Chapter 1 非参数统计

Statistical Inference 的步骤

① 建立模型 $\{P_f : f \in F\}$ 参数 f 在参数空间 F 中.

用估计量 $T(Y)$ 来估计 f, 其中 Y 是我们的观测值

② 建立 test function $\Psi(Y)$ 以对提出的 f 进行假设检验

③ 建立关于参数 f 的置信区间 (confidence sets)

在本书中, 主要考虑参数空间 F 是无穷维的情况.

1.1 Statistical Sampling Models.

1.1.1

我们定义衡量两个 probability 距离远近的 metric. 其中 base probability space 为 $(\mathcal{X}, \mathcal{A})$

① total variational metric

$$\|P-Q\|_{TV} = \sup_{A \in \mathcal{A}} |P(A)-Q(A)|$$

② bounded Lipschitz metric:

$$\beta_{(\mathcal{X},d)}(P,Q) = \sup_{f \in BL(1)} \left| \int_{\mathcal{X}} f(dP-dQ) \right|$$, 其中 $BL(1) = \{f : \sup_{x \in \mathcal{X}} |f(x)| + \sup_{x \neq y} \frac{|f(x)-f(y)|}{d(x,y)} \le 1$

③ kolmogorov distance:

$$\|F_P - F_Q\|_\infty = \sup_{x \in \mathbb{R}} |F_P(x) - F_Q(x)|$$. 其中 $F(x) = P(X \le x)$.

④.

$$\|f_P - f_Q\|_1 = \int_{\mathbb{R}} |f_P(x) - f_Q(x)| dx$$

1.1.2 Indirect observations.

事实上能观测到的都是在原始数据上添加了扰动.

例如: $Y_i = X_i + \varepsilon_i$. ε_i 为观测误差, Y_i 为观测结果.

$P_Y = P_X * P_\varepsilon$. $*$ 代表卷积

1.2. Guassian Models

1.2.2 Some nonparametric Gaussian Models

The Gaussian White Noise Model

(1.5) $$dY(t) = dY_f^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t), \quad t \in [0,1]. \quad f \in L^2[0,1].$$

其中 σ 为 dispersion parameter

dW: standard Gaussian white noise process.

上面式子可说成 We observe the function f in Gaussian model, at the noise level $\frac{\sigma}{\sqrt{n}}$.

在这里,我们本该将 dW 视为 standard Brownian motion $\{W(t): t \in [0,1]\}$ 的弱导数

但是事实上可以这么考虑

(1.6) $$g \longrightarrow \int_0^1 g(t)\, dY^{(n)}(t) \equiv Y_f^{(n)}(g) \sim N\left(\langle f, g\rangle, \frac{\|g\|_2^2}{n}\right), \quad g \in L^2([0,1])$$

(1.7) $$g \longmapsto \int_0^1 g(t)\, dW(t) \equiv W(g) \sim N(0, \|g\|_2^2), \quad g \in L^2([0,1])$$

若在 (1.7) 中,函数 g 取成 L^2 中的有限个标准正交基 (e_k)

$$e_k \longrightarrow \int_0^1 e_k(t)\, dY(t) \equiv Y_f^{(n)}(e_k) \sim N\left(\langle f, e_k\rangle, \frac{1}{n}\right)$$

回顾 Kolmogorov consistency theorem, 知道

[For T indexed set. $s, t \in T$ 若 $\exists$ Gaussian process $X(t)$ 使得

$E(X(t)) = f(t)$, $\quad E[(X(t)-f(t))(X(s)-f(s))] = \Phi(s,t)$, 则 Φ is

the covariance ~~and~~ of this process and with f its expectation.]

那么现在: $T = L^2[0,1]$ $\quad Y(g) \sim N\left(\langle f, g\rangle, \frac{\|g\|_2^2}{n}\right)$, $\quad W(g) \sim N(0, \|g\|^2)$

那么 $$E(Y(t)-f(t))(X(s)-f(s)) = E(Y(t)-f(t))E(X(s)-f(s)) = 0$$

那我们该如何理解 model (1.5) 呢？

$dY(t) = dY_f^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t), \quad t \in [0,1]$

也就是说，给个 g，我们就有一个 $N(\langle f, g \rangle, \frac{\|g\|^2}{n})$ 的正态分布. However, the natural sample space now is hard to work with.

