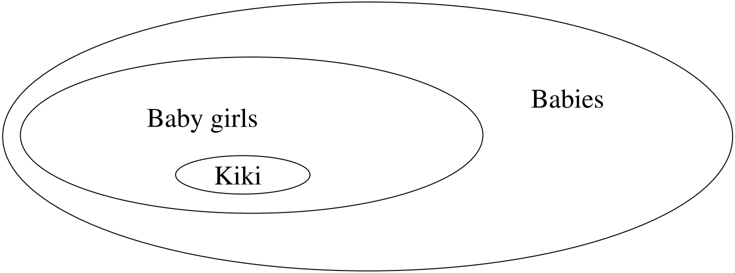
CMPT-413
Computational Linguistics

Anoop Sarkar
http://www.cs.sfu.ca/∼anoop

February 12, 2007

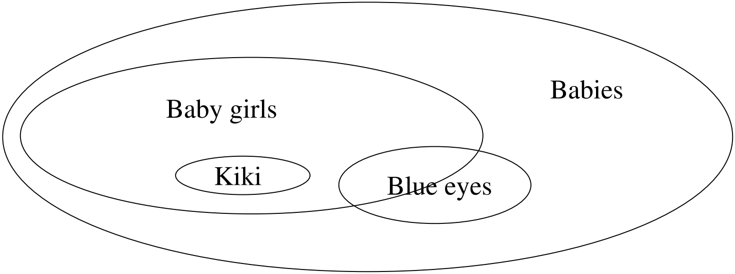
Quick guide to probability theory

- P(X) means probability that X is true
  - P(baby is a girl) $= 0.5$
    percentage of total number of babies that are girls
  - P(baby girl is named Kiki) $= 0.001$
    percentage of total number of babies that are named Kiki



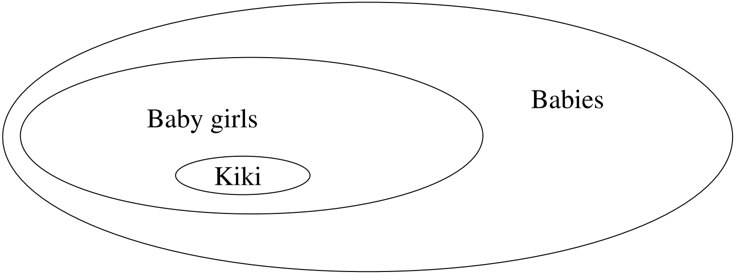
Joint probability

- P(X,Y) means probability that X and Y are both true
  - P(baby girl, blue eyes) percentage of total number of babies that are girls and have blue eyes



Conditional probability

- $P(X \mid Y)$ means probability that X is true when we already know that Y is true
  - $P(\text{baby is named Kiki} \mid \text{baby is a girl}) = 0.002$
  - $P(\text{baby is a girl} \mid \text{baby is named Kiki}) = 1$

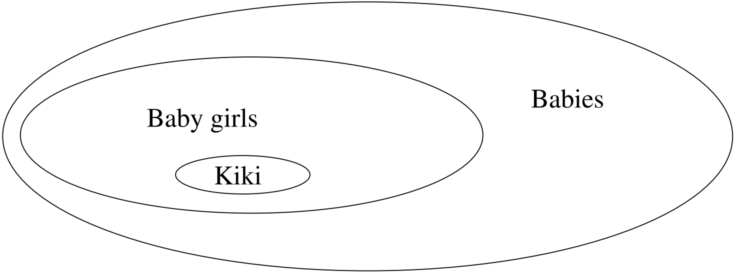


Conditional probability

- Conditional and joint probabilities are related:

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

  - $P(\text{baby is named Kiki} \mid \text{baby is a girl}) = \frac{P(\text{baby is a girl}, \text{baby is named Kiki})}{P(\text{baby is a girl})} = \frac{0.001}{0.5} = 0.002$

