Principles of Information Science

①点名 ②作业报告 ③开卷 ④随堂考试

Day.1:

1. what do you want to learn from the program?
   basic principle, lastest technique & n. (大人物、ABC)

2. What is information science? What is the difference among others disciplines?
   研究信息的性质和运动规律。
   信息的目的是减少熵(不确定性)

   $$H=-\sum_{i=1}^{N} p_i \log p_i$$

3、Entropy:

Definition of entropy H(X) of a discrete random varible X is: $H(x)=-\sum_{x\in X} P(x)\log P(x)$, we define $0\log 0=0$ $\log_e=\ln$ nats

and max entropy as: when $P=\frac{1}{n}$ $H(x)=-\sum_{i=1}^{n}\frac{1}{n}\cdot\log\frac{1}{n}=-\log P=\log n$

We denote the expectation by E, thus, if X~P(x), the expected value of the random varibles g(x). is $E_Y g(x)=\sum_{y\in X} g(x)\cdot P(x)$

The entropy of X can also be interepreted as the expected values of the random varibles $\log\frac{1}{P(X)}$, where X is drawn according to probability P(x). $H(x)=E_P(\log\frac{1}{P(x)})=-E_P(\log P(x))$

$H(x)\geq 0$, $0\leq P(x)\leq 1$,

$\Rightarrow H_b(x)=(\log_b a)\cdot H_a(x)$

P2

Proof 1:

$$H_b(X) = -E[\log_b P(X)] \quad \log_b a \cdot H_a(X) = -E[\log_a P(X)] \cdot \log_b a$$

$$= -E[\log_a P(X)] \cdot E\left[\frac{1}{\log_a b}\right]$$

$$= -E[\log_b P(X)]$$

Proof 2: Given that $\{X_1, X_2, \cdots, X_n\}$

$$\downarrow \quad \downarrow \quad \cdots \quad \downarrow$$

let $\{y_1, y_2, \cdots, y_n\}$, where $y_1 = -\log p(y_1)$

$y_2 = -\log p(y_2)$

we obtain $E(Y) = \sum_{i=1}^{n} y_i \cdot P(y_i)$

$$= \sum_{i=1}^{n} -\log P(y_i) \cdot P(y_i)$$

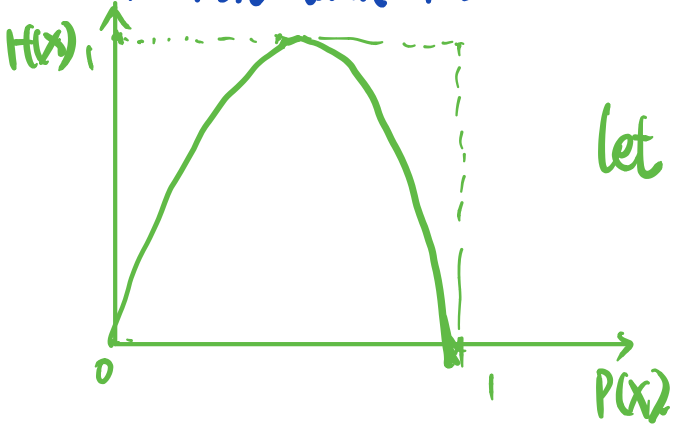
$$= -\sum_{i=1}^{n} P(X_i) \cdot \log P(X_i) = H(X)$$

$\therefore H(X) = E(-\log P(X_i))$

Example: Let $X = \begin{cases} 1 & (p) \\ 0 & (1-p) \end{cases} \Rightarrow H(X) = -p\log p - (1-p)\log(1-p)$

$\triangleq H(p)$

$H(X) = 1$ bit when $P = \frac{1}{2}$



let $X = \begin{cases} \frac{1}{2} & a \\ \frac{1}{4} & b \\ \frac{1}{8} & c \\ \frac{1}{8} & d \end{cases} \Rightarrow H(X) = \frac{7}{4}$ bits

$\downarrow$

$-\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8}$.

P3

joint entropy & conditional entropy.

Definition of joint entropy

$$① \ H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P(X,Y)\log P(X,Y) = -E_{P(x,y)}[\log P(X,Y)]$$

Definition: If $(X,Y) \sim P(X,Y)$, the conditional entropy $H(Y|X)$ is defined as $H(Y|X) = \sum_{x\in X} P(x)H(Y|X=x)$

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

Form ①

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

$$= -\sum_{x\in X, y\in Y} P(x,y)\cdot\log\frac{P(x,y)}{P(x)}$$

Form ②

$$= \sum_{\substack{x\in X\\ y\in Y}} P(x,y)\cdot\log\frac{P(x)}{P(x,y)}$$

Theorem (chain rule)

$$H(X,y) = H(X) + H(Y|X)$$

Proof: $H(X,y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)\cdot\log P(x,y)$

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

$$= -\sum\sum P(x,y)[\log P(x) + \log P(y|x)]$$

$$= -\sum_x\sum_y P(x,y)\log P(X) - \sum\sum P(x,y)\cdot\log P(y|x)$$

$$= -\sum_x \log P(x)\sum_y P(x,y) - \sum\sum P(x,y)\cdot\log P(y|x)$$

$$= -\sum_x \log P(x)\cdot P(x) \overset{\downarrow\,\text{marginal}}{-} \sum\sum P(x,y)\cdot\log P(y|x).$$

$$= H(X) + H(Y|X) = H(y) + H(X|Y).$$

Example: Let (X, Y) have the following distribution:

Y\X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

$\overset{\frac{7}{4}}{H(X)}, \overset{2}{H(y)}, \overset{\frac{11}{8}}{H(X|Y)}, \overset{\frac{13}{8}}{H(Y|X)}$.

$\Rightarrow H(X,y) - \frac{27}{8}$

$$H(X) = -\sum_x P(x) \lg P(x) = -\frac{1}{2} \cdot \lg\left(\frac{1}{2}\right) - \frac{1}{4} \cdot \lg\left(\frac{1}{4}\right) - \frac{1}{8} \cdot \lg\left(\frac{1}{8}\right) - \frac{1}{8} \cdot \lg\left(\frac{1}{8}\right)$$

$$= \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{3}{8} = \frac{7}{4}$$

$$H(y) = -\frac{1}{4} \cdot \log\frac{1}{4} - \frac{1}{4} \cdot \log\frac{1}{4} - \frac{1}{4} \cdot \log\frac{1}{4} - \frac{1}{4} \log\frac{1}{4} = 2$$

$$H(X|Y) = \sum_{x,y \in D} P(x,y) \cdot \log\frac{P(x)}{P(x,y)} = P(X=1, y=1) \cdot \log\frac{P(X=1)}{P(X=1, y=1)} + \cdots$$

$$+ P(X=4, y=4) \cdot \log\frac{P(X=4)}{P(X=4, y=4)} = \frac{11}{8}$$

Notably:

$$H(X|Y) \neq H(Y|X)$$

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

Relative Entropy or K-L Divergence:

Defination: In the context of Machine Learning, DKL(P||Q) is often called the information gain achieved if P is used instead of Q.

Example:

P is often used as true distribution, given [1,0,0].

Q is used for prediction [0.7, 0.2, 0.1].

so we need to make Q as same as P.

Then let $D_{KL}(p\|q) = \sum_{i=1}^{n} p(x_i) \log\left(\frac{p(x_i)}{q(x_i)}\right)$.

$D_{KL}\downarrow$, two mass distribution are more similar.

relative Entropy or **K-L divergence** between two mass distribution P(X) and q(X). Is defined as: $D(p\|q)=\sum_{x\in X} p(x)\cdot\log\frac{p(x)}{q(x)} = E_{p(x)}\log\frac{p(x)}{q(x)}$

$0\log 0/0=0$ if $p=q \Rightarrow D(p\|q)=0$

Proof: $D_{KL}(p\|q)\geq 0$ (via Jensen Inequation.)

$$D_{KL}(p\|q)=\sum_x p(x)\cdot\log\frac{p(x)}{q(x)} = -\sum_x p(x)\cdot\log\frac{q(x)}{p(x)}$$

$$= -E_{p(x)}\left(\log\frac{q(x)}{p(x)}\right) \geq -\log E_{p(x)}\left(\frac{q(x)}{p(x)}\right) = -\log\sum_x p(x)\frac{q(x)}{p(x)}$$

$$= -\log\sum_x q(x)$$

$\because \sum_x p(x)=1 \therefore D_{KL}(p\|q)\geq 0$ ※

Mutual Information:

Definition: consider two random varibles X, Y:

a joint probability mass function P(x,y) and marginal probability mass function P(x) and P(y)

The mutual information: I(X;y) is the relative Entropy between the joint distribution and product information P(x) P(y).

