantics of several planning…”

Categories:

- Machine Learning
- Planning
- Prog. Lang. Semantics
- Garbage Collection
- Multimedia
- GUI
A Probabilistic Approach to Classification: “Naïve Bayes”

Pick the most probable class, given the evidence:

\[ P(c | d) = \frac{P(d | c) P(c)}{P(d)} \]

- a class (like “Planning”)
- a document (like “language intelligence proof...”)

Bayes Rule:

“Naïve Bayes”:

- the \( i \) th word in \( d \) (like “proof”)

Andrew McCallum, UMass Amherst