Mathematical Foundations

Fang-Hui, Chu

References:
1. Foundations of Statistical Natural Language Processing, chap 2

Outline

- 2.1 Elementary Probability Theory
  - Probability spaces
  - Conditional probability and independence
  - Bayes' theorem
  - Random variables
  - Expectation and variance
  - Joint and conditional distributions
  - Standard distributions
  - Bayesian statistics
- 2.2 Essential Information Theory
  - Entropy
  - Joint entropy and conditional entropy
  - Mutual information
  - The noisy channel model
  - Relative entropy or Kullback-Leibler divergence
  - Cross entropy & Perplexity

Probability spaces

- Probability theory deals with predicting how likely it is that something will happen
- An experiment (or trial) is the process by which an observation is made
- Sample space: Ω
  - A collection of basic *outcomes* (or *sample points*) for our experiment
- Event space: the set of all subsets of the sample space
- The foundations of probability theory depend on the set of events F forming a σ-field
- Probability function $P(A) = \frac{|A|}{|\Omega|}$
- $P(\Omega)=1$
- A well-founded probability space consists of a sample space Ω, a σ-field of events F, and a probability function P

Conditional probability and independence(1/2)

- The conditional probability of an event A given that an event B has occurred is

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

- $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$ [The multiplication rule]
- The chain rule is as fellows:

$$P(A_1 \cap \ldots \cap A_n) =$$

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

Conditional probability and independence(2/2)

- Two events A, B are independent of each other if

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

- Two event A and B are conditionally independent given C when

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

Bayes' theorem(1/3)

- Bayes' theorem lets us swap the order of dependence between events. (calculate P(B|A) in terms of P(A|B))

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

The righthand side denominator P(A) can be viewed as a normalizing constant

$$\arg\max_B \frac{P(A \mid B)P(B)}{P(A)} = \arg\max_B P(A \mid B)P(B)$$

Since the denominator is the same in all cases

Bayes' theorem(2/3)

- The set A can be divided into two parts

$$P(A \cap B) = P(A \mid B)P(B), \quad P(A \cap \overline{B}) = P(A \mid \overline{B})P(\overline{B})$$

$$so\ we\ have:$$

$$P(A) = P(A \cap B) + P(A \cap \overline{B}) = P(A \mid B)P(B) + P(A \mid \overline{B})P(\overline{B})$$

[additivity]

- If we have some group of sets B_i that partition A, if $A \subseteq \cup_i B_i$ and the B_i are disjoint, then

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

Bayes' theorem(3/3)

- Bayes' theorem (more elaborated version)

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

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{P(A)} = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$$

Random variables(1/2)

- Random variables is simply a function

  $X: \Omega \rightarrow R^n$ (commonly with n=1)
  where R is the set of real numbers

- A discrete random variable is a function

  $X: \Omega \rightarrow S$
  where S is a countable subset of R

- A indicator random variable or a Bernoulli trial :

  $X: \Omega \rightarrow \{0,1\}$

Because a random variable has a numeric range, we can do mathematics more easily by working with the values of a random variable, rather than directly with events

Random variables(2/2)

- We can define the probability mass function (pmf) for a random variable X, which gives the probability that the random variable has different numeric values:

$$pmf\ p(x) = p(X = x) = P(A_x) \quad where\ A_x = \{\omega \in \Omega : X(\omega) = x\}$$

- For a discrete random variable , we have that

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

Expectation and variance(1/4)

- The expectation is the mean or average of a random variable
- If X is a random variable with a pmf p(x) such that $\sum_x |x| p(x) < \infty$ then the expectation is

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

- Example:

  if Y is the value of face on one rolling die ,then

$$E(Y) = \sum_{y=1}^{6} yp(y) = \frac{1}{6}\sum_{y=1}^{6} y = \frac{21}{6} = 3\frac{1}{2}$$

This is the expected average found by totaling up a large number of throws of the die, and dividing by the number of throws.