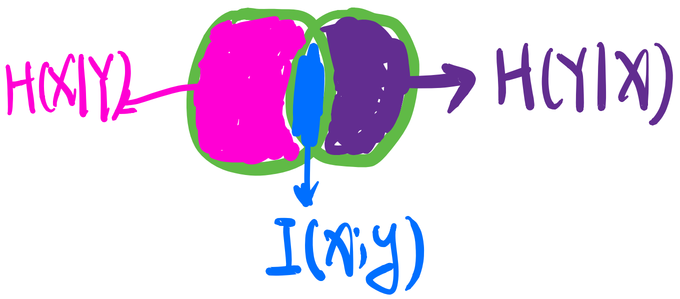
$$I(X;y)=\sum_x\sum_y P(x,y)\cdot\log\frac{P(x,y)}{P(x)\cdot P(y)} = D(P(x,y)\|P(x)\cdot P(y))$$

$$= E_{P(x,y)}\cdot\log\frac{P(x,y)}{P(x)\cdot P(y)}$$

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

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

Venn ⇒



P6 Day 2:

Definition of Information Science

Information Science is a trans-disciplinary science
with information as its domain;
with laws of information process as its content;
with information methodology as its approach;
with strengthening human intelligence as its goal;

The methodology:

M1: Information system analysis approach
M2: information system synthesis approach
M3: information system evolution approach
Criteria 1: Matter-Energy-Information Trinity.
Criteria 2: Structure-Function-Behavior Trinity.

Example: Let $X=\{0,1\}$, and consider two distributions p & q on X.

Let $p(0)=1-r$, $p(1)=r$, $q(0)=1-s$, $q(1)=s$. Then

$$D(p\|q)=(1-r)\cdot\log\frac{1-r}{1-s}+r\cdot\log r/s$$

and $D(q\|p)=(1-s)\log\frac{1-s}{1-r}+s\log s/r$

If $s=r$, then $D(p\|q)=D(q\|p)=0$

If $r=\frac{1}{2}$, $s=\frac{1}{4}$, ✓✓

$D(p\|q)\neq D(q\|p)$

$\parallel$ 0.2075 bit 0.1887 bit

Relationship between entropy & mutual information:

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

$$= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)$$

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

Chain Rule : For Entropy , Relative Entropy and mutual information.

Theorem : Let $x_1, x_2, \cdots, x_n$ be drawn from $p(x_1, x_2, \cdots, x_n)$.

Then $H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, X_2, \cdots, X_{i-1})$

Proof. $H(X_1, X_2) = H(X_1) + H(X_2 | X_1)$

$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3 | X_1)$

$= H(X_1) + H(X_2 | X_1) + H(X_3 | X_2, X_1)$

$\cdots H(X_1, X_2, \cdots, X_n) = H(X_1) + H(X_2 | X_1) \cdots = \sum_{i=1}^{n} (X_i | X_{i-1}, \cdots, X_1)$

Definition : The conditional mutual information of random aribles X & Y given Z is defined by

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

$$= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z) \cdot p(Y|Z)}$$

Chain Rule for information

Theorem : $I(X_1, X_2, \cdots, X_n ; Y) = \sum_{i=1}^{n} I(X_i ; Y | X_{i-1}, X_{i-2}, \cdots, X_1)$

Proof : $I(X_1, X_2, \cdots, X_n ; Y) = H(X_1, \cdots, X_n) - H(X_1, X_2, \cdots, X_n | Y)$

$= \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1) - \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1, Y)$

$= \sum_{i=1}^{n} I(X_i ; Y | X_1, X_2, \cdots, X_{i-1})$

Defination : For joint distribution mass function $P(X,Y)$ and $q(x,y)$, the conditional relative entropy $D(p(y|x) \| q(y|x))$ is the average of the relative entropies between the conditional probability mass function $p(y|x)$ and $p(x|y)$ averaged

P8

over the probability mass function P(X).

$$D(p(y|x)\|q(y|x)) = \sum_x p(x)\cdot\sum_y p(y|x)\cdot\log\frac{p(y|x)}{q(y|x)}$$

$$= E_{p(x,y)}\left[\log\frac{p(y|x)}{q(y|x)}\right]$$

Chain Rule for relative entropy:

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

Proof: $D(p(x,y)\|q(x,y)) = \sum_x\sum_y p(x,y)\cdot\log\frac{p(x,y)}{q(x,y)}$

$$= \sum_x\sum_y p(x,y)\cdot\log\frac{p(x)\cdot p(y|x)}{q(x)\cdot q(y|x)}$$

$$= \sum_x\sum_y \cdot p(x,y)\log\frac{p(x)}{q(x)} + \sum_x\sum_y p(x,y)\log\frac{p(y|x)}{q(y|x)}$$

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

Jensen's inequality and its consequences.

Definition a function f(x) is said to be convex over an interval (a,b) if for every $x_1, x_2 \in (a,b)$, and $0 \le \lambda \le 1$.

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2)$$

A function is said to be strictly convex, if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Definition: A function f is said to be concave if $-f$ is convex.

convex ∪ ✓ $y = x^2, |x|, e^x, x\cdot\log x$

concave ∩ ╱ $y = \sqrt{x}$,

Theorem If the function f has a second derivative that is non-negative (positive) over an interval the function is convex (strictly convex) over this interval.

P9

Proof: $f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{f''(X)}{2}(x-x_0)^2$

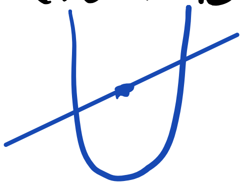
where x^* lies between x_0 and x. By hypothsis $f''(x^*)\geq 0$

we let $x_0=\lambda x_1+(1-\lambda)x_2$ and $x=x_1$. we obtain

$f(x_1)\geq f(x_0)+f'(x_0)\cdot[(1-\lambda)(x_1-x_2)] * \lambda$

$f(x_2)\geq f(x_0)+f'(x_0)\cdot[(1-\lambda)(x_2-x_1)] * (1-\lambda)$

Theorem (Jensen's inequality): If f is a convex function and X is a random varible. $E(f(X))\geq f(E(X))$. Moreover, if f is strictly convex, the equality implies that $X=EX$ with probability 1 (i.e. X is a constant).



Proof: For a two-mass-function; the inequality become $p_1 f(x_1)+p_2 f(x_2)\geq f(p_1x_1+p_2x_2)$

suppose that the theorem is true for distrib with $k-1$ mass points.

Then con $p_i'=p_i/(1-p_k)$, for $i=1,2,\cdots,k-1$, we have

$$\sum_{i=1}^{k} p_i f(x_i)=p_k$$

$$\geq p_k f(x_k)+(1-p_k)f\left(\sum_{i=1}^{k-1} p_i' x_i\right)$$

$$\geq f\left(p_k x_k+(1-p_k)\sum_{i=1}^{k-1} p_i'\cdot x_i\right)=f\left(\sum_{i=1}^{k} p_i x_i\right)$$

Theorem: Let $p(x), q(x)$, $x\in\mathcal{X}$, be two probability mass functions then $D(p\|q)\geq 0$, with equality if and only if $p(x)=q(x)$.

Proof: Let $A=\{x: p(x)>0\}$ be the support set of $p(x)$.

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

$$\geq \log\sum_{x\in A} p(x)\cdot\frac{q(x)}{p(x)}=\log\sum_{x\in A} q(x)\geq \log\sum_{x\in\mathcal{X}} q(x)$$

$$=\log 1=0$$

Corollary: Nonnegativity of mutual information

For two random varibles X, Y, $I(X;Y)\geq 0$, with equality if and only if X and Y are independent.

Proof: $I(x;y) = D(p(x,y)\,\|\,p(x)\cdot p(y)) \geq 0$

$p(x,y) = p(x)\cdot p(y)$, X and y are independent.

Corollary: $D(p(y|x)\,\|\,q(y|x)) \geq 0$, with equality if and only if $p(y|x) = q(y|x)$ for all y and x, such that $p(x)>0$

Corollary: $I(X;Y|Z) \geq 0$, with equality if and only if X and Y are independent given Z.

Theorem: $H(X) \leq \log|X|$, where $|X|$ denotes the number of elements in the range of X, with equality if and only if X has a uniform distribution over X.

Proof: Let $u(x) = \frac{1}{|X|}$ be the uniform mass probability function over X, and let q(x) be the mass probability function for X. Then

$D(p\|u) = \sum p(x)\log\frac{p(x)}{u(x)} = \log|X| - H(X)$

Hence by the relative entropy:

$$0 \leq D(p\|u) = \log|X| - H(X)$$

Day3: Chap 2 The concept & description of information

Problems that should be concerned with

1. How should we define the concept of Information?

   Information is something eliminating untertainty. It is different background that obtaining different definitions/concepts.

2. What are the relationship and difference between Shannon and the Comprehensive Info?

   Shannon is a special situation/state of Comprehensive Info.

