# CS 630
# Basic Probability and Information Theory

Tim Campbell

21 January 2003

# Probability Theory

- Probability Theory is the study of how best to predict outcomes of events.

- An experiment (or trial or event) is a process by which observable results come to pass.

- Define the set D as the space in which experiments occur.

- Define $\mathcal{F}$ to be a collection of subsets of D including both D and the null set. $\mathcal{F}$ must have closure under finite intersection and union operations and complements.

- A probability function (or distribution) is a function P:$\mathcal{F} \rightarrow [\prime, \infty]$ such that $P(D) = 1$ and for disjoint sets $A_i \in \mathcal{F}$ it must be that $P(\bigcup_{\forall i} A_i) = \sum_{\forall i} P(A_i)$.

- A probability space consists of a sample space D, a set $\mathcal{F}$, and a probability function P.

## Continuous Spaces

- The discussion being presented is given in discrete spaces, but they carry over to continuous spaces.

- Probability density functions are zero for any finite union of points, $P(D) = \int_D p(u)du = 1$ and $P * event) = \int_{event} p(u)du$

## Conditional Probability

- Conditional Probability is the (possibly) changed probability of an event given some knowledge.

- Prior Probability of an event is an event's probability before new knowledge is considered.

- Posterior Probability is the new probability resulting from use of new knowledge.

- Conditional probability of event A given B has happened is:

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

- This generalizes to the chain rule:

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

- If events A and B are independent of each-other then $P(A|B) = P(A)$ and $P(B|A) = P(A)$ so it follows that $P(A \cap B) = P(A)P(B)$

- Events A and B are conditionally independent given event C if

  $$P(A,B,C) = P(A,B|C)P(C) = P(A|C)P(B|C)P(C)$$

## Bayes' Theorem

- Bayes' theorem:

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

- The denominator $P(A)$ can be thought of as a normalizing constant and ignored if one is just trying to find a most likely event given A.

- More generally if $\mathcal{B}$ is a group of sets that are disjoint and partition A then

$$P(B|A) = \frac{P(A|B)P(B)}{\sum_{B_i \in \mathcal{B}} P(A|B_i)P(B_i)}$$

## Random Variables

- A random variable is a function $X : D \to \Re^n$

- The probability mass function is defined as

$$p(x) = p(X = x) = P(A_x)$$

  where

$$A_x = |a \in D : X(a) = x|$$

- Expectation is defined as

$$E(x) = \sum_x xp(x)$$

- Variance is defined as

  $Var(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$

- Standard Deviation is defined as the square root of variance.

- Joint probability distributions are possible using many random variables over a sample space. A joint probability mass function is defined $p(x, y) = P(A_x, B_x)$

- Marginal probability mass functions total up the probability masses for the values of each variable separately, for example, $p_x(x) = \sum_y p(x, y)$

- Conditional probability mass function is defined

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_y(y)} p_y(y) > 0$$

- The chain rule for random variables follows

$$p(w, x, y, z) = p(w)p(x|w)p(y|w, x)p(z|w, x, y)$$

## Determining P

- The function P is not always easy to obtain. Methods of construction include Relative Frequency, Parametric construction, and empirical estimation.

- Uniform distribution has the same value for all points in the domain.

- Binomial distribution is the result of a series of Bernoulli trials.

- Poisson distribution distributes points in such a way that the expected number of points in an interval is proportional to the length of the interval.

- Normal distribution or Gaussian distribution.

## Bayesian Statistics

- Bayesian Statistics integrates prior beliefs about probabilities into observations using Bayes' theorem.

- Example: Consider the toss of a possibly unbalanced coin. A sequence of flips s gives i heads and j tails and $\mu_m$ is a model in which P(h) = m, then

  $$P(s|\mu_m) = m^i(1-m)^j$$

  Now suppose the prior belief is modeled by $P(\mu_m) = 6m(1-m)$ which is centered on .5 and integrates to 1. Bayes' theorem gives

  $$P(\mu_m|s) = \frac{P(s|\mu_m)P(\mu_m)}{P(s)} = \frac{6m^{i+1}(1-m)^{i+1}}{P(s)}$$

  P(s) is a marginal probability, which means summing $P(s|\mu_m)$ weighted by $P(\mu_m)$:

  $$P(s) = \int_0^1 P(s|\mu_m)P(\mu_m)dm = \int_0^1 6m^{i+1}(1-m)^{i+1}dm$$

- Bayesian Updating is a process in which the above technique can be used regularly to update beliefs as new data become available.