Expectation and variance(2/4)

- If Y~p(y) is a random variable, any function g(Y) defines a new random variable. If E(g(Y)) is defined, then

$$E\big(g(Y)\big)=\sum_{y} g(y)p(y)$$

  - g(Y)=aY+b, we see that E(g(Y))=aE(Y)+b

- We also have that E(X+Y)=E(X)+E(Y)

  - If X and Y are independent, then E(XY)=E(X)E(Y)

Expectation and variance(3/4)

- The variance of a random variable is a measure of whether the values tend to be consistent over trials or to vary a lot
- One measures it by finding out how much on average the variable's values deviate from the variable's expectation

$$
\begin{aligned}
&Var(X) = E\left((X - E[X])^2\right) \\
&= E[(x-\mu_X)^2] = \sum_X (x-\mu_X)^2 P_X(x) = \sum_X x^2 P_X(x) - \sum_X 2\mu_X x P_X(x) - \sum_X {\mu_X}^2 P_X(x) \\
&= E[X^2] - 2\mu_X x \sum_X x P_X(x) + {\mu_X}^2 \sum_X P_X(x) = E[X^2] - 2{\mu_X}^2 + {\mu_X}^2 \\
&= E[X^2] - E^2[X]
\end{aligned}
$$

- In commonly denotes the mean is μ and the variance is σ^2 the standard deviation is hence written as σ

Expectation and variance(4/4)

- Proof of variance calculation II

$$
\begin{aligned}
Var(X) &= E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E^2(X) \\
&= \sum_x p(x^2)x^2 - 2E^2(X) + E^2(X) \\
&= \sum_x p(x)x^2 - 2E(X)\sum_x p(x)x + 1E^2(X) \\
&= \sum_x p(x)x^2 - \sum_x p(x)x2E(X) + \sum_x p(x)E^2(X) \\
&= \sum_x p(x)\left(x^2 - 2xE(X) + E^2(X)\right) \\
&= \sum_x p(x)\left(x - E(X)\right)^2
\end{aligned}
$$

Joint and conditional distributions(1/2)

- The joint probability mass function for two discrete random variables X,Y is p(x,y)=P(X=x,Y=y)

- The marginal pmfs, which total up the probability masses for the value of each variable separately

$$p_X(x) = \sum_y p(x,y) \text{ , } p_Y(y) = \sum_x p(x,y)$$

  - In general the marginal mass function don't determine the joint mass function.

Joint and conditional distributions(2/2)

- If X and Y are independent, then $p(x,y)=p_X(x)p_Y(y)$.
- To define a conditional pmf in terms of the joint distribution:

$$p_{X|Y}(x \mid y) = \frac{p(x,y)}{p_Y(y)} \quad for\ y\ such\ that\ \ p_Y(y) > 0$$

- And deduce a chain rule in terms of random variables, like

$$p(w,x,y,z) = p(w)p(x \mid w)p(y \mid w,x)p(z \mid w,x,y)$$

Standard distributions(1/3)

- Statisticians refer to the family of functions as a distribution and to the numbers that define the different members of the family as parameters

- Discrete distributions:
  - Binomial distribution

- Continuous distributions:
  - Normal distribution

Standard distributions(2/3)

- The Binomial distribution results when one has a series of trials with only two outcomes, each trial being independent from all the others

- The binomial distributions gives the number *r* of successes out of *n* trials and the probability of success in any trial is p

$$b(r;n,p) = \binom{n}{r} p^r (1-p)^{n-r} \; where \; \binom{n}{r} = \frac{n!}{(n-r)!r!} \quad 0 \le r \le n$$

- Let *R* have as value the number of heads in *n* tosses of a coin, where the probability of a head is p