3. How to reasonably classify Information?

   Information should be classified by property result in: "Syntactic, Semantic, Pragmatic"

4. How to properly represent Information?

   Considering

认识论

Definition of Epistemological Information

The epistemological information about an object concerned by a subject, is what he described concerning the form, the content, and the value of the states and their relations of the object.

Theorem (Conditioning reduces entropy) [Information can't hurt]

$$H(X|Y) \le H(X)$$

with equality if and only if X and Y are independent

Proof: $0 \le I(X;Y) = H(X) - H(X|Y)$

Example: Let (X,Y) have distribution:

$$\left(\begin{array}{c|cc} Y \backslash X & 1 & 2 \\ \hline 1 & 0 & 3/4 \\ 2 & 1/8 & 1/8 \end{array}\right) \Rightarrow$$

Then $H(X)=0.544$ bit $H(X|Y=1)=0$ bit
$H(X|Y=2)=1$ bit $H(X|Y)=0.25$ bit

Thus, the uncertainty in X is increased if $Y=2$ is observed and decreased if $Y=1$ is observed., but uncertainty decreases on the average.

Theorem: (Independence bound on entopy)

Let $X_1, X_2, \cdots, X_n$ be drawn from $P(X_1, X_2, \cdots, X_n)$. Then $H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$ with equality if and only if all X_i are independent.

Proof: By the chain rule for entropies

$$H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \cdots, X_{i-1}) \leq \sum_{i=1}^{n} H(X_i)$$

LOG sum inequality and its applications

Theorem: (log sum inequality): For nonnegative numbers: $a_1, a_2, \cdots, a_n$ and $b_1, b_2, \cdots, b_n$

$$\sum a_i \log\frac{a_i}{b_i} \geq \left(\sum a_i\right) \log\frac{\sum a_i}{\sum b_i}$$

, with equality if and only if $a_i/b_i = \text{const}$.

Proof: Assume without loss of generality, that $a_i > 0$, $b_i > 0$.

The function $f(t) = t \cdot \log t$. is strict convex.

Since $f''(t) = \frac{1}{t} \cdot \log e > 0$ for all positive t.

Hence by Jenson's inequality, we have $\sum \alpha_i f(t_i) \geq f(\sum \alpha_i t_i)$ for $\alpha_i > 0$, $\sum \alpha_i = 1$. Setting $\alpha_i = b_j / \sum_j b_j$ and $t_i = a_i / b_i$

we obtain $\sum \frac{a_i}{\sum b_j} \log \frac{a_i}{b_j} \geq \sum \frac{a_i}{\sum b_j} \log \sum \frac{a_i}{b_j}$

Theorem (Convexity of relative entropy)

D(p||q) is convex in the pair (p,q): that is, if (p_1, q_1) & (p_2, q_2) are two pairs of mass probability function, then

$$D(\lambda, P_1 + (1-\lambda)P_2 \,\|\, \lambda q_1 + (1-\lambda) q_2) \le \lambda D(P_1 \| q_1) + (1-\lambda) D(P_2 \| q_2)$$

for all $0 \le a \le 1$.

Proof: We apply the log sum inequality to a term on the left-hand side $(\lambda P_1(x) + (1-\lambda) P_2(x)) \log \frac{\lambda P_1(x) + (1-\lambda) P_2(x)}{\lambda q_1(x) + (1-\lambda) q_2(x)}$

$$\le \lambda P_1(x) \log \frac{\lambda P_1(x)}{\lambda q_1(x)} + (1-\lambda) P_2(x) \cdot \log \frac{\lambda P_2(x)}{\lambda q_2(x)}$$

suming over all x, we obtain the desired property.

Theorem (Concavity of Entropy)

H(p) is a concave function of P.

Proof $H(p) = \log|X| - D(p\|u)$, where u is the uniform distribution over |X| outcomes. The concavity of H then follows directly the convexity of P.

Data-processing inquality:

Definition: Random Varibles X, Y, Z are said to form a Markov chain in that order (devoted by X→Y→Z), if the conditional distribution of Z depends only on Y and is conditionally independent of X. Specifically, X, Y, Z form a Markov chain X→Y→Z. if the joint probability mass function can be written as

$$P(X,Y,Z) = P(X) p(y|x)) | P(z|y))$$

Theorem (Data-Processing inequality)

If $X\to Y\to Z$, then $I(X;Y)\geq I(X;Z)$

Proof: $I(X;YZ)=I(X;Z)+I(X;Y|Z)$

$$=I(X;Y)+I(X;Z|Y)$$

since X & Z are conditionally independent given Y, we have $I(X;Z|Y)=0$, since $I(X;Y|Z)\geq 0$, we have

$$I(X;Y)\geq I(X;Z)$$

Corollary: If $X\to Y\to Z$, then $I(X;Y|Z)\leq I(X;Y)$

Proof: $I(X;Z|Y)=0$, by Markovity, and $I(X;Z)\geq 0$. Thus $I(X;Y|Z)\leq I(X;Y)$

Definition: A function $T(X)$ is send to be a sufficient statstic relative to the family $\{f_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution on θ.

[i.e. $\theta\to T(X)\to X$ forms a Markov Chain]

Differential Entropy

Definition: The differential entropy $h(X)$ of a continuous random varible X with density $f(x)$ is defined as $h(X)=-\int_S f(x)\cdot\log f(x)\,dx$, where S is the support set of the random varible.

Example: (uniform distribution)

Consider a random varible distributed uniformly from 0 to a so that its density is $1/a$ from 0 to a and 0 elsewhere. Then the entropy is $h(X)\overset{\leq}{=}-\int_0^a\frac{1}{a}\log\frac{1}{a}\,dx=\log a$.

Note: For $a<1$, $\log a<0$ and the differential entropy is negative.

P15

Normal distribution:

Let $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{x^2}{2\delta^2}}$.

$$h(\phi) = -\int \phi \ln \phi = -\int \phi(x)\left[-\frac{x^2}{2\delta^2} - \ln\sqrt{2\pi\delta^2}\right]$$

$$= \frac{EX^2}{2\delta^2} + \frac{1}{2}\ln 2\pi\delta^2 = \frac{1}{2} + \frac{1}{2}\ln 2\pi\delta^2 = \frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\delta^2$$

$$= \frac{1}{2}\ln 2\pi e\delta^2 \text{ nats}$$

$$h(\phi) = \frac{1}{2}\log 2\pi e\delta^2 \text{ bits}$$

Joint and conditional differential entropy

Definition: the differential entropy of a set $X_1, \cdots, X_n$ of random varibles with density $f(x_1, x_2, \cdots, x_n)$ is defined as

$$h(X_1, X_2, \cdots, X_n) = -\int f(x^n) \log f(x^n)\, dx^n$$

Definition: If X, Y have a joint density function $f(x,y)$, we can define the conditional differential entropy $h(X|Y)$ as

$$h(X|Y) = -\int f(x,y) \log f(y|x) \cdot dxdy$$

$$h(X|Y) = h(x,y) - h(Y)$$

Theorem: Entropy of a mul-normal distribution with mean μ and matrix K. Then, $h(X_1, X_2, \cdots, X_n) = h(N_n(\mu, k) = \frac{1}{2}\log(2\pi e)^n |K|$ bit.
where $|K|$ devotes the determinant of K.

Proof: The probability density function of $X_1, X_2, \cdots, X_n$ is

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|K|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^{-1} K^{-1}(x-\mu)\right\}$$

$$h(f) = -\int f(x)\left[-\frac{1}{2}(x-\mu)^T \cdot K^{-1}(x-\mu) - \ln(\sqrt{2\pi})^n |k|^{\frac{1}{2}}\right] dx$$

$$= \frac{1}{2} E\left[\sum_{i,j}(X_i - \mu_i)(K^{-1})_{ij}(X_j - \mu_j)\right] + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2} E\left[\sum_{i,j}(X_i - \mu_i)(X_j - \mu_j)(K_{ij})^{-1}\right] + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2}\sum_{i,j} E(X_j - \mu_j)(X_i - \mu_i)(K^{-1})_{ij}] + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2}\sum\sum K_{ji}(K^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n|K|$$

$$= \frac{1}{2}n + \frac{1}{2}\ln(2\pi)^n|K|$$

Relative Entropy and Mutual Information:

Defination: The relative entropy (K-L divergence) $D(f\|g)$ between two density f and g defined by $D(f\|g)=\int f\log f/g$

Definition: The mutual information $I(X;Y)$ between two random varibles with joint density $f(x,y)$ is defined as $I(X;Y)=\int f(x,y)\log\frac{f(x,y)}{f(x)\,f(y)}\,dxdy$

Example: Mutual information between correlated Gaussian random varibles with correlation ρ. Let $X,Y\sim N(0,K)$

where $K=\begin{bmatrix}\sigma^2 & \rho\sigma^2\\ \rho\sigma^2 & \sigma^2\end{bmatrix}$ $h(x)=h(y)=\frac{1}{2}\log(2\pi e)\sigma^2$

$h(x,y)=\frac{1}{2}\log(2\pi e)^2|K|=\frac{1}{2}\log(2\pi e)^2\sigma^4(1-\rho^2)$

$\sigma_x^2\cdot\sigma_y^2=\sigma^4$

$\sigma_{xy}^2=\sigma^4\rho^2$

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

$$=\frac{1}{2}\ln\left(\frac{\sigma^4}{\sigma^4-\sigma^4\rho^2}\right)$$

$$=\frac{1}{2}\ln\left(\frac{\sigma^4-\sigma^4\rho^2+\sigma^4\rho^2}{\sigma^4-\sigma^4\rho^2}\right)$$

$$=\frac{1}{2}\ln\left(1+\frac{\sigma^4\rho^2}{\sigma^4-\sigma^4\rho^2}\right)=\frac{1}{2}\ln\left(1+\frac{\rho^2}{1-\rho^2}\right)$$

if $\rho=0$, then $I(X;Y)=0$

if $\rho=\pm1$, then $I(X;Y)$ is infinite.

