Mathematical Foundations

Foundations of Statistical Natural Language Processing, chapter2

Presented by Chu Haei-Ming（朱惠銘）
CSIE, NTNU
g92470059@csie.ntnu.edu.tw

Outline

- Elementary Probability Theory
  - Probability spaces
  - Conditional probability and independence
  - Bayes' theorem
  - Random variables
  - Expectation and variance
  - Joint and conditional distributions
  - Standard distributions
  - Bayesian statistics
- Essential Information Theory
  - Entropy
  - Joint entropy and conditional entropy
  - Mutual information
  - Relative entropy or Kullback-Leibler divergence

Elementary Probability Theory

Probability spaces

- Sample space: Ω
- Event A is the subset of Ω
- Probability function $$P(A) = \frac{|A|}{|\Omega|}$$
- $P(\Omega)=1$
- Example :
  A fair coin tossed 3 times. What is th chance of 2 heads?
  - Ω={HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}
  - A={HHT,HTH,THH}
  - So $$P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8}$$

Elementary Probability Theory

Conditional probability and independence

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

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

- Evan if P(B)=0 we have that：

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

- The chain rule is as fellows:

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

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

Elementary Probability Theory

Conditional probability and independence

- Two event 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

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

Elementary Probability Theory

Bayes' theorem

- Bayes' theorem lets us swap the order of dependence between events.

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

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

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

Elementary Probability Theory
Bayes' theorem

- The set A can be divided into two parts

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

$$\textit{so we have}:$$

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

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

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

Elementary Probability Theory

Bayes' theorem

- Bayes' theorem

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

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

Elementary Probability Theory

Bayes' theorem

- Example
  - Let G be the event of the sentence having a parasitic gap, and let T be the event of the test being positive

$$P(G \mid T) = \frac{P(T \mid G)P(G)}{P(T)}$$

$$= \frac{P(T \mid G)P(G)}{P(T \mid G)P(G) + P(T \mid \overline{G})P(\overline{G})}$$

$$= \frac{0.95 \times 0.00001}{0.95 \times 0.00001 + 0.005 \times 0.99999} \approx 0.002$$

  - On average, only 1 in every 500 sentences that the test identifiers will actually contain a parasitic gap.
  - Because the prior probability of a sentence containing a parasitic gap is so low

Elementary Probability Theory

Random variables

- Random variables is simply a function

  $$X : \Omega \rightarrow \mathbf{R}^{n}$$

  R is the set of real numbers, commonly with n=1

- A discrete random variable is a function

  $$X : \Omega \rightarrow \mathbf{S}$$

  where S is a countable subset of $\mathbf{R}$

- A indicator random variable is a function

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

  and X is also called a Bernoulli trial

Elementary Probability Theory

Random variables

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

$$pmf \;\; p(x) = p(X = x) = P(A_x)$$

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

- For a discrete random variable , we have

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

Elementary Probability Theory

Random variables

Example:

Suppose the event are those that result from tossing two dice

The discrete random variable X that is the sum of their faces:

S={2,…,12} as fellows

First die	Second die										
	1	2	3	4	5	6					
6	7	8	9	10	11	12					
5	6	7	8	9	10	11					
4	5	6	7	8	9	10					
3	4	5	6	7	8	9					
2	3	4	5	6	7	8					
1	2	3	4	5	6	7					
x	2	3	4	5	6	7	8	9	10	11	12
p(X=x)	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

Elementary Probability Theory

Expectation and variance

- 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.

Elementary Probability Theory

Expectation and variance

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

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

  - Example：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)

Elementary Probability Theory

Expectation and variance

- The variance is the measure of the random variable 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

$$Var(X) = E\left(\left(X - E(X)\right)^2\right) = E(X^2) - E^2(X)$$

- The standard deviation of a variable is the square root of the variance.
- In commonly denotes the mean is μ and the variance is σ^2 the standard deviation is hence written as σ

Elementary Probability Theory

Expectation and variance

- Proof of variance calculation I

$$
\begin{aligned}
Var(X) &= E\left(\left(X - E(X)\right)^2\right) \\
&= E\left(X^2 - 2XE(X) + \left(E(X)\right)^2\right) \\
&= E(X^2) - E\left(2XE(X)\right) + E\left(\left(E(X)\right)^2\right) \\
&= E(X^2) - 2E(X)E(X) + \left(E(X)\right)^2 \\
&= E(X^2) - E^2(X)
\end{aligned}
$$

