**Probability**

**Lecture #4**

**Introduction to Natural Language Processing**

**CMPSCI 585, Fall 2004**

**University of Massachusetts Amherst**

Andrew McCallum

---

**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”?

---

**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, Ω**, (set of all possible basic outcomes), e.g.
  - One coin toss, sample space Ω = { H, T }
  - Three coin tosses, Ω = { HHH, HHT, HTH, THH, THT, TTH, TTT }
  - Part-of-speech of a word, Ω = { CC, CD, CT, …, WRB }
  - Lottery tickets, |Ω| = 10^7
  - Next word in Shakespeare play, |Ω| = size of vocabulary
  - Number of words in your Ph.D. thesis, Ω = { 0, 1, …, ∞ }
  - Length of time of “a” sounds when I said “sample”.

---

**Events and Event Spaces**

- An **event, A**, is a set of basic outcomes, i.e., a subset of the sample space, Ω.
  - Intuitively, a question you could ask about an outcome.
  - Ω = { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT }
  - E.g. basic outcome = THH
  - E.g. event = “has exactly 2 H’s”, A={THH, HHT, HTH}
  - A=Ω is the certain event, A=∅ is the impossible event.
  - For “not A”, we write $\overline{A}$
- A common **event space, F**, is the power set of the sample space, Ω. (power set is written $2^Ω$)
  - 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 → [0,1]
  - P(Ω) = 1
  - Countable additivity: For disjoint events $A_i$ in F $P(\bigcup A_i) = \sum P(A_i)$
- We call $P(A)$ “the probability of event A”, well-defined **probability space** consists of
  - Sample space Ω
  - 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 \( P(A) = \frac{373}{1000} = 0.373 \)

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, \( F. \)
  - \( P : F \rightarrow [0,1] \)
  - 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:
  - \( P(\emptyset) = 0, \ P(\Omega) = 1 \)
  - \( A \subseteq B \Rightarrow P(A) \leq P(B), \)
  - \( \sum_a P(a) = 1, \) for \( a = \) basic outcome.

Joint and Conditional Probability

- Joint probability of \( A \) and \( B: \)
  \[ P(A \cap B) \text{ is usually written } P(A,B) \]
- Conditional probability of \( A \) given \( B: \)
  \[ P(A|B) = \frac{P(A,B)}{P(B)} \]

Joint Probability Table

<table>
<thead>
<tr>
<th>Precipitation, Temperature</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>0.07</td>
<td>0.03</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>

Conditional Probability Table

<table>
<thead>
<tr>
<th>Precipitation, Temperature</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>0.3</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>0.7</td>
<td>0.3</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>
**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)**
  \[ P(A) = P(A, B) + P(A, \overline{B}) \]
  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 \( P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \), and thus
  \[ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \]
- **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!
  
Same year Mozart wrote his symphony #1 in E-flat.

**Independence**

- Can we compute \( P(A, B) \) from \( P(A) \) and \( P(B) \)?
- Recall: \( P(A, B) = P(A|B) \cdot P(B) \) (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 a Bill Clinton had for breakfast.
- Two events \( A, B \) are **independent** from each other if
  \[ P(A, B) = P(A) \cdot P(B) \] (Equivalent to \( P(B) = P(B|A) \) if \( P(A) \neq 0 \))
  - Otherwise they are **dependent**.

**Joint Probability with Independence**

<table>
<thead>
<tr>
<th>Precipitation, Temperature</th>
<th>Precipitation</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>Rain</td>
<td>Sleet</td>
</tr>
<tr>
<td>10°</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>20°</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>30°</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td>40°</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>50°</td>
<td>0.08</td>
<td>0.03</td>
</tr>
<tr>
<td>60°</td>
<td>0.08</td>
<td>0.03</td>
</tr>
<tr>
<td>70°</td>
<td>0.07</td>
<td>0.03</td>
</tr>
<tr>
<td>80°</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>90°</td>
<td>0.06</td>
<td>0.03</td>
</tr>
</tbody>
</table>

**Chain Rule**

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

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

\[ P(A_1, A_2, A_3, A_4, \ldots, A_n) = \frac{P(A_1|A_2, A_3, A_4, \ldots, A_n) \cdot P(A_2|A_3, A_4, \ldots, A_n) \cdot P(A_3|A_4, \ldots, A_n) \cdots P(A_n)}{P(A_1)} \]

Furthermore, if \( A_1, \ldots, A_n \) are all independent from each other:

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

**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, \( \Omega = \{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\} \)
  - Event: basic outcome has exactly 3 H's
    \( A = \{\text{HHH}\} \)
  - Chain rule
    \[ P(\text{HHH}) = P(H) \cdot P(H|H) \cdot P(H|HH) = P(H) \cdot P(H) \cdot P(H) = (0.5)^3 = \frac{1}{8} \]
  - Size of event spaces
    \[ P(\text{HHH}) = \frac{|A|}{|\Omega|} = \frac{1}{8} \]

**Finding most likely posterior event**

- \( P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \) (for example, P("lies"|Noun)="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 \( X : W \rightarrow Q \)
  
- 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 \( Q = \{1,2,3,4,5,6\} \)

- Called an indicator variable or Bernoulli trial if \( Q \subset \{0,1\} \)

- Example:
  
  - Suppose event space comes from tossing two dice.
  