P17 Day 4:

Disscussion:

① Why is semantic information important ? Understanding.

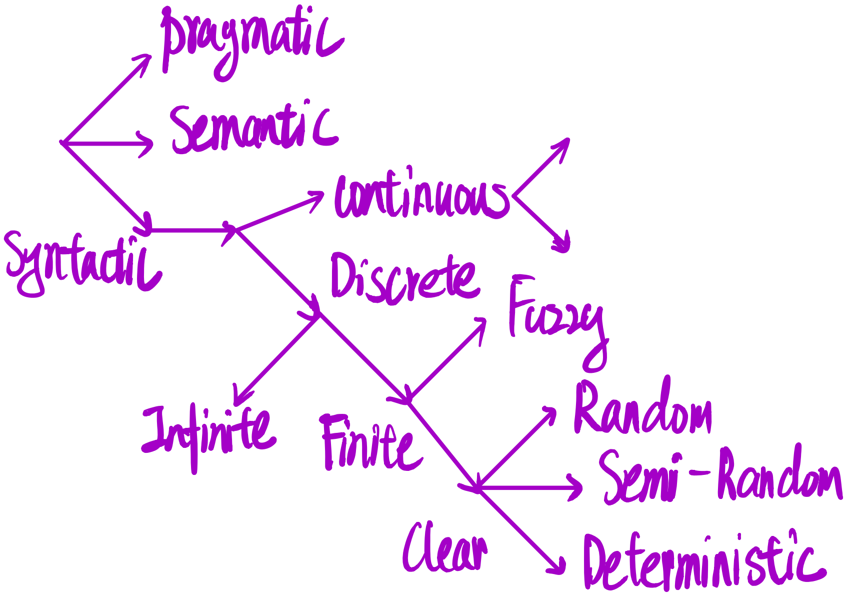
② Why is pragmatic information necessary?

A great challenge:

The sensor can only recieve the Syntactic Information while the Cognitive Thinking needs Semantic Information.

⟶ How can the Semantic Information be produced?

Classification of Information:



Properties of Differential Entropy, Relative Entropy and Mutual Entropy

Theorem: $D(f\|g) \geq 0$, with equality if and only if $f=g$ almost everywhere.

Proof: Let S be the support set of f. Then

$$-D(f\|g) = \int_S f \log g/f$$

$$\leq \log \int_S f \cdot g/f \leq \log \int_S g \leq \log 1 = 0$$

Corollery: $I(X;Y)\geq 0$, with equality iff X and Y are independent.

Corollery: $h(X|Y)\leq h(X)$ with equality iff X and Y are independent.

Theorem: (Chain Rule for differential entropy)

$$h(X_1, X_2, \cdots, X_n)=\sum_{i=1}^{n} h(X_i|X_1, X_2, \cdots, X_{i-1})$$

Theorem: $H(X+c)=H(X)$: Translation doesnot change the differential entropy.

Theorem: $H(aX)=H(X)+\log|a|$

Proof: Let $Y=aX$, then $f_Y(y)=\frac{1}{|a|}f_X(y/a)$, and

$$H(aX)=-\int_Y f_Y(y)\cdot \log f_Y(y)\cdot d$$

$$=-\int \frac{1}{|a|}f_X(y/a)\log\left(\frac{1}{|a|}f_X(y/a)\right)\cdot dy$$

$$=-\int f_X(x)\cdot \log f_X(x)\cdot dx+\log|a| = h(X)+\log|a|$$

Corollery: $h(AX)=h(X)+\log|A|$

Theorem: Let the random vector $X\in R^n$, have zero mean and convariance $K=E[XX^t]$. Then $h(X)\leq\frac{1}{2}\log(2\pi e)^n|K|$ with equality iff $X\sim N(0,K)$. $K_{ij}=E[X_i, X_j]$. where $1\leq i$, $1\leq j<n$.

Proof: Let $g(x)$ be any density satisfy $\int g(x)\cdot x_i x_j\, dx=K_{ij}$ for all i,j. Let ϕ_K be the density of a $N(0,K)$ vector as given in where we set $\mu=0$. Note that $\log\phi_K$ is a quadratic form and $\int x_i x_j \phi_K(x)\cdot dx=K_{ij}$. Then

$$0\leq D(g\|\phi_K)=\int g\log(g/\phi_K)=-h(g)-\int g\log\phi_K$$

$$=-h(g)-\int\phi_K d\phi_K=-h(g)+h(\phi_K)$$

where the substition $\int g\log\phi_K=\int\phi_K\cdot\log\phi_K$ follows from the fact that g and ϕ_K yields the same moments.

of the quadratic form $\log p_K(x)$.

Theorem: Estimation Error and differential Entropy.

For any random varible X and estimated $\hat{X}$.

$$E[X-\hat{X}]^2 \geqslant \frac{1}{2\pi e} e^{2h(X)}$$

. with equality iff X is Gaussian and $\hat{X}$ is the means of X.

Proof: Let $\hat{X}$ be any estimator of X; then

$$E(X-\hat{X})^2 \geqslant \min_{\hat{X}} E[X-\hat{X}]^2 = E(X-EX)^2 = var(X)$$
$$\geqslant 2\pi e\, e^{2h(X)}$$

where follows from the fact that the mean of X is the best estimator for X and the last quantity follows from the fact that the Gaussian distrib has the maximum entropy for a given variance. We have eqality iff $\hat{X}$ is the best estimator.

Corolley: Given side information Y and estimator $\hat{X}(Y)$, it follows that $E[X-\hat{X}(Y)]^2 \geqslant \frac{1}{2\pi e} e^{2h(X|Y)}$

Maximum Entropy Distributions:

Consider the following problem: Maximum the entropy $h(f)$ over all probability densities f satisfying:

1. $f(x) \geqslant 0$ with equality outside the support set S.
2. $\int_S f(x)dx = 1$
3. $\int_S f(x) r_i(x) \cdot dx = d_i$ for $i \leq 1 \leq m$

Thus, f is a density on support set S meeting certain moment constraint $\alpha_1, a_2, \cdots, \alpha_m$.

(Calculus): The differential entropy $h(f)$ is a concave function over a convex set. We form the functional

$J(f) = -\int f \ln f + \lambda_0 \int f + \sum_{i=1}^{m} \lambda_i \int f r_i$. and differentiate. with respect to $f(x)$, the xth component of f. to obtain $\frac{\partial J(f)}{\partial f(x)} = -\ln f(x) - 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x)$.

setting this equal to zero, we obtain the form of the maximum density: $f(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)}$ $x \in S$. where $\lambda_0, \lambda_1, \cdots, \lambda_m$ are chosen so that f satisfy the constrains

Information inequality: If g satisfies conditions and if f^* is of the form exp $\cdots$, then $0 \le D(g\|f^*) = -h(g) + h(f^*)$. Thus. $h(g) \le h(f^*)$, for all g satisfies the constrains

Theorem: (Maximum entropy distribution): Let $f^*(x) = f_\lambda(x) = e^{\lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x)}$; $x \in S$, where $\lambda_0, \cdots, \lambda_m$ are chosen so that f^* satisfied conditions. Then f^* uniquely maximize $h(f)$ over all probability densities f satisfying constrains.

Proof: Let g satisfy the constrains. Then $h(g) = -\int g \log g$

$$= -\int g \ln \frac{g}{f^*} \cdot f^* = -D(g\|f^*) - \int g \ln f^* \overset{(a)}{\le} -\int g \ln f^*$$

$$\overset{(b)}{=} -\int g\left(a_0 + \sum_{i=1}^{m} \lambda_i r_i\right) \overset{(c)}{=} -\int f^*\left(\lambda_0 + \sum_{i=1}^{m} \lambda_i r_i\right)$$

where (a) follows from the nonnegative of f^*

(b) follows from the definition of f^*

(c) follows from the fact that both f^* and g satisfy the constrains.

