April 1, 2004

Manning & Schütze 2.1

Introduction/organization

Probability theory

Overview

- Introductions
- CSE talk post-mortem
- Administrivia
- Probability theory

Administrivia

- Goals of the course
- Course requirements:
  - Participation
  - In-class presentation (of 1-2 papers, on one day)
  - Term paper or project (3 credit students only)
- URL: http://courses.washington.edu/ling580g
- Email me re: topics you're particularly keen on presenting

Important dates

- 4/22 Term paper/project topic choice
- 5/13 Term paper outlines/term project specs
- 5/20 (optional) paper/project presentations
- 6/9 Final papers/projects due, 5pm

Probability Theory: Overview

- Probability spaces
- Conditional probability and (in)dependence
- Bayes' theorem
- Random variables
- Expectation and variance
- Joint and conditional distributions
- Determining P
- Standard distributions
- Bayesian statistics

Probability Theory: Overview

- Probability spaces
- Conditional probability and (in)dependence
- Bayes' theorem
- Random variables
- Expectation and variance
- Joint and conditional distributions
- Determining P
- Standard distributions
- Bayesian statistics

Probability spaces (1/3)

- Experiment (trial): the process by which an observation is made
- Basic outcome: a possible observation (first heads, then tails)
- Event (A): equivalence class of basic outcomes (one heads, one tails)
- Sample space (Ω): set of all possible outcomes (for a given experiment)
- $A \subseteq \Omega$

Probability spaces (2/3)

- Event space ($\mathcal{F}$): set of events
- (For technical reasons, $\mathcal{F}$ needs to be a σ-*field*, satisfied by making it the powerset of the sample space: $2^{\otimes}$)
- Probability: A number between 0 (impossibility) and 1 (certainty).
- Probability mass: Total probability available to distribute over a set of events.
- Probability function/probability distribution (P): Function from $\mathcal{F}$ to the interval [0,1] such that $P(\Omega) = 1$.

Probability spaces (3/3)

- $P(A)$: The probability of event A.
- Countable additivity (property of probability functions):

  For disjoint sets $A_j \in \mathcal{F}$:

$$P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$$

- Addition rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Well-founded probability space: Ω, $\mathcal{F}$ (σ-field) and P

Probability spaces: Example

- A fair coin is tossed 3 times. What is the chance of 2 heads?
- Ω = ?
- Probability distribution of outcomes with Ω is a uniform distribution.
- A = ?
- $P(A) = \frac{|A|}{|\Omega|} =?$
- ($|A|$ is the number of elements in the set A.)

Probability Theory: Overview

- Probability spaces
- Conditional probability and (in)dependence
- Bayes' theorem
- Random variables
- Expectation and variance
- Joint and conditional distributions
- Determining P
- Standard distributions
- Bayesian statistics

Conditional probability (1/2)

- Conditional probability of A given B ($P(A|B)$): The updated probability of an event given some knowledge.
- Prior probability of A: The probability before the knowledge is gained.
- Posterior probability of A: The new probability after.

Conditional probability (2/2)

- If $P(B) > 0$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Multiplication rule: Even if $P(B) = 0$,

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

- Chain rule:

  $P(A_1 \cap \ldots \cap A_n) =$
  $P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n| \cap_{i=1}^{n-1} A_i)$

Conditional probability: Example

- Experiment: Tossing three fair coins.
- A: Two heads, one tails
- Prior probability of A = ?
- B: The first coin came up heads
- $P(A|B)$ = ?

(In)dependence

- A and B are independent of each other if

$$P(A \cap B) = P(A)P(B)$$

- If A and B are independent and $P(B) \neq 0$, $P(A) = P(A|B)$.
- Conditional independence: A and B are conditionally independent given C when

$$P(A \cap B|C) = P(A|C)P(B|C)$$

- Examples?

Probability Theory: Overview

- Probability spaces
- Conditional probability and (in)dependence
- Bayes' theorem
- Random variables
- Expectation and variance
- Joint and conditional distributions
- Determining P
- Standard distributions
- Bayesian statistics

Bayes' theorem

- A means of calculating $P(B|A)$ in terms of $P(A|B)$.

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

- $P(B)$ functions as a normalizing constant.
- If B is given, we can ignore it.

$$\text{argmax}_B \frac{P(A|B)P(B)}{P(A)} = \text{argmax}_B P(A|B)P(B)$$

Partitions

$$P(A \cap B) = P(A|B)P(B)$$
$$P(A \cap \bar{B}) = P(A|\bar{B})P(\bar{B})$$

- By additivity:
- $P(A) = P(A \cap B) + P(A \cap \bar{B})$
- Substitute the above:

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$

- $\bar{B}$ and B split A into two disjoint parts.
- For some group of sets B_i which partition A ($A \subseteq \cup_i B_i$):

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

Another statement of Bayes' theorem

If $A \subseteq \cup_{i=1}^{n} B_i$, $P(A) > 0$, and $B_i \cap B_j = \phi$ for $i \neq j$, then:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1} nP(A|B_i)P(B_i)}$$

Bayes' theorem: Example (1/2)

- Parasitic gaps appear once in 100,000 sentences.
- If a sentence contains a pg, the pattern matcher will say so with probability 0.95 (0.05 probability of false negatives)
- If it doesn't, will wrongly say it does with probability 0.005 (false positives).
- If the pattern matcher says a sentence contains a parasitic gap (event T), what is the probability that this is true (event G)?
- $P(G|T)$ = ?