Elementary Probability Theory

Expectation and variance

- Proof of variance calculation II

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

Elementary Probability Theory

Expectation and variance

- Example：What is the expectation and variance for the random variable introduced in example3, the sum of the numbers on two dies?

$$E(X) = E(Y+Y) = E(Y) + E(Y) = 3\frac{1}{2} + 3\frac{1}{2} = 7$$

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

Elementary Probability Theory

Joint and conditional distributions

- 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

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

Elementary Probability Theory

Joint and conditional distributions

- If X and Y are independent, then

$$p(x,y)=p_X(x)p_Y(y)$$

  – Example:
  getting two sixes from rolling two dice, since the events are independent, we can compute that:

$$p(Y=6,Z=6)=p(Y=6)\,p(Z=6)=\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$$

- The conditional pmf in terms of the joint distribution

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

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

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

Elementary Probability Theory
Standard distributions

- Discrete distributions:
  - Binomial distribution
- Continuous distributions:
  - Normal distribution

Elementary Probability Theory

Standard distributions

- 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

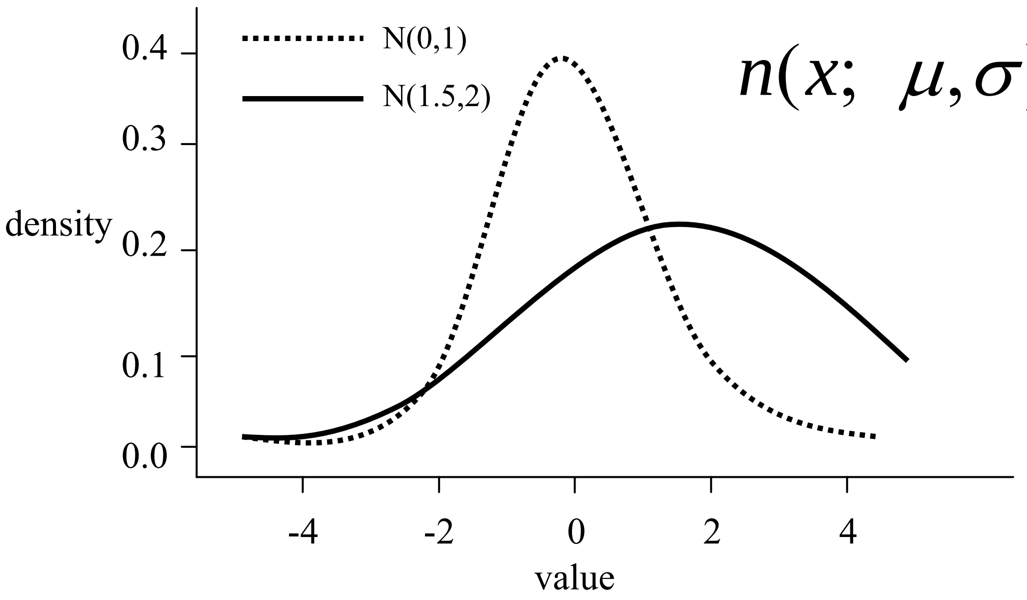
$$p(R = r) = b(r; n, p)$$

Elementary Probability Theory
Standard distributions

- Multinomial distribution
  - The generalization of a binomial trial to the case where each of the trial has more than two basic outcomes is called multinomial experiment and modeled by it.
  - A zeroth order n-gram model is a straightforward example of a multinomial distribution.

Elementary Probability Theory
Standard distributions

- Normal distribution
  - 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): $\mu=0$ and $\sigma=1$: standard normal distribution

Elementary Probability Theory

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.
    - This is the maximum likelihood estimate
  - But he belief the coin would come down equally head and tails over the long run this is called a prior belief
  - Bayesian statistics
    - Measure degree of belief
    - Starting with prior belief
    - updating tem in the face of evidence
    - By use of Bayes' theorem

Elementary Probability Theory

Bayesian statistics

- μ_m be the model that asserts P (head) = m
- s be a sequence of observations:
  i heads and j tails
- For any m, $0 \le m \le 1$

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

- 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 i / i+j, or 0.8 for the case of 8 heads and 2 tails

Elementary Probability Theory

Bayesian statistics

- Assume one's prior belief is modeled by

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

because this distribution is centered on 1/2

- By bayes' theorem

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

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

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

Elementary Probability Theory

Bayesian statistics

- P(s) is the prior probability of s
- s doesn't depend on μ_m so we can ignore it
- Then we can determine the case for 8 heads and 2 tails