Examples: Let the constraints be $EX=0$ & $EX^2=\delta^2$, The form of the maximum distribution is

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2}$$

maximize the entropy is $N(0, \delta^2)$ distribution.

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}$$

Example: Let $S=\{1,2,\cdots,6\}$. The distribution the maximize the Entropy is the uniform distrib $P(X)=\{\frac{1}{6}\}$. for all $\in S$.

Example: Let $S=[a,b]$. The distribution the maximize the Entropy is the uniform distrib over the range

Example: Let $S=[0,+\infty)$, $EX=\mu$. Then the entropy-maximization distrib is $f(x)=\frac{1}{\mu}e^{-\frac{x}{\mu}}$ $(x\geq 0)$

Example: $S=(-\infty,+\infty)$, $EX=\alpha_1$, $EX^2=\alpha_2$. The max entropy distribution: $N(\alpha_1,\alpha_2-\alpha_1^2)$

Example: $S=R^n$, $EX_iX_j=K_{ij}$ $1\leq i,j\leq n$,

$$f(x)=e^{\lambda_0+\sum_{i=1}^{n}\lambda_{ij}x_ix_j}$$

$$f(x)=\frac{1}{(\sqrt{2\pi})^n|K|^{1/2}}e^{-\frac{1}{2}x^TK^{-1}x}$$

$$h(N_n(0,K))=\frac{1}{2}\log(2\pi e)^n|K|$$

Anomelous maximum entropy distribution. Maximize the entropy subject to the constraints: $\int_{-\infty}^{+\infty}f(x)\cdot dx=1$. $\int_{-\infty}^{+\infty}xf(x)\cdot dx=\alpha_1$

$$\int x^2f(x)\cdot dx=\alpha_2 \quad \int x^3f(x)\cdot dx=\alpha_3$$

$$f(x)=e^{\lambda_0+\lambda_1x+\lambda_2x^2+\lambda_3x^3}$$

But if X_3 is nonzero $\int f(x)=\infty$, the density cannot be normalized.

Entropy rate of A Gaussian Process.

Definition: The differential entropy rate of a stochastic process $\{X_i\}$, $X_i\in R$ is defined to be

$$h(X)=\lim_{n\to+\infty}\frac{h(X_1,X_2,\cdots,X_n)}{n}$$ if the limit exists.

$$h(X)=\lim_{n\to+\infty}\frac{h(X_1,X_2,\cdots,X_n)}{n}=\lim_{n\to+\infty}h(X_n|X_{n-1},\cdots,X_1)$$

For a stationary Gaussian stochastic process. we have $h(X_1, X_2, \cdots, X_n) = \frac{1}{2}\log(2\pi e)^n |K^{(n)}|$

where the convariance matrix $K^{(n)}$

$h(X) = \frac{1}{2}\log 2\pi e + \frac{1}{4\pi}\int_{-\pi}^{\pi} \log S(\lambda)\, d\lambda$

The entropy rate is also $\lim\limits_{n\to\infty} h(X_n|X^{n-1})$. Since the stochastics process is Gaussian, the conditional distribution is also Gaussian. And hence the Conditional entropy is $\frac{1}{2}\log 2\pi e \sigma_\infty^2$, where σ_∞^2 is the variance of the error in the best estimate of X_n give the infinite past. Thus $\sigma_\infty^2 = \frac{1}{2\pi e} 2^{2h(X)}$

Hence, the entropy rate compresses to the minimium mean-squared error of the best estimator of a sample of the process give the information past.

P23 Day5：考试：开放题型 + 计算题

Exercises for examination:

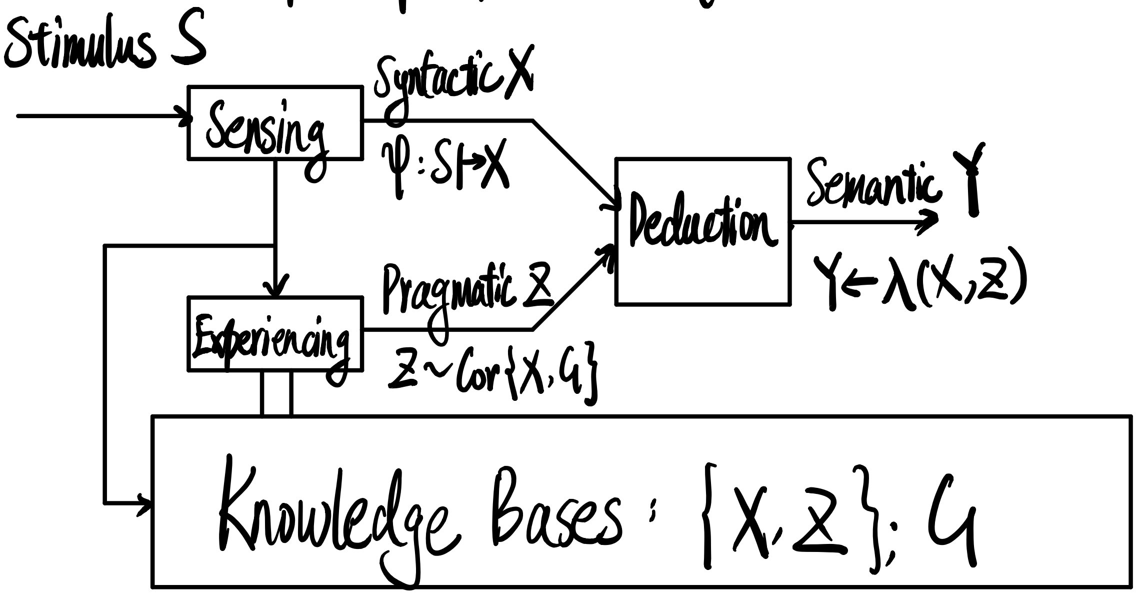
1. Why can only syntactic information be sensed?

2. Please state the principles of pattern recognition.

3. What are the principles of machine learning? Do you find any problem in present of machine learning?

4. Do you have any idea to improve the quality of machine learning in general?

Fundamental Principles of Information Acquisition



Synatic Information Acquisition:

Definition Acquisition:

A process able to obtain the syntactic information concerned.

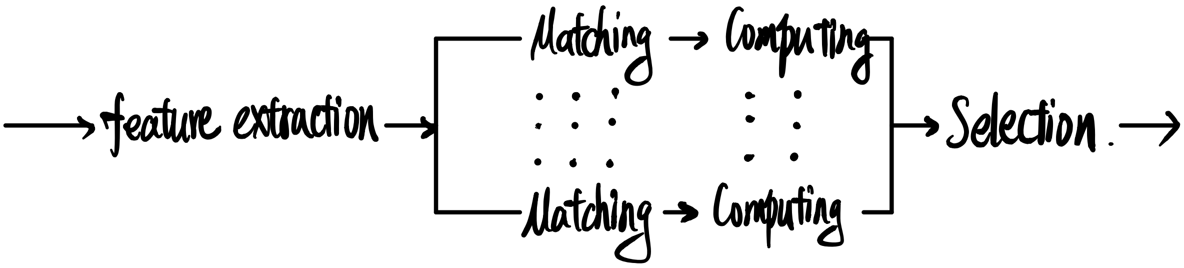
It may consist of a number of steps:

1) Sensing: exist?
2) Recognition: which class according to formal information?
3) Representation: how to display?

Category of Information Acquisition:

①: sensing ② Classification ③ Machine Learning.

Model of Classification:



Definition: A stochastic process is said to be statis if the joint distribution of any subset of the sequences of random varibles is invariant with respect to shifts in the time index; that is.

$$Pr\{X_1=x_1, X_2=x_2, \cdots, X_n=x_n\} = Pr\{X_{1+l}=x_1, X_{2+l}=x_2, \cdots, X_{n+l}=x_2\}$$

for every n and every shift l and for all $x_1, x_2, \cdots, x_n \in X$

Definition: A discrete stochastic process $X_1, X_2, \cdots$ is said to be a Markov chain or a markov process if for $n=1,2,\cdots$

$$\cdot\ Pr(X_{n+1}=x_{n+1} \mid X_n=x_n, X_{n-1}=x_{n-1}, \cdots, X_1=x_1) = Pr(X_{n+1}=x_n \mid X_n=x_n)$$

for all $x_1, x_2, \cdots, x_n, x_{n+1} \in X$, $P(x_1, x_2, \cdots, x_n) = P(x_1)P(x_2|x_1)\cdots P(x_n|x_{n-1})$

Definition: The Markov chain is said to be time invariant if the conditional probability $P(X_{n+1}|X_n)$ does not depend on X_n, that is, for $n=1,2,\cdots$

$$Pr[X_{n+1}=b|X_n=a]=Pr(X_2=b|X_1=a) \text{ for all } a,b\in X$$

Usually, Markov chain is time invariant unless otherwise stated.

