Probability

Lecture #2

Introduction to Natural Language Processing

CMPSCI 585, Spring 2004

University of Massachusetts Amherst

Andrew McCallum

Syllabus

- Probability and Information Theory
  - Spam filtering or email categorization
- Language models and clustering
  - Word sense disambiguation
- Hidden Markov models
  - Part-of-speech tagging
- SPRING BREAK
- Parsing
  - Build a Probabilistic Context Free Parser
- Information extraction
- Machine translation
- Semantics & discourse
- Final project!

Grading

- 32% Homework (4 programs & writeup)
- 25% Final project
- 10% Midterm Exam
- 15% Final Exam
- 5% Classroom participation
- 3% + E.C. "Pen/Pencil Quizzes"

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 }; *basic outcome* = H or T
  - three coin tosses, Ω = { HHH, HHT, HTH,…, TTT }
  - Part-of-speech of a word, Ω = { 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 $\Omega = \{ 0, 1, \ldots \infty \}$ 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, 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}
  - $A=\Omega$ is the certain event, $A=\varnothing$ is the impossible event.
  - For "not A", we write $\bar{A}$
- 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*) 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
    $P(\cup_j A_j) = \Sigma_j P(A_j)$
- 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,
    Ω = {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

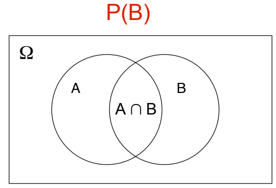
- "A *fair* coin is tossed 3 times. What is the likelihood of 2 heads?"
  - Experiment: Toss a coin three times,
    Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
  - Event: basic outcome has exactly 2 H's
    A = {THH, HTH, HHT}
- Assume a *uniform distribution* over outcomes
  - Each basic outcome is equally likely
  - P({HHH}) = P({HHT}) = … = P({TTT})
- $P(A) = |A| / |\Omega| = 3 / 8 = 0.375$

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
    $P(\cup_j A_j) = \Sigma_j P(A_j)$
- The above are axioms of probability theory
- Immediate consequences:
  - $P(\emptyset) = 0$, $P(\overline{A}) = 1 - P(A)$, $A \subseteq B \rightarrow P(A) \leq P(B)$,
    $\Sigma_{a \in \Omega} P(a) = 1$, for a = basic outcome.

Joint and Conditional Probability

- *Joint probability* of A and B:
  $P(A \cap B)$ 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)

	sun	rain	sleet	snow
10s	0.09	0.00	0.00	0.01
20s	0.08	0.00	0.00	0.02
30s	0.05	0.01	0.01	0.03
40s	0.06	0.03	0.01	0.00
50s	0.06	0.04	0.00	0.00
60s	0.06	0.04	0.00	0.00
70s	0.07	0.03	0.00	0.00
80s	0.07	0.03	0.00	0.00
90s	0.08	0.02	0.00	0.00
100s	0.08	0.02	0.00	0.00

it takes 40 numbers

Conditional Probability Table

What does it look like "under the hood"?

P(precipitation I temperature)

	sun	rain	sleet	snow
10s	0.9	0.0	0.0	0.1
20s	0.8	0.0	0.0	0.2
30s	0.5	0.1	0.1	0.3
40s	0.6	0.3	0.1	0.0
50s	0.6	0.4	0.0	0.0
60s	0.6	0.4	0.0	0.0
70s	0.7	0.3	0.0	0.0
80s	0.7	0.3	0.0	0.0
90s	0.8	0.2	0.0	0.0
100s	0.8	0.2	0.0	0.0

it takes 40 numbers

Two Useful Rules

- *Multiplication Rule*
  $P(A,B) = P(A|B)\,P(B)$

  (equivalent to conditional probability definition from previous slide)

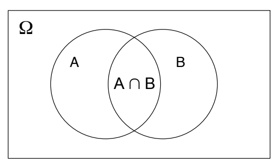
- *Total Probability Rule* (Sum Rule)
  $P(A) = P(A,B) + P(A,\bar{B})$
  or more generally, if B can take on *n* values
  $P(A) = \Sigma_{i=1..n}\, P(A,B_i)$

  (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…
- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

"Normalizing constant"



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

calculate P(AIB) in terms of P(BIA).

Reverend Thomas Bayes

1702 - 1761

- 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!
- *"Essay towards solving a problem in the doctrine of chances"* published in the *Philosophical Transactions of the Royal Society of London* in 1764.

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

Independence

- Can we compute P(A,B) from P(A) and P(B)?
- Recall:
  $P(A,B) = P(B|A)\,p(A)$ (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
  $P(A,B) = P(A)\,P(B)$ Equivalent to $P(B) = P(B|A)$ (if $P(A) \neq 0$)
- Otherwise they are dependent.

Joint Probability with Independence

Independence means we need far fewer numbers!

P(precipitation, temperature)

	sun	rain	sleet	snow
10s	0.09	0.00	0.00	0.01
20s	0.08	0.00	0.00	0.02
30s	0.05	0.01	0.01	0.03
40s	0.06	0.03	0.01	0.00
50s	0.06	0.04	0.00	0.00
60s	0.06	0.04	0.00	0.00
70s	0.07	0.03	0.00	0.00
80s	0.07	0.03	0.00	0.00
90s	0.08	0.02	0.00	0.00
100s	0.08	0.02	0.00	0.00

it takes 40 numbers

P(precipitation) P(temperature)

sun	rain	sleet	snow
0.5	0.3	0.05	0.15

10s	0.1
20s	0.1
30s	0.1
40s	0.1
50s	0.1
60s	0.1
70s	0.1
80s	0.1
90s	0.1
100s	0.1

it takes 14 numbers

Chain Rule

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

$P(A_1 | A_2, A_3, A_4, \ldots A_n)$

$P(A_2, A_3, A_4, \ldots A_n)$

Analogous to P(A,B) = P(A|B) 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