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

$$= \arg\max_m 6m^{i+1}(1-m)^{j+1} = \arg\max_m 6m^{8+1}(1-m)^{2+1}$$

$$= \arg\max_m 6m^{9}(1-m)^{3} = \frac{9}{9+3} = \frac{3}{4}$$

- 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

Elementary Probability Theory

Bayesian statistics

- Marginal probability
  - Adding up all the $P(s|\mu_m)$ weighted by the probability of μ_m
- For the continuous case

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

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

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

- It is a normalization factor, for $P(\mu_m|s)$ is a actually a probability function

Elementary Probability Theory

Bayesian statistics

- Bayesian decision theory
  - To evaluate which model better explains some data
- Example:
  comparing two models ν and μ
  - Tossing two fair coins and called out "tails" if both of tem come down tails this is called theory ν and the theory μ above

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

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

Elementary Probability Theory

Bayesian statistics

- 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)
  - If the ratio is greater than 1, we should prefer μ

Outline

- Elementary Probability Theory
  - Probability spaces
  - Conditional probability and independence
  - Bayes' theorem
  - Random variables
  - Expectation and variance
  - Joint and conditional distributions
  - Gaussian distributions
- Essential Information Theory
  - Entropy
  - Joint entropy and conditional entropy
  - Mutual information
  - Relative entropy or Kullback-Leibler divergence

Essential Information Theory

Entropy

- 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

$$0 \log_2 0 = 0$$

Essential Information Theory

Entropy

- Example:

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

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

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

Essential Information Theory
Entropy

- Entropy:
  - The average number of bits used for identifying the transmission of the information
  - We hope the entropy is lower in the system

Essential Information Theory

Entropy

- Properties of Entropy:

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

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

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

Essential Information Theory

Joint Entropy and Conditional Entropy

- Joint Entropy:

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

- Conditional Entropy:

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

Essential Information Theory

Joint Entropy and Conditional Entropy

- Proof of Conditional Entropy:

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

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

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

Essential Information Theory

Joint Entropy and Conditional Entropy

- Chain rule for Entropy:

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

- Proof:

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

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

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

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

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

Essential Information Theory

Entropy rate

- Per-letter or per-word entropy
- For a message of length n the entropy rate

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

- Assume that a language is a stochastic process consisting of a sequence of tokens L=(Xi)

$$H_{rate}(L) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

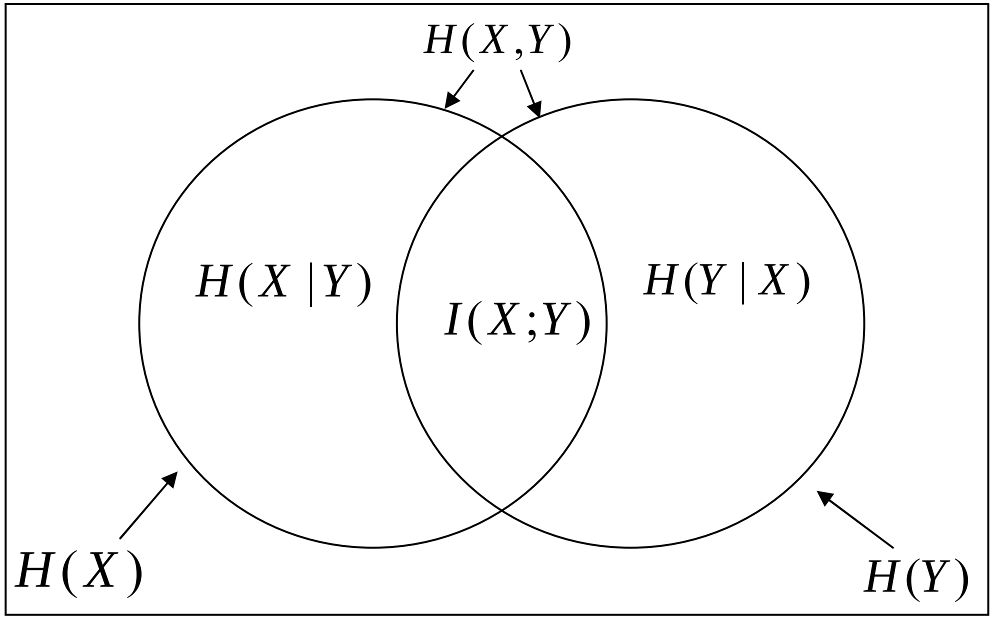
Essential Information Theory

Mutual Information

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

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

This difference is called the mutual information between X and Y



Essential Information Theory

Mutual Information

- The likeness of Information 。
- This difference is called the ***mutual information*** between X and Y.
- The amount of information one random variable contains about another.
- It is 0 only when two variables are independent.
  The mutual Information is 0 for two independent events 。