- Bayesian Decision Theory is a method by which multiple models can be evaluated. Given two models $\mu$ and $v$, $P(\mu|s) = \frac{P(s|\mu)P(\mu)}{P(s)}$ and $P(v|s) = \frac{P(s|v)P(v)}{P(s)}$. The likelihood ratio between these models is

  $$\frac{P(\mu|s)}{P(v|s)} = \frac{P(s|\mu)P(\mu)}{P(s|v)P(v)}$$

  If the ratio is greater than 1 then $\mu$ is preferable, otherwise $v$ is preferable.

## Information Theory

- Developed by Claude Shannon

- Addresses the questions of maximizing data compression and transmission rate for any source of information and any communication channel.

## Entropy

- Entropy measures the amount of information in a random variable and is defined

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x) = E(\log_2 \frac{1}{p(x)})$$

- Joint Entropy of a pair of discrete random variables X and Y is defined

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

- Conditional Entropy of a random variable Y given X expresses the amount of information needed to communicate Y if X is already universally known.

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X=x) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y|x)$$

- The chain rule for entropy is defined

$$H(X_1, ..., X_n) = H(X_1) + H(X_2|X_1) + ... + H(X_n|X_1, ..., X_{n-1})$$

## Mutual Information

- Mutual Information is the reduction in uncertainty of a random variable caused by knowing about another. Using the chain rule for $H(X,Y)$,

  $$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

  Denote mutual information for random variables $X$ and $Y$ $I(X;Y)$,

  $$I(X;Y) = H(X) - H(X|Y)$$

  $$= H(X) + H(Y) - H(X,Y)$$

  $$= \sum_{x \in X, y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

- Conditional mutual information is defined:

  $$I(X;Y|Z) = I((X;Y)|Z) = H(X|Z) - H(X|Y,Z)$$

- The chain rule for mutual information is defined:

$$I(X_1, ..., X_n; Y) = I(X_1; Y) + ... + I(Xn; Y|X_1, ..., X_{n-1})$$

$$= \sum_{i=1}^{n} I(X_i; Y|X_1, ..., X_{i-1})$$

## The Noisy Channel Model

- There is a trade-off between compression and transmission accuracy. The first reduces space, the second increases it.

- Channels are characterized by their capacity, which (in a memoryless channel) can be expressed $C = max_{p(X)} I(X;Y)$ where $X$ is input to the channel and $Y$ is channel output.

- Channel capacity can be reached if an input code $X$ is designed that maximizes mutual information between $X$ and $Y$ over all possible input distributions $p(X)$.

## Relative Entropy

- Given two probability mass functions $p$ and $q$, relative entropy is defined

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

- Relative Entropy gives a measure of how different two probability distributions are.

- Mutual Information is really a measure of how far a joint distribution is from independence

$$I(X;Y) = D(p(x,y)||p(x)P(y))$$

- Conditional relative entropy and a chain rule are also defined.

# The Relation to Language

- Given a history of words h, the next word w, and a model m, define point-wise entropy as $H(w|h) = -\log_2 m(w|h)$. If the model is correct point-wise entropy is 0, if the model is incorrect point-wise entropy is infinite. In this sense a model's accuracy is tested, and one would hope to keep these 'surprises' to a minimum.

- In practice $p(x)$ may not be known, so a model m is best when $D(p||m)$ is minimal. Unfortunately if $p(x)$ is unknown, $D(p||m)$ can only be approximated using techniques like cross entropy and perplexity.

## Cross Entropy

- The cross entropy between $X$ with actual probability distribution $p(x)$ and a model $q(x)$ is

  $$H(X,q) = H(X){+}D(p||q) = -\sum_{x \in X} p(x) \log q(x)$$

- If a large sample body is available cross entropy can be approximated

  $$H(X,q) \approx \frac{1}{n} \log q(x_{1,n})$$

- Minimizing cross entropy is equivalent to minimizing relative entropy, which brings the model's probability distribution closer to the actual probability distribution.

## Perplexity

- 'A perplexity of k means that you are as surprised on average as you would have been if you had had to guess between k equiprobable choices at each step.' It is defined

$$perplexity(x_{1n}, m) = 2^{H(x_{1,n},m)} = m(x_{1n})^{\frac{1}{n}}$$

## The Entropy of English

- English can be modeled using n-gram models, or Markov chains. They assume the probability of the next word relies on the previous k in the stream.

- Models have exhibited cross entropy with English as low as 2.8 bits, and experiments with humans have resulted in cross entropy of 1.34 bits.