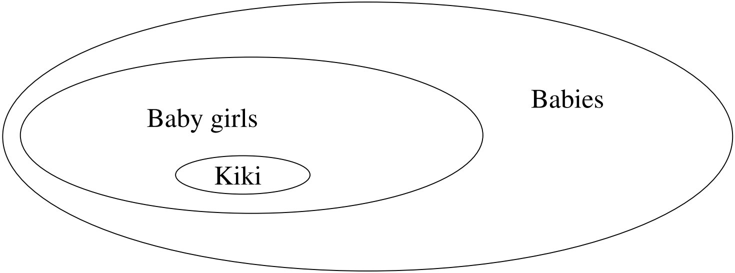


Bayes rule

- Conditional probability re-written as likelihood times prior:

$$P(X \mid Y) = \frac{P(Y \mid X) \times P(X)}{P(Y)}$$

  - $P(\text{named Kiki} \mid \text{girl}) = \frac{P(\text{girl}|\text{named Kiki}) \times P(\text{named Kiki})}{P(\text{girl})} = \frac{1.0 \times 0.001}{0.5} = 0.002$



Bayes Rule

$$
\begin{aligned}
P(X \mid Y) &= \frac{P(X, Y)}{P(Y)} && (1)\\
P(Y \mid X) &= \frac{P(Y, X)}{P(X)} && (2)\\
P(X, Y) &= P(Y, X) && (3)\\
P(X \mid Y) \times P(Y) &= P(Y \mid X) \times P(X) && (4)\\
P(X \mid Y) &= \frac{P(Y \mid X) \times P(X)}{P(Y)} && (5)\\
P(X \mid Y) &= P(Y \mid X) \times P(X) && (6)
\end{aligned}
$$

Basic Terms

- $P(e)$ – *a priori* probability or just *prior*
- $P(f \mid e)$ – *conditional* probability. The chance of f given e
- $P(e, f)$ – *joint* probability. The chance of e and f both happening.
- If e and f are *independent* then we can write $P(e, f) = P(e) \times P(f)$
- If e and f are not *independent* then we can write
  $P(e, f) = P(e) \times P(f \mid e)$
  $P(e, f) = P(f) \times$?

Basic Terms

- Addition of integers:

$$\sum_{i=1}^{n} i = 1 + 2 + 3 + \ldots + n$$

- Product of integers:

$$\prod_{i=1}^{n} i = 1 \times 2 \times 3 \times \ldots \times n$$

- Factoring:

$$\sum_{i=1}^{n} i \times k = k + 2k + 3k + \ldots + nk = k \sum_{i=1}^{n} i$$

Probability: Axioms

- P measures total probability of a set of events
  - $P(\emptyset) = 0$
  - $P(\text{all events}) = 1$
  - $P(X) \leq P(Y)$ for any $X \subseteq Y$
  - $P(X) + P(Y) = P(X \cup Y)$ provided that $X \cap Y = \emptyset$
- P(GC drives drunk & GC is in Hawaii) + P(GC drives drunk & GC is not in Hawaii) = P(GC drives drunk)

Probability Axioms

- All events sum to 1:

$$\sum_e P(e) = 1$$

- Conditional probability:

$$\sum_e P(e \mid f) = 1$$

- Computing $P(f)$ from axioms:

$$P(f) = \sum_e P(e) \times P(f \mid e)$$

Probability: Bias and Variance

- P(GC drives drunk | GC is in Hawaii, GC is alone, GC is low in polls, ...)
- As we add more material to the right of | :
  - probability could increase or decrease
  - probability usually gets more relevant (less **bias**)
  - probability usually gets less reliable (more **variance**)
  - removing items from the right of | makes it easier to get an estimate (more bias but less variance)

Probability: The Chain Rule

- P(GC is in Hawaii,GC is alone,GC is low in polls | GC drives drunk)
- We cannot remove items from the left of |
  (verify that it violates the definitions we have given based on sets)
- In this case we can use the chain rule of probability to rescue us
- P(GC in Hawaii,GC alone,GC low in polls | GC drives drunk) $=$
  P(GC in Hawaii | GC alone,GC low in polls,GC drives drunk) $\times$
  P(GC alone | GC low in polls, GC drives drunk) $\times$
  P(GC low in polls | GC drives drunk)