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

$P(A_1 | A_2, A_3, A_4, \ldots A_n)$

$P(A_2 | A_3, A_4, \ldots A_n)$

$P(A_3 | A_4, \ldots A_n)$

…

$P(A_n)$

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

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

$P(A_1)$

$P(A_2)$

$P(A_3)$

…

$P(A_n)$

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, Ω = {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) = $(0.5)^3$ = 1/8
- Size of event spaces
  P(HHH) = |A| / |Ω| = 1/8

Finding most likely posterior event

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

because B is constant while changing A

Random Variables

- A *random variable* is a function $X : \Omega \rightarrow Q$
  - in general $Q=\Re^n$, but more generally simply $Q=\Re$
  - 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 $Q \in \{0,1\}$
- Example:
  - Suppose event space comes from tossing two dice.
  - We can define a random variable X that is the sum of their faces
  - $X : \Omega \rightarrow \{2,..12\}$

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 ~ 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
  $E[X] = \Sigma_{x \in X(\Omega)} x \cdot p(x)$
- 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 ~ p(Y=y) is a random variable, then any function g(Y) defines a new random variable, with *expected value*
  $E[g(Y)] = \Sigma_{y \in Y(\Omega)} g(y) \cdot p(y)$
- 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 σ^2
- Measures how consistent the value is over multiple trials
  - "How much on average the variable's value differs from the its mean."
- $Var[X] = E[(X-E[X])^2]$
- *Standard deviation* = $\sqrt{Var[X]} = \sigma$

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)
  $p(x) = \Sigma_y p(x,y)$ (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…
- …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

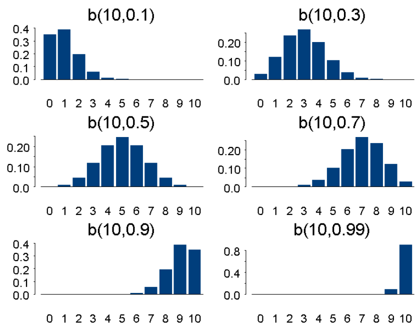
- A discrete distribution with *two* outcomes
  $\Omega = \{0, 1\}$ (hence *bi*-nomial)
- Make *n* experiments.
- "Toss a coin *n* times."

- Interested in the probability that *r* of the *n* experiments yield 1.
- Careful! It's not a *uniform* distribution.
- $p(R = r \mid n,q) = \binom{n}{r} q^r (1-q)^{n-r}$

  where $\binom{n}{r} = \frac{n!}{(n-r)!\,r!}$

Pictures of Binomial Distribution

binomial (n,q):



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 $\mathbf{c} = c_1, c_2, \ldots c_m$ from the *n* experiments.

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

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(Heads) = 300/1000
- Make 3 coin flips, observe 2 Heads
  - P(Heads) = 2/3 ??
- Make 1 coin flips, observe 1 Tail
  - P(Heads) = 0 ???
- Make 0 coin flips
  - P(Heads) = ???

- We have some "*prior*" belief about P(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(data|parameters),

  not the parameters that maximize
  - p(parameters|data) **(parameters are random variables!)**
- $p(q|n,r) = \dfrac{p(r|n,q)\,p(q|n)}{p(r|n)} = \dfrac{p(r|n,q)\,p(q)}{\text{constant}}$
- p(q) = 6 q(1-q)

Maximum A Posteriori Parameter Estimation Example: Binomial

$$\text{posterior} = p(r \mid n,q)\,p(q) = \binom{n}{r} q^r (1-q)^{n-r} (6q(1-q))$$

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

$$\frac{\partial L}{\partial q} = \frac{(r+1)}{q} - \frac{(n-r+1)}{1-q} \Rightarrow (r+1)(1-q) = (n-r+1)q \Rightarrow q = \frac{r+1}{n+2}$$

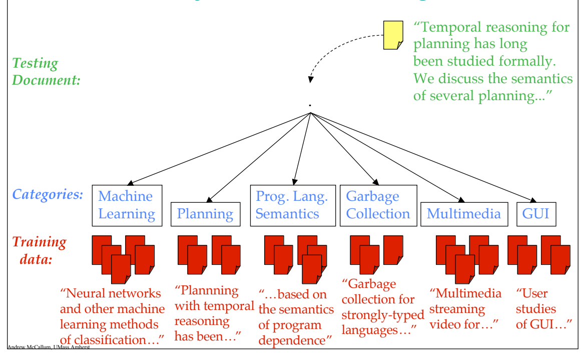
Bayesian Decision Theory

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

- Likelihood Ratio

$$\frac{P(\text{model1} \mid \text{data})}{P(\text{model2} \mid \text{data})} = \frac{P(\text{data}|\text{model1})\,P(\text{model1})}{P(\text{data}|\text{model2})\,P(\text{model2})}$$

Document Classification by Machine Learning



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 - a class (like "Planning")

d - a document (like "language intelligence proof...")

Bayes Rule:

$$\Pr(c_j \mid d) = \frac{\Pr(c_j)\Pr(d \mid c_j)}{\Pr(d)}$$

"Naïve Bayes":

$$\approx \frac{\Pr(c_j)\prod_{i=1}^{|d|}\Pr(w_{d_i} \mid c_j)}{\sum_{c_k}\Pr(c_k)\prod_{i=1}^{|d|}\Pr(w_{d_i} \mid c_k)}$$

w_{d_i} - the i th word in d (like "proof")