- Gaussian Sequence Space Model

回顾我们的 Gaussian process $\{Y_f^{(n)}(g) : g \in L^2\}$ just means, we observe $Y_f^{(n)}(g)$ for all $g \in L^2$ simultanuous. Now take $\{e_k : k \in \mathbb{Z}\}$ is orthonormal basis of L^2. Y

(1.8) $Y_k = Y_{f,k}^{(n)} = \langle f, e_k \rangle + \frac{\sigma}{\sqrt{n}} g_k, \quad k \in \mathbb{Z}, \; n \in \mathbb{N}$. 被称为 GSSM.

g_k: i.i.d r.v of law $W(e_k) \sim N(0, \|e_k\|_2^2) = N(0,1)$

(1.5) and (1.8) are observationally equivalent to each other.

(1.8) 的 special form: $Y_k = \theta_k + \frac{\sigma}{\sqrt{n}} g_k, \; k = 1, 2, \dots n$.

(1.8) 的进一步说明：

$\{e_k : k \in \mathbb{Z}\}$ 诱 a sequence space isometry from L^2 to the sequence space l_2 of all square-summable infinite sequence through the mapping $f \mapsto \langle f, e_k \rangle$. the law $\{Y_{f,k}^{(n)} : k \in \mathbb{Z}\}$ completely characterise the finite-dimensional distibutions, and thus the law, of the process $Y_f^{(n)}$.

Q：为何说 有限维分布 determines the law of process $Y_f^{(n)}$？

1.2.3 Equivalence of Statistical Experiences

The Le Cam Distance of Statistical Experiences.

记统计实验为$\mathcal{E}^{(i)}$，其中$\mathcal{E}^{(i)}$由样本空间$\mathcal{Y}_i$与$\mathcal{Y}_i$上的测度$P_f^{(i)}$构成.

定义损失函数（用于观测 the performance of a decision procedure $T^{(i)}(Y^{(i)}) \in \mathcal{J}$ based on observations $Y^{(i)}$）

$$L: \mathcal{F} \times \mathcal{J} \longrightarrow [0,\infty)$$ 其中$\mathcal{J}$代表由决策规则所组成的集合.

$$f, T^{(i)}(Y^{(i)}) \longmapsto L(f, T^{(i)}Y^{(i)})$$

举例来说，如若 $T^{(i)}Y^{(i)}$ 表示对于参数f本身的估计量，此时$\mathcal{F}=\mathcal{J}$. $L(f,T)=d(f,T)$，d为参数空间$\mathcal{F}$上定义的某种距离.

在概率$P_f^{(i)}$下，记录此损失函数的数学期望为 $R^{(i)}(f,T^{(i)},L) = E_{P_f^{(i)}}[L(f,T^{(i)}(Y^{(i)})]$

"Le Cam distance" (between 2 experiments)

$$\Delta_{\mathcal{F}}(\mathcal{E}^{(1)},\mathcal{E}^{(2)}) = \max_f \Big[\sup_{T^{(2)}} \inf_{T^{(1)}} \sup_{f,L:|L|=1} |R^{(1)}(f,T^{(1)},L) - R^{(2)}(f,T^{(2)},L)|,$$
$$\sup_{T^{(1)}} \inf_{T^{(2)}} \sup_{f,L,|L|=1} |R^{(1)}(f,T^{(1)},L) - R^{(2)}(f,T^{(2)},L)|\Big].$$

其中 $|L| = \sup\{L(f,T): f\in\mathcal{F}, T\in\mathcal{J}\}$

如若 $\Delta_{\mathcal{F}}(\mathcal{E}^{(1)},\mathcal{E}^{(2)})=0$，说明任意决策规则 $T^{(2)}$ in experiment $\mathcal{E}^{(1)}$ 都可被转换为实验$\mathcal{E}^{(2)}$下的决策规则 $T^{(1)}$.