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

Essential Information Theory

Mutual Information

- How to simply calculate Mutual Information？

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

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

$$= \sum_{x} p(x)\log\frac{1}{p(x)} + \sum_{y} p(y)\log\frac{1}{p(y)} + \sum_{x,y} p(x,y)\log p(x,y)$$

$$= \sum_{x,y} p(x,y)\log\frac{1}{p(x)} + \sum_{x,y} p(x,y)\log\frac{1}{p(y)} + \sum_{x,y} p(x,y)\log p(x,y)$$

$$= \sum_{x,y} p(x,y)\left[\log\frac{1}{p(x)} + \log\frac{1}{p(y)} - \log\frac{1}{p(x,y)}\right]$$

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

Essential Information Theory

Mutual Information

- Conditional mutual information

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

- Chain rule

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

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

Essential Information Theory

Mutual Information

- 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.

Essential Information Theory

Relative Entropy or Kullback-Leibler divergence

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

Essential Information Theory

Relative Entropy or Kullback-Leibler divergence

- 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.

- Some authors use the name "KL distance", but note that relative entropy isn't a metric (it doesn't satisfy the triangle inequality)

Essential Information Theory

Relative Entropy or Kullback-Leibler divergence

Properties of KL-divergence:

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

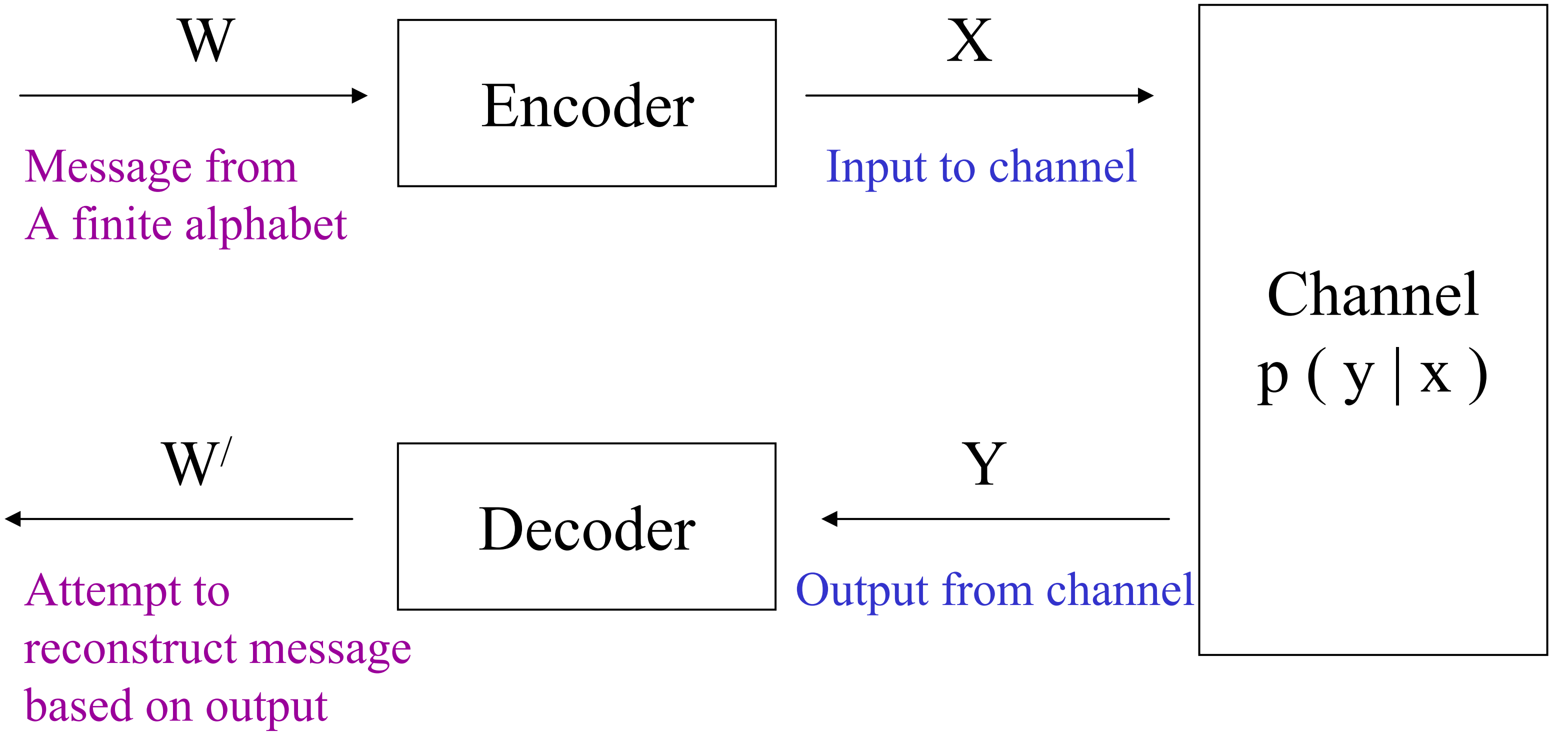
$$= D(p(x,y) \,\|\, p(x)p(y))$$

Define the Conditional Relative Entropy:

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

Essential Information Theory

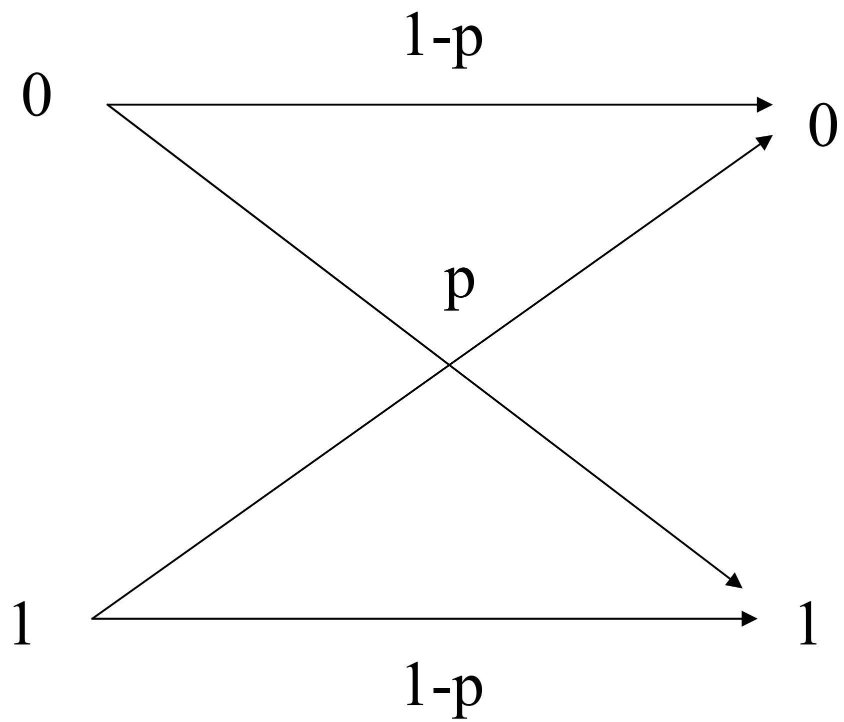
The noisy channel model



The noisy channel model

Essential Information Theory

The noisy channel model



A binary symmetric channel

Essential Information Theory

The noisy channel model