Bayes' theorem: Example (2/2)

$$P(G|T) = \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|\bar{G})P(\bar{G})}$$

$$P(G|T) = \frac{0.95 \times 0.00001}{0.95 \times 0.00001 + 0.005 \times 0.99999} \approx 0.002$$

Probability Theory: Overview

- Probability spaces
- Conditional probability and (in)dependence
- Bayes' theorem
- Random variables
- Expectation and variance
- Joint and conditional distributions
- Determining P
- Standard distributions
- Bayesian statistics

Random variables (1/2)

- Random variable: a function $X : \Omega \to \mathbb{R}^n$
- (commonly, n = 1)
- Used to model event spaces in a numerical domain, which is easier to manipulate mathematically.
- Conceptually, an abstract *stochastic process* which generates numbers with a certain probability distribution.

Random variables (2/2)

- Discrete random variable: $X : \Omega \to S$, where S is a countable subset of $\mathbb{R}$.
- (We're not worrying about continuous random variables right now.)
- Indicator random trial/Bernoulli trial: $X : \Omega \to \{0, 1\}$
- Example: If the experiment is tossing two dice, then the event space can be modeled by a discrete random variable $X : \Omega \to \{2, \ldots, 12\}$.

Probability mass functions (pmf) (1/2)

- Functions from possible values of the random variable to probabilities of that value being the result on any given experiment.

  $\mathrm{p}(x) = \mathrm{p}(X = x) = P(A_x)$

  Where $A_x = \{\omega \in \Omega : X(\omega) = x\}$

- $X \sim \mathrm{p}(x)$: The random variable X is distributed according to the pmf $\mathrm{p}(x)$

Probability mass functions (pmf) (2/2)

- $\mathrm{p}(x) > 0$ at only a countable number of points
- For discrete random variables:

$$\sum_i \mathrm{p}(x_i) = \sum_i P(A_{x_i}) = P(\Omega) = 1$$

- How are pmfs (p) different from probability functions (P)?

Probability Theory: Overview

- Probability spaces
- Conditional probability and (in)dependence
- Bayes' theorem
- Random variables
- Expectation and variance
- Joint and conditional distributions
- Determining P
- Standard distributions
- Bayesian statistics

Expectation (1/2)

- Expectation $E(X)$: the mean of a random variable X
- For X with pmf $\mathrm{p}(x)$ such that $\sum_x |x|\mathrm{p}(x) < \infty$,

$$E(X) = \sum_x x\mathrm{p}(x)$$

- Example: If rolling one die and Y is the value on its face, what is $E(Y)$?

Expectation (2/2)

- If $Y \sim \mathrm{p}(y)$ is a random variable, any function $g(Y)$ defines a new random variable.
- If $E(g(Y))$ is defined, then:

$$E(g(Y)) = \sum_y g(y)\mathrm{p}(y)$$

- $E(X+Y) = E(X) + E(Y)$
- If X and Y are independent:

$$E(XY) = E(X)E(Y)$$

Variance

- Variance (Var(X),σ^2): A measure of how much values of a random variable tend to vary across trials.

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$

- Standard deviation (σ): The square root of the variance.

Variance: Example

- Example: What are the expectation and variance of X, the sum of the numbers on two dice?

$$E(X) = E(Y + Y) = E(Y) + E(Y) = 7.$$

$$\mathrm{Var}(X) = E((X - E(X))^2) = \sum_x \mathrm{p}(x)(x - E(X))^2 = 5\frac{5}{6}$$

- (NB: Not the same as the variance for an 11 sided die.)

Probability Theory: Overview

- Probability spaces
- Conditional probability and (in)dependence
- Bayes' theorem
- Random variables
- Expectation and variance
- Joint and conditional distributions
- Determining P
- Standard distributions
- Bayesian statistics

Joint and conditional distributions: Joint pmfs

- Joint pmf for discrete random variables X, Y:

$$\mathrm{p}(x, y) = P(X = x, Y = y)$$

- Marginal pmfs, looking at each of the variables separately:

$$\mathrm{p}_X(x) = \sum_y \mathrm{p}(x, y) \qquad \mathrm{p}_Y(y) = \sum_x \mathrm{p}(x, y)$$

- If X and Y are independent:

$$\mathrm{p}(x, y) = \mathrm{p}_X(x)\mathrm{p}_Y(y)$$

Joint and conditional distributions: Conditional pmfs

- Conditional pmf: For y such that $\mathrm{p}_Y(y) > 0$,

$$\mathrm{p}(x|y) = \frac{\mathrm{p}(x, y)}{\mathrm{p}_Y(y)}$$

- Chain rule in terms of random variables:

$$\mathrm{p}(w, x, y, z) = \mathrm{p}(w)\mathrm{p}(x|w)\mathrm{p}(y|w, x)\mathrm{p}(z|w, x, y)$$

- What's new here compared to the other chain rule?

Probability Theory: Overview

- Probability spaces
- Conditional probability and (in)dependence
- Bayes' theorem
- Random variables
- Expectation and variance
- Joint and conditional distributions
- Determining P
- Standard distributions
- Bayesian statistics

Determining P

- For (imaginary, perfectly fair) coins, dice, etc: logic.
- For natural language phenomena??
- $\rightarrow$ Estimate P on the basis of frequency counts.

$$\frac{C(u)}{N}$$

- (The fact that counts tend to stabilize after a large number of trials supports doing this.)

Parametric v. non-parametric estimation

- Parametric: Assume that P is suitably modeled by a well-known family of distributions, and figure out (empirically) which member of that family is the closest fit. This involves setting (numeric) parameters.
- Parametric approaches require less data (less counting)
- Non-parametric: Make no assumptions about the underlying distribution, and do a lot of counting.
- Non-parametric approaches require more data.
- ... and often end up needing to assume things about the distribution in order to handle data sparsity.

Next time

- Whatever we didn't get to today, plus:
  - Standard distributions
  - Bayesian statistics
- And then elementary information theory (2.2)