- propositions

(1) $\mathcal{Y}^{(1)}=\mathcal{Y}^{(2)}=\mathcal{Y}$（为一完备可分度量空间时），当 $P_f^{(1)}, P_f^{(2)}$ 有共同的 dominating measure μ on $\mathcal{Y}$ 时，有 $\Delta_{\mathcal{F}}(\mathcal{E}^{(1)},\mathcal{E}^{(2)}) \le \sup_{f\in\mathcal{F}} \int_{\mathcal{Y}} \left|\frac{dP_f^{(1)}}{d\mu} - \frac{dP_f^{(2)}}{d\mu}\right| d\mu \equiv \|P^{(1)}-P^{(2)}\|_{1,\mu,\mathcal{F}}$

这里是由于：

$$\Delta_F(\mathcal{E}^{(1)},\mathcal{E}^{(2)}) = \max_{\text{其中}i=1,2}\Big[\sup_{T^{(2)}}\inf_{T^{(1)}}\sup_{f,|L|=1} |R^{(1)}(f,T^{(1)},L) - R^{(2)}(f;T^{(2)},L)|,$$

$$\sup_{T^{(1)}}\inf_{T^{(2)}}\sup_{f,|L|=1} |R^{(1)}(f,T^{(1)},L) - R^{(2)}(f,T^{(2)},L)|\Big]$$

而 $\sup_{T^{(2)}}\inf_{T^{(1)}}\sup_{f,|L|=1} |R^{(1)}(f,T^{(1)},L) - R^{(2)}(f,T^{(2)},L)|$

$$\leq \sup_{T^{(2)}}\sup_{f,|L|=1} |R^{(1)}(f,T^{(2)},L) - R^{(2)}(f,T^{(2)},L)|$$

$$= \sup_{T^{(2)}}\sup_{f,|L|=1} \left|\int_{\mathcal{Y}} L(f,T^{(2)}(Y^{(2)}))\,dP_f^{(1)}(Y) - \int_{\mathcal{Y}} L(f,T^{(2)}(Y^{(2)}))\,dP_f^{(2)}(Y)\right|$$

$$\leq \int_{\mathcal{Y}} |L(f,T)|\,|dP_f^{(1)}(Y) - dP_f^{(2)}(Y)| \leq \|P^{(1)} - P^{(2)}\|_{\text{u.w.}F}.$$

以上我们须要求 $\mathcal{Y}^{(1)}=\mathcal{Y}^{(2)}=\mathcal{Y}$. 但 $\mathcal{Y}^{(1)}\neq\mathcal{Y}^{(2)}$ 时怎么办呢？

- two experiments are defined on different sample space.

  构建同构 from $\mathcal{Y}^{(1)}$ to $\mathcal{Y}^{(2)}$, independent of f.

  使 $P_f^{(1)} = P_f^{(2)}\circ B$, $\forall f\in F$

  因此, Given $Y^{(2)}\in\mathcal{Y}^{(2)}$, $T^{(2)}(Y^{(2)}) = T^{(1)}(B^{-1}(Y^{(2)}))$ in $\mathcal{E}^{(2)}$

  此时 $\Delta_F(\mathcal{E}^{(1)},\mathcal{E}^{(2)}) = \Delta_F(\mathcal{E}^{(1)},B^{-1}(\mathcal{E}^{(2)}))=0$

- In absense of such a bijection:

  运用充分统计量:

  $\mathcal{Y}^{(i)}$ 为样本空间. $\mathcal{E}^{(1)}$ 为 experiment give rise to observations $Y^{(1)}$ in law $P_f^{(1)}$ on $\mathcal{Y}^{(1)}$ 设 $S:\mathcal{Y}^{(1)}\longrightarrow\mathcal{Y}^{(2)}$ 使

  $Y^{(2)} = S(Y^{(1)})$, $Y^{(2)}\sim P_f^{(2)}$ on $\mathcal{Y}^{(2)}$

  且 $S(Y^{(1)})$ 为 $Y^{(1)}$ 的充分统计量.

  则 $\Delta_F(\mathcal{E}^{(1)},\mathcal{E}^{(2)})=0$.

证明: $\Delta_F(\mathcal{E}^{(1)},\mathcal{E}^{(2)}) = \max\Big[\sup_{T^{(2)}}\inf_{T^{(1)}}\sup_{f,L,|L|=1}|R^{(1)}(f,T^{(1)},L)-R^{(2)}(f,T^{(2)},L)|,$

$$\sup_{T^{(1)}}\inf_{T^{(2)}}\sup_{f,L:|L|=1}|R^{(1)}(f,T^{(1)},L)-R^{(2)}(f,T^{(2)},L)|\Big]$$

为此我们要先回顾一下充分统计量之定义:

直观上:给定统计量的值 $T=t$,它所对应的分布 $F_\theta(X|T=t)$ 是一个与参数 θ 无关的分布.也就是说,只要T值确定,不论参数 θ 选择什么,样本的概率分布都是一成不变的.