- 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.

$$C = \max_{p(X)} I(X;Y)$$

$$= \max_{p(X)} H(Y) - H(Y \mid X)$$

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

$$if \quad p = 0 \;\; or \;\; p = 1 \Rightarrow C = 1$$

$$if \quad p = \frac{1}{2} \Rightarrow C = 0$$

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

Essential Information Theory

The noisy channel model

Application: (In speech recognition)

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

Bayes' theorem

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

Essential Information Theory

Cross entropy

- If a model captures more of the structure of a language, then the entropy of the model should be lower
- Entropy is a measure of the quality of our models

p	t	k	a	i	u	p	t	k	a	i	u
1/8	1/4	1/8	1/4	1/8	1/8	100	00	101	01	110	111

$$H(P) = -\sum_{i \in \{p,t,k,a,i,u\}} P(i) \log P(i)$$

$$= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}] = 2\frac{1}{2} \ \ bits$$

Essential Information Theory

Cross entropy

- 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:*

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

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

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

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

Essential Information Theory

Cross entropy

Cross entropy of a language :

suppose

Language L = $(X_i) \sim p(x)$ 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:

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

Essential Information Theory

Cross entropy

- Expectation is a weighted average over all possible sequence
- 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

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

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

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

Essential Information Theory

Cross entropy

- 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.*

Essential Information Theory

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."