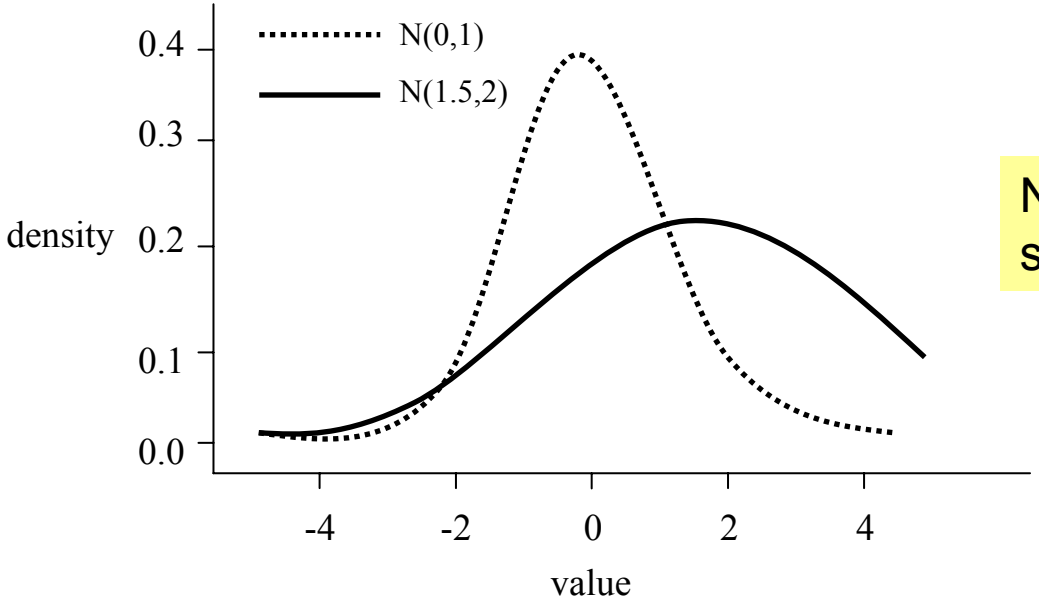
$$p(R = r) = b(r;n,p)$$

- The binomial distribution has an expectation of *np* and a variance of *np*(1-*p*)

Standard distributions(3/3)

- Normal distribution:
  (pdf, do not directly give the probabilities of the points)
  - With two parameters : μ : mean (variance)
    σ : standard deviation
  - And the bell curve is given by:

$$n(x;\ \mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/\left(2\sigma^2\right)}$$



N(0,1): μ =0 and σ =1:
standard normal distribution

Bayesian statistics(1/8)

- Frequentist Statistics vs. Bayesian Statistics

- Bayesian updating
  - A coin is tossed in times and gets 8 heads then this coin comes down heads 8 times out of 10. (from a frequent point of view)
    - This is the maximum likelihood estimate

  - Bayesian statistics：measure degree of belief, and are calculated by starting with prior belief and updating tem in the face of evidence, by use of Bayes' theorem

Bayesian statistics(2/8)

- μ_m be the model that asserts P (head) = m

  s be a sequence of observations, i heads and j tails

  For any m, $0 \leq m \leq 1$:

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

- From a frequentist point of view, we wish to find the MLE

$$\arg\max_m P(s \mid \mu_m)$$

- We can differentiate the above polynomial then the answer is $\frac{i}{i+j}$, or 0.8 for the case of 8 heads and 2 tails

Bayesian statistics(3/8)

- Bayesian undating—

  – Assume one's prior belief is modeled by

$$P(\mu_m) = 6m(1-m)$$

  because this distribution is centered on 1/2

  – When one sees an observation sequence s one wants to know one's new belief in the fairness of the coin. By Bayes' theorem

$$P(\mu_m \mid s) = \frac{P(s \mid \mu_m)P(\mu_m)}{P(s)}$$

$$= \frac{m^i(1-m)^j \times 6m(1-m)}{P(s)}$$

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

Bayesian statistics(4/8)

- P(s) is the prior probability of s
- s doesn't depend on μ_m so we can ignore it (normalization factor)
- If we then differentiate the numerator so as find its maximum, then we can determine that the case of 8 heads and 2 tails

$$\arg\max_m P(\mu_m \mid s) = \frac{6m^{i+1}(1-m)^{j+1}}{P(s)} \approx \arg\max_m 6m^{i+1}(1-m)^{j+1}$$

$$\frac{\partial P(\mu_m \mid s)}{\partial m} \approx \frac{\partial 6m^{i+1}(1-m)^{j+1}}{\partial m} = 6(i+1)m^i(1-m)^{j+1} - 6(j+1)m^{i+1}(1-m)^j$$

$$\frac{\partial P(\mu_m \mid s)}{\partial m} = 0 \Rightarrow m = \frac{i+1}{i+j+2}$$

$$\arg\max_m P(\mu_m \mid s) = \frac{3}{4}$$

We have moved a long way in the direction of believing that the coin is biased, but we haven't moved all the way to 0.8

Bayesian statistics(5/8)

- Marginal probability
  - Adding up all the P(s| μ_m) weighted by the probability of μ_m

