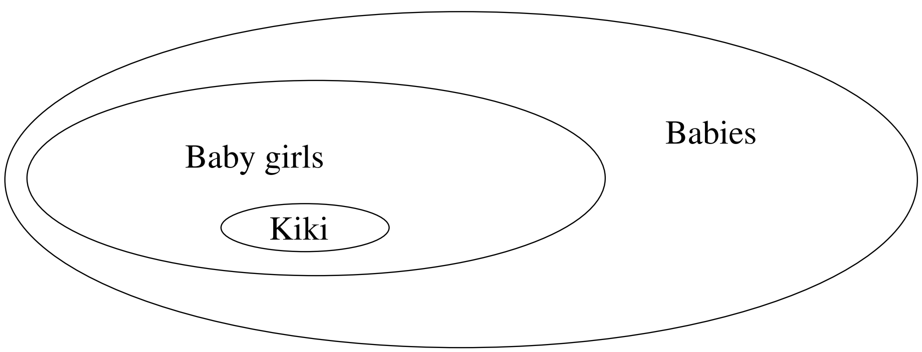
CMPT-413: Computational Linguistics

Anoop Sarkar
anoop@cs.sfu.ca
www.sfu.ca/˜anoop/courses/CMPT-413-Spring-2003.html

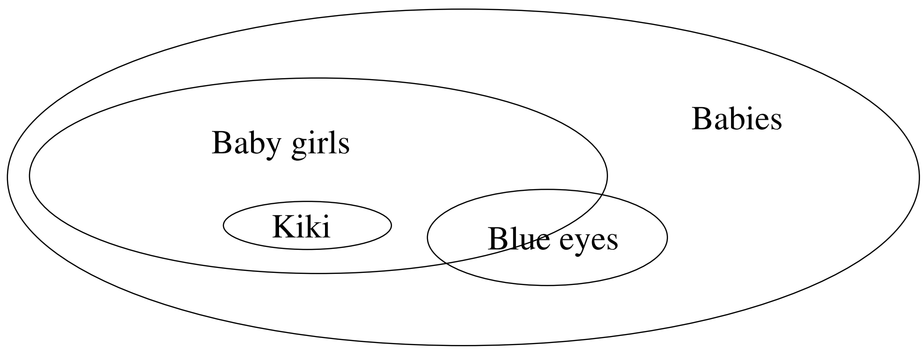
Everything you need to know about probability

- 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



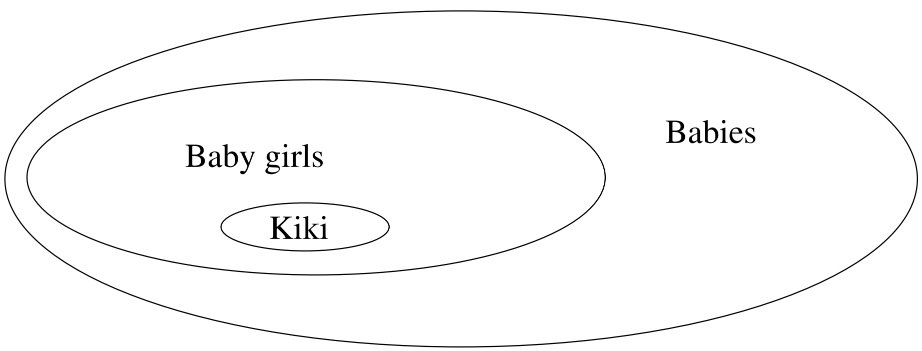
Joint probability

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

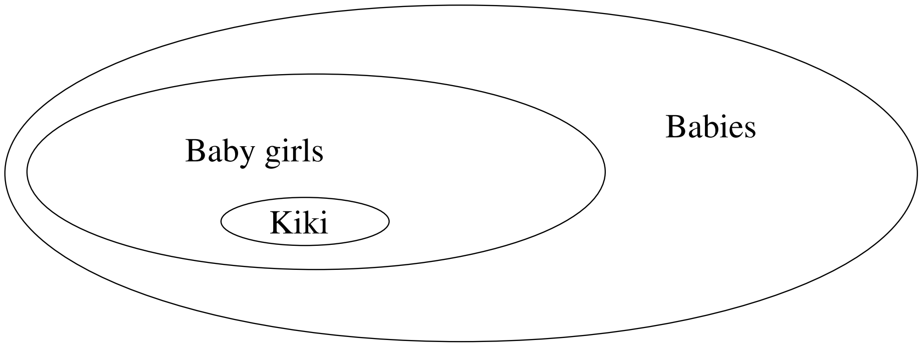


Conditional probability

- Conditional and joint probabilities are related:

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

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

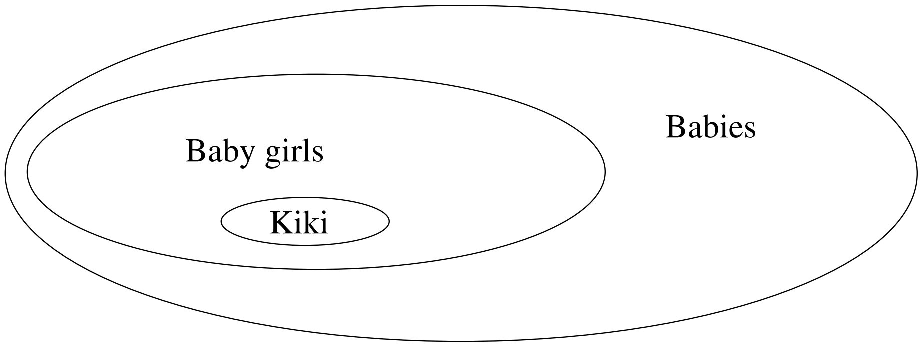


Bayes rule

- 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

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} \tag{1}$$

$$P(Y \mid X) = \frac{P(Y,X)}{P(X)} \tag{2}$$

$$P(X,Y) = P(Y,X) \tag{3}$$

$$P(X \mid Y) \times P(Y) = P(Y \mid X) \times P(X) \tag{4}$$

$$P(X \mid Y) = \frac{P(Y \mid X) \times P(X)}{P(Y)} \tag{5}$$

$$P(X \mid Y) = P(Y \mid X) \times P(X) \tag{6}$$

Probability: What does it really mean?

- P(GC drinks and drives $|$ GC is in Hawaii) $= 0.9$
  - GC drove drunk 90% of the time when in Hawaii
  - If GC visited Hawaii infinitely many times . . .
  - I would bet \$90 to win \$100 (strength of belief)
  - Just the output of a computation based on sets

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: 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 $\mid$ GC drives drunk) =
  P(GC in Hawaii $\mid$ GC alone, GC low in polls, GC drives drunk) $\times$
  P(GC alone $\mid$ GC low in polls, GC drives drunk) $\times$
  P(GC low in polls $\mid$ 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:

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

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*. . .. In this case we can save space by simply sending the length of the message and the single character.

- 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. Note that this is a very abstract notion and applies to many situations other than the examples given here.

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 $\mathtt{001}, \mathtt{010}, \ldots$ (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:

$$
\begin{aligned}
H(X) = \\
&= -\sum_{i=1}^{8} p(i)\log_2 p(i) \\
&= -\frac{1}{2}\log_2 \frac{1}{2} - \frac{1}{4}\log_2 \frac{1}{4} - \frac{1}{8}\log_2 \frac{1}{8} - \frac{1}{16}\log_2 \frac{1}{16} - 4(\frac{1}{64}\log_2 \frac{1}{64}) \\
&= -\frac{1}{2} \times -1 - \frac{1}{4} \times -2 - \frac{1}{8} \times -3 - \frac{1}{16} \times -4 - 4(\frac{1}{64} \times -6) \\
&= -(-\frac{1}{2} - \frac{1}{2} - \frac{3}{8} - \frac{1}{4} - \frac{3}{8}) \\
&= 2 \textit{ bits}
\end{aligned}
\tag{7}
$$

- 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:

$$H(X) = -\sum_{x \in \mathcal{E}} p(x) \mathrm{log}_2 p(x)$$

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

$$H(X \mid Y) = -\sum_{x,y \in \mathcal{E}} p(x,y) \mathrm{log}_2 p(x \mid y)$$

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

$$I(X;Y) = KL(p(x,y) \| p(x)p(y)) = \sum_x \sum_y p(x,y) \mathrm{log}_2 \frac{p(x,y)}{p(x)p(y)}$$