If $\{X_i\}$ is a Markov chain, X_n is called the state of time n.

A time-invariant Markov chain is characterized by its limitial state and a probability transfer matrix $P=[P_{ij}]$,

$$\text{where } P_{ij}=Pr\{X_{n+1}=j|X_n=i\}$$

If it is possible to go with position probability from any state of the Markov chain to any state in a finite number steps. The Markov chain is said to be irreducible. If the longest connection factor the lengths of different path from a state to itself is 1. The Markov chain is said to be aperiodic.

Definition: The entropy rate of a stochastic process $\{X_i\}$ is defined by

$$H(\chi)=\lim_{n\to\infty}\frac{1}{n}H(X_1,X_2,\cdots,X_n)$$, when the limit exists.

1. Typewritter

Consider the case of a typewritter that has m equally likely output letter. The typewritter can produce m^n sequence of length n, all of them equally likely. Hence $H(X_1,X_2,\cdots,X_n)=\log m^n$ and the entropy rate is $H(\chi)=\log m$ bits per symbol.

2. $X_1,X_2\cdots$ are i.i.d. random varibles. Then

$$H(X)=\lim\frac{H(X_1,X_2,\cdots,X_n)}{n}=\lim\frac{nH(X_1)}{n}=H(X_1)$$

Therom : For a stationary stochastic process, the limits:

① $H(X)=\lim_{n\to\infty}\frac{1}{n}H(X_1,X_2,\cdots,X_n)$ ② $H'(X)=\lim_{n\to\infty}H(X_n|X_{n-1},X_{n-2},\cdots,X_1)$

exist and are equal : $H(X)=H'(X)$

Models, statistical inference and Learing

Statistical inference, or "learning" as it it called in computer science, is the process of using data to infer the distribution that generated the data. A typical statistical inference question is:

Given a sample $X_1,\cdots,X_n\sim F$, how do we infer F?

In some cases, we may want to infer only some feature of F such as its mean.

Parametric and Nonparametric Models:

A statistical model F is a set of distributions (or density or regression functions).

A parametric model is a set F that can be parameterized by a finite number of parameters. For example, if we assume that the data come from a Normal distribution, then the model is

$$F=\left\{f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},\ \mu\in\mathbb{R},\ \sigma>0\right\}.$$

This is two-parameter model: $F=\{f(x;\theta):\theta\in\Theta\}$, where θ is an unknown parameter (or vector of parameter) that can take values in parameter space Θ. If θ is a vector but we are only interested in one component of θ, we call the remaining parameters nuisance parameters.

A nonparametric model is a set F that cannot be parameterized by a finite number of parameters. For example, $F_{all}=\{\text{all CDF's}\}$ is nonparametric.

Example (One-dimensional Parametric Estimation): Let $X_1, \cdots, X_n$ be independent Bernoulli(p) observations. The problem is to estimate the parameter p.

Example: (Two-dimensional Parametric Estimation): Suppose that $X_1, \cdots, X_n \sim F$ and we assume that the PDF $f \in F$, where F is given in $\{\mu, \delta\}$. The goal is to estimate parameters from the data. If we are only interested in estimating μ, then μ is the parameter of interest and δ is a nuisance parameter.

Example: (Regression, prediction, and classification).

Suppose we observe pairs of data $(X_1, Y_1), \cdots, (X_n, Y_n)$. Perhaps X_i is the blood pressure of subject i and Y_i is how long they live. X is called a predictor or regressor or feature or the dependent variable. We call $r(x) = E(Y|X=x)$ the regression function. If we assume that $r \in F$, where F is finite dimensional — the set of straight lines for example — then we have a parametric regression. If we assume that $r \in F$, where F is not define dimensional then we have a nonparametric regression model. The goal of predicting Y for a new patient based on their X value is called prediction. If Y is discrete (for example, live or die) then prediction is instead called classification. If our goal is to estimate the function r, then we call this regression or curve estimation. Regression models are sometimes written as $Y = r(X) + \epsilon$, where $E(\epsilon) = 0$.

define $\epsilon = Y - r(X)$ and hence $Y = Y + r(X) - r(X) = r(X) + \epsilon$. Moreover,

$$E(\epsilon) = EE(\epsilon|X) = E(E(Y - r(X)|X) = E(E(Y|X) - r(X))$$
$$= E(r(X) - r(X)) = 0.$$

Point Estimation:

Definition: A point estimator $\hat{\theta}_n$ of a parameter θ is consistent if $\hat{\theta}_n \xrightarrow{P} \theta$.

The distribution of $\hat{\theta}_n$ is called the sampling distribution. The standard deviation of $\hat{\theta}_n$ is called the standard error. denoted by se:

$$se = se(\hat{\theta}_n) = \sqrt{V(\hat{\theta}_n)}$$

Often, the standard error depends on the unknown F. In those cases, se is an unknown quantity but we usually can estimate it. The estimated standard error is denoted by $\hat{se}$:

Example. Let $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$ and let $\hat{p}_n = n^{-1}\sum_i X_i$.

Then $E(\hat{p}_n) = n^{-1}\sum_i E(X_i) = p$, so $\hat{p}_n$ is unbiased. The standard error is $se = \sqrt{V(\hat{p}_n)} = \sqrt{p(1-p)/n}$, The estimated standard error is $\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$.

The quality of a point estimate is sometimes assessed by the mean squared error, or MSE defined by: $MSE = E_\theta(\hat{\theta}_n - \theta)^2$. Keep in mind that $E_\theta(\cdot)$ refers to expectation with respect to the distribution:

$$f(x_1, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

that generated the data. It does not mean we are averaging over a distribution for θ.

Theorem: The MSE can be written as $MSE = bias^2(\hat{\theta}_n) + V_\theta(\hat{\theta}_n)$

Proof: Let $\bar{\theta}_n = E_\theta(\hat{\theta}_n)$. Then

$$E_\theta(\hat{\theta}_n - \theta)^2 = E_\theta(\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2 = E_\theta(\hat{\theta}_n - \bar{\theta}_n)^2 + 2(\bar{\theta}_n - \theta) \cdot E_\theta(\hat{\theta}_n - \bar{\theta}_n) + E_\theta(\bar{\theta}_n - \theta)^2 = (\theta_n - \theta)^2 + E_\theta(\hat{\theta}_n - \bar{\theta}_n)^2$$

$= bias^2(\hat{\theta}_n) + V(\hat{\theta}_n)$, where we have used the fact that

$$E_\theta(\hat{\theta}_n - \bar{\theta}_n) = \bar{\theta}_n - \bar{\theta}_n = 0$$

Theorem: If bias → 0 and se → 0 as $n \to \infty$ then $\hat{\theta}_n$ is consistent. that is $\hat{\theta}_n \xrightarrow{P} \theta$

proof: If bias → 0 and se → 0 then, theorem above, MSE → 0. It follows that $\hat{\theta}_n \xrightarrow{qm} \theta$.

Example: Returning to the coin flipping example, we have that $E_p(\hat{p}_n) = p$ so the bias $= p - p = 0$ and $se = \sqrt{p(1-p)/n} \to 0$. Hence, $\hat{p}_n \xrightarrow{P} p$, that is, $\hat{p}_n$ is a consistent estimator.

Definition: An estimator is asymptotically Normal if

$$\frac{\hat{\theta}_n - \theta}{se} \sim N(0,1)$$

Confidence Sets:

A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a,b)$, where $a = a(X_1, \cdots, X_n)$ and $b = b(X_1, \cdots, X_n)$ are functions of the data such that $P_\theta(\theta \in C_n) \geq 1-\alpha$, for all $\theta \in \Theta$.

In words, (a,b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the converage of the confidence interval.

Warning! C_n is random and θ is fixed.

Day 6: Congnition based on CLT.

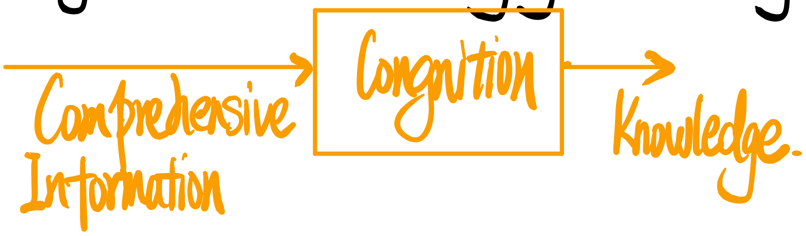
1.Definition of Cognition:

Congition: action, or process of acquiring knowledge, by reasoning or by intuision or through the sense

Cognition is the process for producing knowledge from information through the ways of refining, mainly various kinds of induction.

Information: The state & the varying manner among the states.

Knowledge: The state & the varying law among the states.



2. Knowledge Theroy: Part one

Knowledge can only be the product of epistemological Information, i.e., comprehensive information.

Estimating the CDF and Statistical Functionals

We will consider nonparametric estimation of the CDF F.