Probability: The Chain Rule

- P(GC in Hawaii,GC alone,GC low in polls | GC drives drunk) =
  P(GC in Hawaii | GC alone,GC low in polls,GC drives drunk) $\times$
  P(GC alone | GC low in polls, GC drives drunk) $\times$
  P(GC low in polls | GC drives drunk)
- Remember: $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$
- $\frac{HALD}{D} = \frac{HALD}{ALD} \times \frac{ALD}{LD} \times \frac{LD}{D}$
  (simply cancel out the matching terms)

Probability: The Chain Rule

- $P(e_1, e_2, \ldots, e_n) = P(e_1) \times P(e_2 \mid e_1) \times P(e_3 \mid e_1, e_2) \ldots$

$$P(e_1, e_2, \ldots, e_n) = \prod_{i=1}^{n} P(e_i \mid e_{i-1}, e_{i-2}, \ldots, e_1)$$

Probability: Random Variables and Events

- What is y in $P(y)$?
- Shorthand for value assigned to a random variable Y, e.g. $Y = y$
- y is an element of some implicit **event space**: $\mathcal{E}$

Probability: Random Variables and Events

- The *marginal probability* $P(y)$ can be computed from $P(x, y)$ as follows:

$$P(y) = \sum_{x \in \mathcal{E}} P(x, y)$$

- Finding the value that maximizes the probability value:

$$\hat{x} = \underset{x \in \mathcal{E}}{\arg\max} \; P(x)$$

Log Probability Arithmetic

- Practical problem with tiny $P(e)$ numbers: underflow
- One solution is to use log probabilities:

$$\begin{aligned} \log(P(e)) &= \log(p_1 \times p_2 \times \ldots \times p_n) \\ &= \log(p_1) + \log(p_2) + \ldots + \log(p_n) \end{aligned}$$

- Note that:

$$x = \exp(\log(x))$$

- Also more efficient: addition instead of multiplication

Log Probability Arithmetic

p	$\log(p)$
0.0	$-\infty$
0.1	-3.32
0.2	-2.32
0.3	-1.74
0.4	-1.32
0.5	-1.00
0.6	-0.74
0.7	-0.51
0.8	-0.32
0.9	-0.15
1.0	-0.00

Log Probability Arithmetic

- So: $(0.5 \times 0.5 \times \ldots 0.5) = (0.5)^n$ might get too small but $(-1 - 1 - 1 - 1) = -n$ is manageable
- Another useful fact when writing code ($\log_2$ is *log to the base 2*):

$$\log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}$$

Log Probability Arithmetic

- Adding probabilities is expensive to compute: $logadd(x, y) = \log(\exp(x) + \exp(x))$
- A more efficient soln, let *big* be a large constant e.g. 10^{30}:

  function $logadd(x, y)$: # returns $\log(\exp(x) + \exp(y))$
  if $(y - x) > \log(big)$ return y
  elsif $(x - y) > \log(big)$ return x
  else return
  $min(x, y) + \log(\exp(x - min(x, y)) + \exp(y - min(x, y)))$
  endif
- There is a more efficient way of computing $\log(\exp(x - min(x, y)) + \exp(y - min(x, y)))$

Log Probability Arithmetic

function $\mathit{logadd}(x, y)$:
    if $(y - x) > \log(\mathit{big})$ return y
    elsif $(x - y) > \log(\mathit{big})$ return x
    elsif $(x \geq y)$ return $x + \log(1 + \exp(y - x))$
        # note that $max(x, y) = x$ and $y - x \leq 0$
    else return $y + \log(\exp(x - y) + 1)$
        # note that $max(x, y) = y$ and $x - y \leq 0$
    endif

Also, in ANSI C, `log1p` efficiently computes $\log(1 + x)$

`http://www.ling.ohio-state.edu/~jansche/src/logadd.c`

Information Theory

- Information theory is the use of probability theory to quantify and measure "information".
- Consider the task of efficiently sending a message. Sender Alice wants to send several messages to Receiver Bob. Alice wants to do this as efficiently as possible.
- Let's say that Alice is sending a message where the entire message is just one character a, e.g. $aaaa\ldots$. In this case we can save space by simply sending the length of the message and the single character.