DEF (Sufficient statistics)

A statistic $t=T(X)$ is sufficient for underlying parameter θ .if the conditional prob distribution of data X, given the statistic $t=T(X)$, doesn't depend on the parameter θ.

Characteristic: ① The data processing inequality: $I(\theta;T(X))=I(\theta;X)$

② Fisher-Neyman. factorization theorem:

$f_\theta(x)$: prob density function , T is sufficient for $\theta \Leftrightarrow \exists\ g,h\geq 0$,

$$f_\theta(x) = h(x)\,g_\theta(T(x))$$

现在,

$$\sup_{T^{(1)}}\inf_{T^{(2)}}\sup_{f,L,|L|=1}|R^{(1)}(f,T^{(1)},L)-R^{(2)}(f,T^{(2)},L)|$$

$$\sup_{T^{(1)}}\sup_{f,L,|L|=1}|R^{(1)}(f,T^{(1)},L)-R^{(2)}(f,SY^{(1)},L)|$$

$$= \left|\int_{y^{(1)}} L(f;T^{(1)}(Y^{(1)}))\,dP_f^{(1)}(Y)-\int_{y^{(2)}}L(f;T^{(2)}(Y^{(2)}))\,dP_f^{(2)}(Y)\right|$$

$$\leq \left|\int_{y^{(1)}} L(f,T^{(1)}(Y^{(1)}))\,dP_f^{(1)}(Y) - \int_{y^{(2)}}L(f,T^{(1)}(S(Y^{(1)}))g(S(Y^{(1)},f)\,h(Y^{(1)})\,dy^{(2)}\right.$$

$$\leq \left|\int_{y^{(1)}} L(f,T^{(1)}(Y^{(1)})dP_f^{(1)} - L(f,\hat{T}^{(1)}(Y^{(1)}))\,g(S(Y^{(1)},f)\,h(Y^{(1)})\,S(Y^{(1)})\,dy^{(1)}\right.$$

$$P_f^{(1)}(Y) = f(Y^{(1)})\,dy^{(1)}$$

$=0$ (暂未想到证明)

Asymptotic Equivalence for Nonparametric Gaussian Regression Models

这里首先定义 $\mathcal{F}(a,M)=\{f:[0,1]\to\mathbb{R},\ \sup_{x\in[0,1]}|f(x)|+\sup\frac{|f(x)-f(y)|}{|x-y|^a}\le M\}$,

$0<a\le 1,\ 0<M<\infty$

Theorem. (三个模型之等价性)

$\mathcal{E}_n^{(1)}$: $\quad Y_i=f(x_i)+\varepsilon_i,\quad x_i=\frac{i}{n},\quad \varepsilon_i\overset{\text{i.i.d}}{\sim}N(0,\sigma^2)$

$\mathcal{E}_n^{(2)}$: $\quad dY(t)=f(t)dt+\frac{\sigma}{\sqrt{n}}dW_t\quad, t\in[0,1], n\in\mathbb{N}$

$\mathcal{E}_n^{(3)}$: $\quad Y_k=\langle f,e_k\rangle+\frac{\sigma}{\sqrt{n}}g_k,\ g_k\sim N(0,\|e_k\|_2^2)=N(0,1)$

(1) for $\mathcal{F}$ any family of bdd functions on $[0,1]$,

$\Delta_{\mathcal{F}}(\mathcal{E}_n^{(2)},\mathcal{E}_n^{(3)})=0,\quad \Delta_{\mathcal{F}}(\mathcal{E}_n^{(1)},\mathcal{E}_n^{(2)})\le\sqrt{\frac{n\sigma^2}{2}}\sup_{f\in\mathcal{F}}\|f-\Pi_n(f)\|_2$

(2). $\mathcal{F}=\mathcal{F}(a,M)$，其中 $a>\frac{1}{2}$，$M>0$. 则以上三个实验是在 Le Cam 距离的意义下渐近等价的.