  - We can define a random variable \( X \) that is the sum of their faces
  
  - \( X : W \rightarrow \{2,3,4,5,6,7,8,9,10,11,12\} \)

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 \( A_x = \{ a \in W : X(a) = x \} \)
- Often written just \( p(x) \), when \( X \) is clear from context.
- 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. \( a = \langle 2,5 \rangle \), \( |W| = 36 \)
  - Random variable \( X \) is the sum of the two faces
  - \( p(X=4) = P(A_4) \), \( A_4 = \{ \langle 1,3 \rangle, \langle 2,2 \rangle, \langle 3,1 \rangle \} \), \( P(A_4) = 3/36 \)

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

Expected Value
- \( \ldots \) is a weighted average, or mean, of a random variable \( E[X] = \sum_{x \in X(W)} x \cdot p(x) \)
- 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 \( E[g(Y)] = \sum_{y \in Y(W)} g(y) \cdot p(y) \)
- For example,
  - let \( g(Y) = aY + b \), then \( E[g(Y)] = a E[Y] + b \)
  - \( E[X+Y] = E[X] + E[Y] \)
  - if \( X \) and \( Y \) are independent, \( E[XY] = E[X] E[Y] \)

Variance
- Variance, written \( s^2 \)
- Measures how consistent the value is over multiple trials
  - “How much on average the variable’s value differs from its mean.”
- \( \text{Var}[X] = E[(X-E[X])^2] \)
- Standard deviation = \( \sqrt{\text{Var}[X]} = s \)

Joint and Conditional Probabilities
- 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
    \( p(x|y) = p(y|x) p(x) / p(y) \)
  - 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...
  - \( \ldots \) 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 \( W = \{0, 1\} \) (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.
  - \( p(R=r|n,q) = \binom{n}{r} \cdot q^r \cdot (1-q)^{n-r} \)
  - where \( \binom{n}{r} = \frac{n!}{r! \cdot (n-r)!} \)
Pictures of Binomial Distribution

Multinomial Distribution

- A discrete distribution with \( m \) outcomes
  \( W = \{0, 1, 2, \ldots, m\} \)
- 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 \( c_1, c_2, \ldots, c_m \) from the \( n \) experiments.

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

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 \( \arg \max_q p(r | n, q) \)

\[
\text{likelihood} = p(R = r | n, q) = \frac{n!}{r! (n-r)!} q^r (1-q)^{n-r}
\]

\[
\log \text{likelihood} = \log(p(r | n, q)) = \log(n! | r! (n-r)! q^r (1-q)^{n-r}) = \log(n!) + r \log(q) + (n-r) \log(1-q)
\]

Our familiar ratio-of-counts is the maximum likelihood estimate!

Binomial Parameter Estimation Examples

- Make 1000 coin flips, observe 300 Heads
  \( \hat{p} = 300/1000 \)
- Make 3 coin flips, observe 2 Heads
  \( \hat{p} = 2/3 \)
- Make 1 coin flips, observe 1 Tail
  \( \hat{p} = 0 \)
- Make 0 coin flips
  \( \hat{p} = ??? \)

We have some “prior” belief about \( P(\text{Heads}) \) before we see any data.
After seeing some data, we have a “posterior” belief.
Bayesian 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!)

- \( p(q|r,n) = p(r|n,q) p(q|n) \)
  \( p(r|n) \) constant

- \( p(q) = 6q(1-q) \)

Bayesian Decision Theory

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

- Likelihood Ratio
  \[
  \frac{P(\text{model1} | \text{data})}{P(\text{model2} | \text{data})} = \frac{P(\text{data}|\text{model1}) P(\text{model1})}{P(\text{data}|\text{model2}) P(\text{model2})}
  \]

Maximum A Posteriori Parameter Estimation Example: Binomial

posterior = \( p(r|n,q)p(q) \)

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

Bayesian Parameter Estimation

A Probabilistic Approach to Classification: “Naïve Bayes”

Pick the most probable class, given the evidence:

\[
c^* = \arg \max_{c} P(c_j | d)
\]

Bayes Rule:

\[
P(c_j | d) = \frac{P(c_j) P(d | c_j)}{P(d)} = \frac{P(c_j) \prod_{i=1}^{m} P(w_{ij} | c_j)}{\sum_{c_i} P(c_i) \prod_{i=1}^{m} P(w_{ij} | c_i)}
\]

\( w_{ij} \) is the \( j \)th word in \( d \) (like “proof”)

Document Classification by Machine Learning

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

Training data:

Testing Document:

Categories:

“Planning with temporal reasoning has been...”
“Garbage collection for strongly-typed languages...”
“Multimedia streaming video for...”
“User studies of GUI...”