For the continuous case

$$P(s) = \int_0^1 P(s \mid \mu_m) P(\mu_m) dm$$

$$= \int_0^1 6m^{i+1}(1-m)^{j+1} dm$$

$$= \frac{6(i+1)!(j+1)!}{(i+j+3)!}$$

The denominator is just a normalization factor, which ensures that what we calculate is actually a probability function

This just happens to be an instance of the beta integral

Bayesian statistics(6/8)

- Beta Function:

$$B(\alpha,\beta)=\int_0^1 x^{\alpha}(1-x)^{\beta}=B(\beta,\alpha)$$

$$B(\alpha,\beta)=\frac{\alpha}{\beta+1}B(\alpha-1,\beta+1)=\frac{\beta}{\alpha+1}B(\alpha+1,\beta-1)$$

$$=\frac{\alpha}{\beta+1}\cdot\frac{\alpha-1}{\beta+2}B(\alpha-2,\beta+2)=\ldots$$

$$=\frac{\alpha!}{\frac{(\beta+\alpha-1)!}{\beta!}}B(1,\beta+\alpha-1)$$

$$=\frac{\alpha!\beta!}{(\beta+\alpha-1)!}\int_0^1 x^{\beta+\alpha-1}(1-x)^1dx$$

$$=\frac{\alpha!\beta!}{(\beta+\alpha-1)!}\left[\int_0^1 x^{\beta+\alpha-1}dx-\int_0^1 x^{\beta+\alpha}dx\right]$$

$$=\frac{\alpha!\beta!}{(\beta+\alpha-1)!}\cdot\frac{1}{(\alpha+\beta)(\alpha+\beta+1)}$$

$$=\frac{\alpha!\beta!}{(\alpha+\beta+1)!}$$

Integration by part:

$$\int udv=uv-\int vdu$$

$$\int_0^1 x^{\alpha}(1-x)^{\beta}=x^{\alpha}\cdot\frac{-1}{\beta+1}(1-x)^{\beta+1}-\int_0^1\frac{-1}{\beta+1}(1-x)^{\beta+1}\alpha x^{\alpha-1}dx$$

$$=\frac{-1}{\beta+1}x^{\alpha}(1-x)^{\beta+1}+\frac{\alpha}{\beta+1}\int_0^1 x^{\alpha-1}(1-x)^{\beta+1}dx$$

??

Bayesian statistics(7/8)

- Bayesian decision theory
  - To evaluate which model better explains some data
- Example:
  comparing two models ν and μ
  - The results reported truly reflect the results of tossing a single. This is the theory μ
  - Tossing two fair coins and calling out “tails” if both of them come down tails and otherwise. this is called theory ν

$$we\ have\ P(s\,|\,\nu)=\left(\frac{3}{4}\right)^{i}\left(\frac{1}{4}\right)^{j}\ \ and\ \ P(\mu)=P(\nu)=\frac{1}{2}$$

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

Bayesian statistics(8/8)

- Bayesian decision theory

$$\frac{P(\mu \mid s)}{P(\nu \mid s)} = \frac{P(s \mid \mu)P(\mu)}{P(s)} \times \frac{P(s)}{P(s \mid \nu)P(\nu)}$$

$$= \frac{P(s \mid \mu)P(\mu)}{P(s \mid \nu)P(\nu)} = \frac{\frac{6(i+1)!(j+1)!}{(i+j+3)!}}{\left(\frac{3}{4}\right)^i \left(\frac{1}{4}\right)^j} = \frac{\frac{6(8+1)!(2+1)!}{(8+2+3)!}}{\left(\frac{3}{4}\right)^8 \left(\frac{1}{4}\right)^2} = 0.33$$

  - The quantity we are now describing as P(s| μ) is the quantity that we wrote as just P(s) (assuming that theory μ_m was true and we were just trying to determine m)
  - If the ratio is greater than 1, we should prefer μ , and otherwise we should prefer ν

Outline

- 2.1 Mathematical Foundations
  - Probability spaces
  - Conditional probability and independence
  - Bayes' theorem
  - Random variables
  - Expectation and variance
  - Joint and conditional distributions
  - Standard distributions
  - Bayesian statistics
- 2.2 Essential Information Theory
  - Entropy
  - Joint entropy and conditional entropy
  - Mutual information
  - The noisy channel model
  - Relative entropy or Kullback-Leibler divergence
  - Cross entropy & Perplexity