Definition: The empirical distribution function $\hat{F}_n$ is the CDF that puts mass $1/n$ at each data point X_i. Formally,

$$\hat{F}_n(x) = \frac{\sum_{i=1}^{n} I(X_i \le x)}{n}, \text{ where } I(X_i \le x) = \begin{cases} 1 & \text{if } X_i \le x \\ 0 & \text{if } X_i > 0 \end{cases}$$

Theorem: At any fixed value of x.

$$E(\hat{F}_n(x)) = F(x), \qquad \hat{F}_n(x) \xrightarrow{P} F(x)$$

$$V(\hat{F}_n(x)) = \frac{F(x)(1-F(x))}{n}$$

$$MSE = \frac{F(x)(1-F(x))}{n} \longrightarrow 0$$

P31

A statistical functional $T(F)$ is any function of F. Examples are the mean $\mu = \int x \, dF(x)$, the variance $\sigma^2 = \int (x-\mu^2) dF(x)$ and the median $m = F^{-1}(1/2)$

Definition: If $T(F) = \int r(x) \cdot dF(x)$ for some function $r(x)$ then T is called a linear functional.

$$T(aF+bG) = aT(G)+bT(G)$$

Parametric Inference.:

Parametric models, that is, models of the form:

$\mathfrak{F} = \{ f(x;\theta) : \theta \in \Theta \}$, where the $\Theta \in R^k$ is the parameter space and $\theta = (\theta, \cdots, \theta_k)$ is the parameter. The problem of inference then reduces to the problem of estimating the parameter θ.

Often, we are only interested in some function $T(\theta)$. For example, if $X \sim N(\mu, \sigma^2)$ then the parameter of interest and σ is called a nuisance parameter.

Example: Let $X_1, \cdots, X_n \sim \text{Normal}(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the parameter space is $\Theta = \{(\mu, \sigma) : \mu \in R, \sigma > 0\}$. Suppose that X_i is the outcome of a blood test and suppose we are interested in τ, the fraction of the population whose test score is larger than 1. Let Z denote a standard Normal random varible. Then

$$\tau = P(X>1) = 1 - P(X<1) = 1 - P\left(\frac{X-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right)$$

$$= 1 - P\left(Z < \frac{1-\mu}{\sigma}\right) = 1 - \phi\left(\frac{1-\mu}{\sigma}\right)$$

The parameter of interest is $\tau = T(\mu, \sigma) = 1 - \Phi((1-\mu)/\sigma))$

Example: Recall that X has a Gamma(α, β) distribution if

$$f(x; \alpha, \beta) = \frac{1}{\beta^{\alpha} \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

where $\alpha, \beta > 0$ and: $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$.

is the Gamma function. The parameter is $\theta = (\alpha, \beta)$. The Gamma distribution is sometimes used to model lifetimes to people, animals, and electronic equipment. Suppose we want to estimate the mean lifetime. Then $T(\alpha, \beta) = E_\theta(X_1) = \alpha\beta$.

Method of Moments:

Suppose that the parameter $\theta = (\theta_1, \cdots, \theta_k)$ has k components. For $1 \le j \le k$,

Define the j^{th} moment:

$$\alpha_j \equiv \alpha_i(\theta) = E_\theta(X^j) = \int x^j \, dF_\theta(x)$$

and the j^{th} sample moment:

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

Definition: The method of moments estimator $\hat{\theta}_n$ is defined to be the value of θ such that:

$$\begin{aligned} \alpha_1(\hat{\theta}_n) &= \hat{\alpha}_1 \\ \alpha_2(\hat{\theta}_n) &= \hat{\alpha}_2 \\ &\vdots \\ \alpha_k(\hat{\theta}_n) &= \hat{\alpha}_k \end{aligned}$$

Formula above defines a system of k equations with k unknowns.

Example: Let $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$. Then $\alpha_1 = E_p(X) = p$ and $\hat{\alpha}_1 = n^{-1}\sum_{i=1}^{n} X_i$.

By equating these we get the estimator: $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Example: Let $X_1, \cdots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then, $\alpha_1 = E_\theta(X_1) = \mu$, and $\alpha_2 = E_\theta(X_i^2) = V_\theta(X_1) + (E_\theta(X_1))^2 = \sigma^2 + \mu^2$. We need to solve the equations:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \\ \hat{\sigma}^2 + \hat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \end{aligned}$$

Maximum Likelihood:

Definition: The likelihood function is defined by $\mathcal{L}_n(\theta)=\prod_{i=1}^{n} f(X_i;\theta)$.

The log-likelihood function is defined by $\ell_n(\theta)=\log\mathcal{L}_n(\theta)$.

The likelihood function is just the joint density of the data, except that we treat it is a function of the parameter θ. Thus, $\mathcal{L}: \Theta \to [0,\infty)$. The likelihood function is not a density function: in general, it is not true that $\mathcal{L}_n(\theta)$ integrates to 1 (with respect to θ).

Definition: The maximum likelihood estimator MLE, denoted by $\hat{\theta}_n$, is the value of θ that maximizes $\mathcal{L}_n(\theta)$.

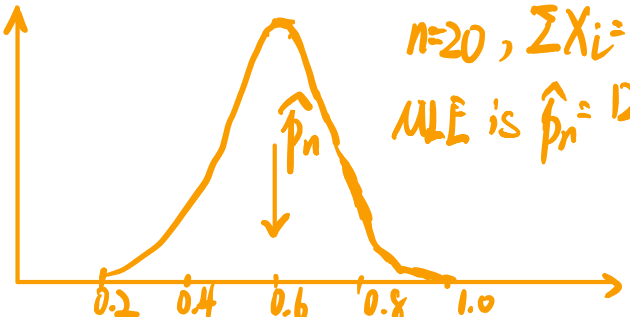
The maximum of $\ell_n(\theta)$ occurs at the same place as the maximum or $\mathcal{L}_n(\theta)$, so maximizing the log-likelihood leads to the same answer as maximizing the likelihood. Often, it is easier to work with log-likelihood. If we multiply $\mathcal{L}_n(\theta)$ by any positive constant c (not depending on θ) then this will not change the MLE. Hence, we shall often drop constants in the likelihood function.

Example: Suppose that $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$. The probability function is $f(x;p)=p^x(1-p)^{1-x}$ for $x=0,1$. The unknown parameter is p. Then,

$$\mathcal{L}_n(\theta)=\prod_{i=1}^{n} f(X_i;p)=\prod_{i=1}^{n} p^{X_i}\cdot(1-p)^{1-X_i}=p^S(1-p)^{n-S}, \text{ where } S=\sum_i X_i.$$

Hence, $\ell_n(p)=S\log p+(n-S)\log(1-p)$.

Take the derivative of $\ell_n(p)$, set it equal to 0 to find that the MLE is $\hat{p}_n=S/n$



$n=20$, $\sum X_i=12$

MLE is $\hat{p}_n=12/20=0.6$

P34

Example : Let $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. The parameter $\theta = (\mu, \sigma)$ and the likelihood function (ignoring some constants) is :

$$\mathcal{L}_n(\mu, \sigma) = \prod_i \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(X_i - \mu)^2\right) = \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2\right\}$$

$$= \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\bar{X} - \mu)^2}{2\sigma^2}\right\},$$

where $\bar{X} = \frac{1}{n} \sum_i X_i$ & $S^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2$

the log likelihood is :

$$\ell(\mu, \sigma) = -n \log \sigma - nS^2 / 2\sigma^2$$

Solving the equations : $\frac{\partial \ell(\mu, \sigma)}{\partial \mu} = 0$ & $\frac{\partial \ell(\mu, \sigma)}{\partial \sigma} = 0$

$\therefore \mu = \bar{X}, \hat{\sigma} = S,$

The frequentist method:

F1: Probability refers to limiting relative frequencies. Probability are objective properties of the real world.

F2: Parameter are fixed, unknown constants. Because they are not fluctuating, no useful probability statements can be made about parameter.

F3: Statistical procedures should be designed to have well-defined long run frequency properties. For example, a 95 percent confidence interval should trap the true value of the parameter with limiting frequency at least 95 percent.

The Bayesian Inference:

B1: Probability describes degree of belief, not limiting frequency. As such, we can make probability statement about lots of things, not just data which are subject to random variation.

B2: We can make probability statements about parameters, even though they are fixed constants.

B3: We make inferences about a parameter θ by producing a probability distribution for θ. Inferences, such as point estimates, may then be extracted from this distribution.

Day 7: The Bayesian Method:

Bayesian inference is usually carried out in the following way:

1. We choose a probability density $f(\theta)$ — called the **prior distribution** — that expresses our beliefs about a parameter θ befor we see any data.
2. We choose a statistical model $f(x|\theta)$ that reflects our beliefs about x given θ. Notice that we know write this as $f(x|\theta)$ instead of $f(x;\theta)$.
3. After observing data $X_1,\cdots,X_n$, we update beliefs and calculate the posterior distribution $f(\theta|X_1,\cdots,X_n)$.