Information Theory

- Now let's say that Alice is sending a completely random signal to Bob. If it is random then we cannot exploit anything in the message to compress it any further.
- The *lower bound* on the number of bits it takes to transmit some infinite set of messages is what is called entropy. This formulation of entropy by Claude Shannon was adapted from thermodynamics.
- Information theory is built around this notion of message compression as a way to evaluate the amount of information.

Entropy

- Consider a random variable X
- Entropy of X is:

$$H(X) = -\sum_{x \in \mathcal{E}} p(x) \log_2 p(x)$$

- Any base can be used for the log, but base 2 means that entropy is measured in bits.
- Entropy answers the question: How many bits are needed to transmit messages from event space $\mathcal{E}$, where $p(x)$ defines the probability of observing $X = x$.

Entropy

- Alice wants to bet on a horse race. She has to send a message to her bookie Bob to tell him which horse to bet on.
- There are 8 horses. One encoding scheme for the messages is to use a number for each horse. So in bits this would be $001, 010, \ldots$
  (lower bound on message length $=$ 3 bits in this encoding scheme)
- Can we do better?

Entropy

Horse 1	$\frac{1}{2}$	Horse 5	$\frac{1}{64}$
Horse 2	$\frac{1}{4}$	Horse 6	$\frac{1}{64}$
Horse 3	$\frac{1}{8}$	Horse 7	$\frac{1}{64}$
Horse 4	$\frac{1}{16}$	Horse 8	$\frac{1}{64}$

- If we know how likely we are to bet on each horse, say based on the horse's probability of winning, then we can do better.
- Let X be a random variable over the horse (chances of winning). The entropy of X is $H(X)$

Entropy

$$
\begin{aligned}
H(X) &= \\
&= -\sum_{i=1}^{8} p(i)\log_2 \; p(i) \\
&= -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{16}\log_2\frac{1}{16} + 4(\frac{1}{64}\log_2\frac{1}{64})\right) \\
&= -\left(\frac{1}{2}\times -1 + \frac{1}{4}\times -2 + \frac{1}{8}\times -3 + \frac{1}{16}\times -4 + 4(\frac{1}{64}\times -6)\right) \\
&= -\left(-\frac{1}{2} - \frac{1}{2} - \frac{3}{8} - \frac{1}{4} - \frac{3}{8}\right) \\
&= 2 \; bits
\end{aligned}
$$

- e.g., most likely horse gets code 0, then $10, 110, 1110, \ldots$ What happens when the horses are equally likely to win?

Perplexity

- The value 2^H is called **perplexity**
- Perplexity is the weighted average number of choices a random variable has to make.
- Choosing between 8 equally likely horses (H=3) is $2^3 = 8$.
- Choosing between the biased horses from before (H=2) is $2^2 = 4$.

Cross Entropy

- In real life, we cannot know for sure the exact winning probability for each horse. Let's say p_t is the true probability and p_e is our estimate of the true probability (say we got p_e by observing a limited number of previous races with these horses)
- Cross entropy is a distance measure between p_t and p_e.

$$H(p_t, p_e) = -\sum_{x \in \mathcal{E}} p_t(x) \log_2 p_e(x)$$

- Cross entropy is an upper bound on the entropy:

$$H(p) \leq H(p, m)$$

Relative Entropy or Kullback-Leibler distance

- Another distance measure between two probability functions p and q is:
$$KL(p\|q) = \sum_{x \in \mathcal{E}} p(x) log_2 \frac{p(x)}{q(x)}$$
- KL distance is asymmetric (not a *true* distance), that is: $KL(p, q) \neq KL(q, p)$

Conditional Entropy and Mutual Information

- *Entropy* of a random variable X:

$$H(X) = -\sum_{x \in \mathcal{E}} p(x) \log_2 p(x)$$

- *Conditional Entropy* between two random variables X and Y:

$$H(X \mid Y) = -\sum_{x,y \in \mathcal{E}} p(x,y) \log_2 p(x \mid y)$$

- *Mutual Information* between two random variables X and Y:

$$I(X;Y) = KL(p(x,y) \| p(x)p(y)) = \sum_{x} \sum_{y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$