Entropy(1/4)

- Entropy measures the amount of information in a random variable. It is normally measured in bits

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

- We define $0 \log_2 0 = 0$

- Entropy:
  - The average uncertainty of a single random variable
  - The average length of the message needed to transmit an outcome of that variable
  - We hope the entropy is lower in the system

Entropy(2/4)

- Example:

  Suppose you are reporting the result of rolling an 8-sided die. Then the entropy is:

$$H(X) = -\sum_{i=1}^{8} p(i) \log p(i) = -\sum_{i=1}^{8} \frac{1}{8} \log \frac{1}{8}$$

$$= -\log \frac{1}{8} = \log 8 = 3 bits$$

Entropy(3/4)

- Properties of Entropy:

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

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

$$= E\left(\log \frac{1}{p(x)} \right)$$

Entropy(4/4)

- Entropy can be interpreted as a measure of the size of the '*search space*' (the possible values of a random variable and its associated probabilities)

- Note that:
  - H(X)>=0
  - H(X)=0 only when the value of X is determinate
  - Entropy increase with the message length

Joint Entropy and Conditional Entropy(1/3)

- Joint Entropy:

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

- Conditional Entropy:

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

Joint Entropy and Conditional Entropy(2/3)

- Proof of Conditional Entropy:

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

$$= \sum_{x \in X} p(x) \left[- \sum_{y \in Y} p(y|x) \log p(y|x) \right]$$

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

Joint Entropy and Conditional Entropy(3/3)

- Chain rule for Entropy:

$$H(X,Y) = H(X) + H(Y \mid X)$$

- Proof:

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\big(p(y\mid x)p(x)\big) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\big(\log p(y\mid x) + \log p(x)\big) \\
&= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y\mid x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) \\
&= H(Y\mid X) + H(X)
\end{aligned}$$

Entropy rate

- A Chain rule of entropy:

  H(X1,…,X2) = H(X1) + H(X2|X1) + …… + H(Xn|X1,…,Xn-1)

- For a message of length n the ,entropy rate is:

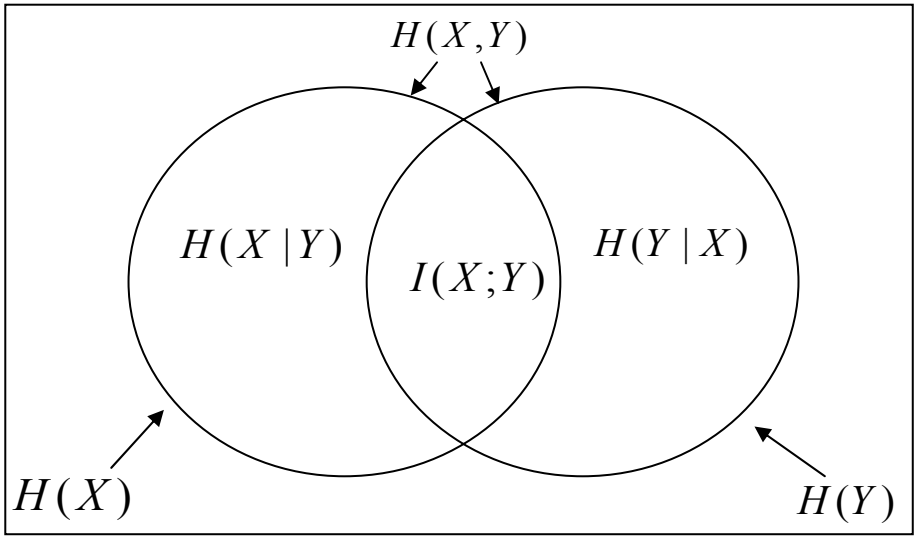
$$H_{rate} = \frac{1}{n} H(X_{1n}) = -\frac{1}{n} \sum_{X1n} p(X_{1n}) \log p(X_{1n})$$

Mutual Information(1/5)

$$H(Y,X) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$
$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

This difference is called the mutual information between X and Y

(The amount of information one random variable contains about another)



Mutual Information(2/5)

- Mutual information is a symmetric, non-negative measure of the common information in the two variables