Suppose that θ is discrete and that there is a single, discrete observation X. Bayes' Theorem:

$$P(\Theta=\theta|X=x) = P(X=x,\Theta=\theta)/P(X=x)$$

$$= P(X=x|\Theta=\theta)P(\Theta=\theta)/\sum_\theta P(X=x|\Theta=\theta)\cdot P(\Theta=\theta)$$

The version for continuous variables is:

$$f(\theta|x) = \frac{f(x|\theta)\cdot f(\theta)}{\int f(x|\theta)\cdot f(\theta)\cdot d\theta}$$

If we have n IID observations $X_1,\cdots,X_n$, we replace $f(x|\theta)$ with ⇒

$$f(x_1,\cdots,x_n|\theta) = \prod_{i=1}^n f(x_i|\theta) = \mathcal{L}(\theta)$$

Notation: We will write X^n to mean $(X_1,\cdots,X_n)$ and x^n to mean $(x_1,\cdots,x_n)$.

$$\Rightarrow f(\theta|x^n) = f(x|\theta)f(\theta)/\int f(x|\theta)f(\theta)d\theta = \frac{\mathcal{L}_n(\theta)f(\theta)}{c_n} \propto \mathcal{L}_n(\theta)f(\theta)$$

where $c_n = \int \mathcal{L}_n(\theta)f(\theta)\,d\theta$ is called normalizing constant.

Note that c_n does not depend on θ. We can summarize by writing.

Posterior is proportional to likelihood times Prior:

$$f(\theta|X^n) \propto \mathcal{L}(\theta) f(\theta).$$

Example: Let $X_1, X_2, \cdots, X_n \sim$ Bernoulli (p). Suppose we take the uniform distribution $f(p)=1$ as a prior. By Bayes' theorem, the posterior has the form:

$$f(p|x^n) \propto f(p)\mathcal{L}_n(p) = p^s (1-p)^{n-s} = p^{s+1-1} \cdot (1-p)^{n-s+1-1},$$

where $s = \sum_{i=1}^{n} x_i$ is the number of successes. Recall that a random variable has a Beta distribution with parameter α and β if its density is $f(p;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot p^{\alpha-1}(1-p)^{\beta-1}$. we see that the posterior for p is a Beta distribution with parameters $s+1$ and $n-s+1$. That is. $f(p|x^n) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \cdot p^{(s+1)-1} \cdot (1-p)^{(n-s+1)-1}$

We write this as $p|x^n \sim \text{Beta}(s+1, n-s+1)$

Notice that we have figured out the normalizing constant without actually doing the integral $\int \mathcal{L}_n(p) f(p) \cdot dp$. The mean of a Beta (α,β) distribution is $\alpha/(\alpha+\beta)$ so the Bayes estimator is $\bar{p} = \frac{s+1}{n+2}$.

It is instructice to rewrite the estimator as $\bar{p} = \lambda_n \hat{p} + (1-\lambda_n) \cdot \tilde{p}$

The posterior mean is $\bar{p} = \alpha+s/\alpha+\beta+n = (n/\alpha+\beta+n)\hat{p} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right) \cdot p_0$

where $p_0 = \alpha/\alpha+\beta$ is the prior mean.

Example: Let $X_1, \cdots, X_n \sim N(\theta, \sigma^2)$. For simplicity, let us assume that σ is known. Suppose we take as a prior $\theta \sim N(a, b^2)$. The posterior for θ is $\theta|X^n \sim N(\bar{\theta}, r^2)$, where $\bar{\theta} = w\bar{X} + (1-w)a$, $w = \frac{1/se^2}{1/se^2 + 1/b^2}$, $\frac{1}{\tau^2} = \frac{1}{se^2} + \frac{1}{b^2}$ and $se = \sigma/\sqrt{n}$ is the standard error of the MLE $\bar{X}$.

Statistical Decision Theory.

We have considered several point estimators such as the maximum likelihood estimator, the method of moment estimator, and the posterior mean.

In fact, there are many other ways to generate estimators. How do we choose among them?

When the loss function is squared error, the risk is just MSE (mean squared error): $R(\theta, \hat{\theta}) = E_\theta(\theta - \hat{\theta})^2 = MSE = V_\theta(\hat{\theta}) + bias^2_\theta(\hat{\theta})$

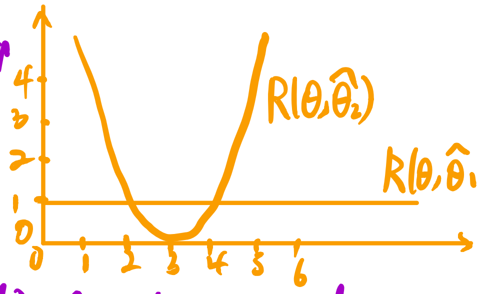
Comparing Risk Functions:

To compare two estimators we can compare their risk functions. However, this does not provide a clear answer as to which estimator is better.

Example: Let $X \sim N(\theta, 1)$ and assume we are using squared error loss.

Consider two estimators: $\hat{\theta}_1 = X$ and $\hat{\theta}_2 = 3$. The risk functions are $R(\theta, \hat{\theta}_1) = E_\theta(X - \theta)^2 = 1$ and $R(\theta, \hat{\theta}_2) = E_\theta(3-\theta)^2 = (3-\theta)^2$

If $2 < \theta < 4$ then $R(\theta, \hat{\theta}_2) < R(\theta, \hat{\theta}_1)$. Neither estimator uniformly dominates the other.



Example: Let $X_1, \cdots, X_n \sim Bernoulli(p)$. Consider squared error loss and let $\hat{p}_1 = \bar{X}$. Since this has 0 bias, we have that

$$R(p, \hat{p}_1) = V(\bar{X}) = p(1-p)/n$$

Another estimator is $\hat{p}_2 = Y+\alpha / \alpha+\beta+n$, where $Y = \sum_{i=1}^{n} X_i$ and α and β are positive constants. This is the posterior mean using a Beta(α, β) prior. Now,

$$\begin{aligned} R(p, \hat{p}_2) &= V_p(\hat{p}_2) + (bias_p(\hat{p}_2))^2 \\ &= V_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(E_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2 \end{aligned}$$

Let $\alpha = \beta = \sqrt{n/4}$. The resulting estimator is $\hat{p}_2 = Y + \sqrt{n/4} / n + \sqrt{n}$ and the risk funtion is $R(p, \hat{p}_2) = n / 4(n+\sqrt{n})^2$

Neither estimator uniformly dominates the other.

Definition: A decision rule that minimizes the Bayes risk is called a Bayes rule. Formally, $\hat{\theta}$ is minimax if $r(f,\hat{\theta})=\inf_{\tilde{\theta}} r(f,\tilde{\theta})$, where the infimum is over all estimator $\tilde{\theta}$.

An estimator that minimizes the maximum risk is called a minimax rule. Formally, $\hat{\theta}$ is minimax if $\sup_{\theta} R(\theta,\hat{\theta})=\inf_{\tilde{\theta}} \sup_{\theta} R(\theta,\tilde{\theta})$, where the infimum is over all estimators $\tilde{\theta}$.

Bayes Estimators:

Let f be a prior. From Bayes' theorem, the posterior density is $f(\theta|x)=f(x|\theta)\cdot f(\theta)/m(x)=f(x|\theta)\cdot f(\theta)/\int f(x|\theta)f(\theta)d\theta$. , where

$m(x)=\int f(x,\theta)d\theta=\int f(x|\theta)f(\theta)d\theta$ is the marginal distribution of X.

Define the posterior risk of an estimator $\hat{\theta}(x)$ by $r(\hat{\theta}|x)=\int L(\theta,\hat{\theta}(x))f(\theta|x)d\theta$.

Theorem: The Bayes risk satisfies $r(f,\hat{\theta})=\int r(\hat{\theta}|x)m(x)dx$.

Let $\hat{\theta}(x)$ be the value of θ that minimizes $r(\hat{\theta}|x)$. Then $\hat{\theta}$ is the estimator.

Proof: we can rewrite the Bayes Risk as follows:

$$r(f,\hat{\theta})=\int R(\theta,\hat{\theta})f(\theta)d\theta=\int\left(\int L(\theta,\hat{\theta}(x))f(x|\theta)\cdot dx\right)f(\theta)\cdot d\theta.$$

$$=\iint L(\theta,\hat{\theta}(x))f(x,\theta)dx\,d\theta=\iint L(\theta,\hat{\theta}(x))f(\theta|x)\,m(x)\cdot dx\,d\theta.$$

$$=\int\left(\int L(\theta,\hat{\theta}(x))f(\theta|x)d\theta\right)m(x)\cdot dx=\int r(\hat{\theta}|x)m(x)\cdot dx.$$