- It is 0 only when two variables are independent.
  The mutual Information is 0 for two independent events

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

Mutual Information(3/5)

- How to simply calculate Mutual Information ?

$$I(X;Y) = H(X) - H(X \mid Y)$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= \sum_x p(x) \log \frac{1}{p(x)} + \sum_y p(y) \log \frac{1}{p(y)} + \sum_{x,y} p(x,y) \log p(x,y)$$
$$= \sum_{x,y} p(x,y) \log \frac{1}{p(x)} + \sum_{x,y} p(x,y) \log \frac{1}{p(y)} + \sum_{x,y} p(x,y) \log p(x,y)$$
$$= \sum_{x,y} p(x,y) \left[\log \frac{1}{p(x)} + \log \frac{1}{p(y)} + \log p(x,y) \right]$$
$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual Information(4/5)

- Conditional mutual information

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

- Chain rule

$$I(X_{1n};\ Y) = I(X_1;\ Y) + \ldots + I(X_n;\ Y \mid X_1,\ldots,X_{n-1})$$

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

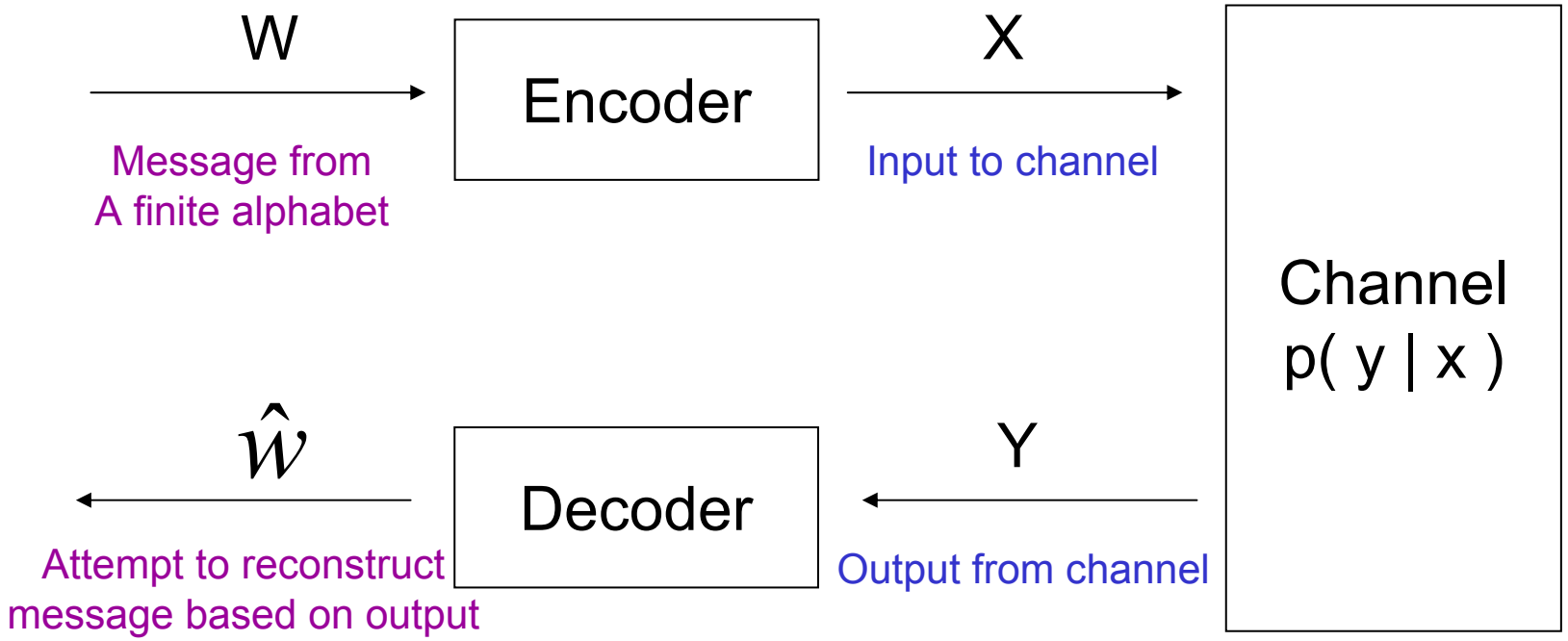
Mutual Information(5/5)

- Define the pointwise mutual information between two particular points

$$I(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

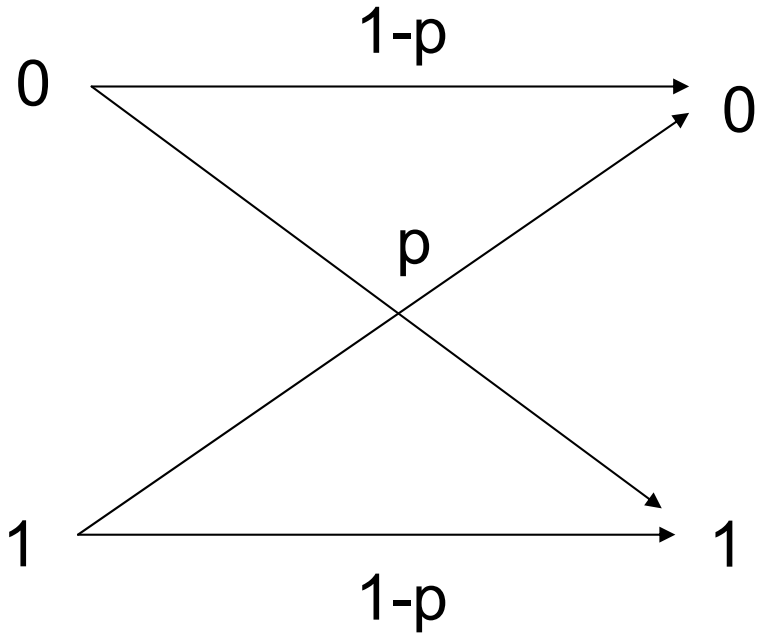
This has sometimes been used as a measure of association between elements

The noisy channel model(1/4)



The noisy channel model

The noisy channel model(2/4)



A binary symmetric channel. A 1 or 0 in the input gets flipped on transmission with probability p.

The noisy channel model(3/4)

- Capacity
  - The channel capacity describes the rate at which one can transmit information through the channel with an arbitrarily low probability of being unable to recover the input from the output

$$C = \max_{p(X)} I(X;Y) \qquad \text{if } p=0 \text{ or } p=1 \Rightarrow C=1$$

$$= \max_{p(X)} H(Y) - H(Y \mid X) \qquad \text{if } p=\frac{1}{2} \Rightarrow C=0$$

$$= \max_{p(X)} H(Y) - H(p) = 1 - H(p) \quad 0 < C \le 1$$

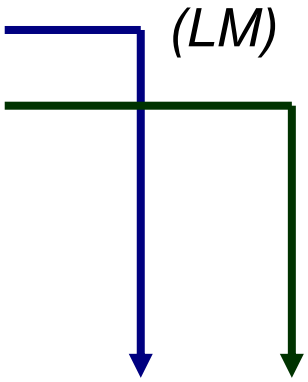
The capacity is used to measured the likeness of X and Y .
If the mutual information is 1 then the X and Y are the same or bits are inverted completely.

The noisy channel model(4/4)

Application: (In speech recognition)

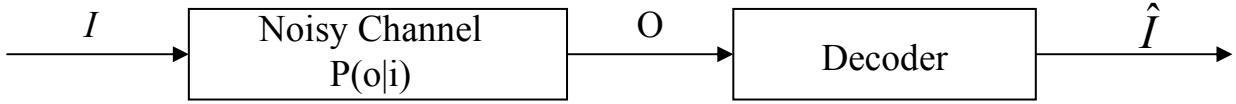
Researchers cast both speech recognition and machine translation as a noisy channel problem

Input: *word sequences*
Output: *observed speech signal*
P(input): *probability of word sequences*
P(output|input): *acoustic model (channel prob.)*



Bayes' theorem

$$\hat{I} = \arg\max_i p(i \mid o) = \arg\max_i \frac{p(i)\,p(o \mid i)}{p(o)} = \arg\max_i p(i)\,p(o \mid i)$$



Relative Entropy or Kullback-Leibler divergence(1/3)

- For two probability mass functions, *p(x)* , *q(x)* their relative entropy is given by:

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

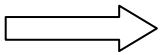
define:

$$0 \log \frac{0}{q} = 0 \quad and \quad p \log \frac{p}{0} = \infty$$

Relative Entropy or Kullback-Leibler divergence(2/3)

- It is a measure of how different two probability distributions are (over the same event space)

- Meaning ：It is the average number of bits that are wasted by encoding events from a distribution *p* with a code based on a not-quite-right distribution *q*
  - *Non-negative*
  - *D(p||q) = 0 iff p = q*



KL distance

But:

- not symmetric in p and q
- not satisfy the triangle inequality

$$d(x,y) \le d(x,z) + d(z,y)$$

Relative Entropy or Kullback-Leibler divergence(3/3)

Mutual information is actually just a measure of how far a joint distribution is from independence:

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

$$= D(p(x,y) \,||\, p(x)p(y))$$

Define the Conditional Relative Entropy and a chain rule for relative entropy:

$$D(p(y \mid x) \,||\, q(y \mid x)) = \sum_{x} p(x) \sum_{y} p(y \mid x) \log \frac{p(y \mid x)}{q(y \mid x)}$$

$$D(p(x,y) \,||\, q(x,y)) = D(p(x) \,||\, q(x)) + D(p(y \mid x) \,||\, q(y \mid x))$$

Cross entropy(1/4)

- Cross entropy:
  - *The cross entropy between a random variable X with true probability distribution p(X) and another pmf q (normally a model of p) is given by:*

$$H(X,q) = H(X) + D(p \| q)$$

$$= \sum_{x \in X} p(x) \log \frac{1}{p(x)} + \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

$$= \sum_{x \in X} p(x) \left[\log \frac{1}{p(x)} + \log \frac{p(x)}{q(x)} \right]$$

$$= \sum_{x \in X} p(x) \left[\log \frac{1}{q(x)} \right] = -\sum_{x \in X} p(x) \log q(x)$$

Cross entropy(2/4)

Cross entropy of a language :

suppose

Language L = (X_i) ~ p(x) according to a model m by

$$H(L,m) = -\lim_{n\to\infty}\frac{1}{n}\sum_{x_{1n}} p(x_{1n})\log m(x_{1n}) = -\lim_{n\to\infty}\frac{1}{n}E\left(\log m(x_{1n})\right)$$

We cannot calculate this quantity without knowing p. But if we make certain assumptions that the language is 'nice,' then the cross entropy for the language can be calculated as:

$$H(L,m) = -\lim_{n\to\infty}\frac{1}{n}\log m(x_{1n})$$

Cross entropy(3/4)

- Expectation is a weighted average over all possible sequence.
- But we are using a limit and looking at longer and longer sequences of language use.
- If we have seen a huge amount of the language, what we have seen is “typical”
- We no longer need to average over all samples of the language; the value for the entropy rate given by this particular sample will be roughly right

$$H(L,m) = -\lim_{n\to\infty}\frac{1}{n}\sum_{x_{1n}} p(x_{1n})\log m(x_{1n}) = -\lim_{n\to\infty}\frac{1}{n}E\left(\log m(x_{1n})\right)$$

$$= \lim_{n\to\infty}\frac{1}{n}E\left(\log\frac{1}{m(x_{1n})}\right)$$

$$= \lim_{n\to\infty}\frac{1}{n}\log\frac{1}{m(x_{1n})} \approx -\frac{1}{n}\log m(x_{1n})$$

Cross entropy(4/4)

- Cross entropy of a language :
  - We do not actually attempt to calculate the limit, but approximate it by calculating for a sufficiently large n:

$$H(L,m) \approx -\frac{1}{n}\log m(x_{1n})$$

  - This measure is just the figure for our average surprise

- *Our goal will be to try to minimize this number*
  - *Because H(X) is fixed, this is equivalent to minimizing the relative entropy, which is a measure of how much our probability distribution departs from actual language use*

Perplexity

In the speech recognition community, people tend to refer to perplexity rather than cross entropy. The relationship between the two is simple:

$$Perplexity(x_{1n}, m)$$

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

$$= m(x_{1n})^{-\frac{1}{n}}$$

Why we use perplexity not cross entropy?

Because it is much easier to impress funding bodies by saying that "we've managed to reduce perplexity from 950 to only 540" than by saying that "we've reduced cross entropy from 9.9 to 9.1 bits."