- Probability

1. probability mass function (PMF) $P(X=x)=f(x)\le 1$
   cumulative distribution function (CDF) $P(X\le x)=f(x)$
   probability density function (PDF) $P(a\le x\le b)=\int_a^b f(x)dx$ (f(x)可大于1)

2. $E(\sum X_i)=\sum E(X_i)$
   $Var(\sum x_i)=\sum Var(X_i)$
   Proof: $Y_i = X_i - E(X_i)$ 常数
   $E(Y_i) = E(X_i) - E(X_i) = 0$
   $Var(\sum x_i) = E\left[\left[\sum(x_i - EX_i)\right]^2\right]$
   $= E\left[(\sum Y_i)^2\right] = \sum_{i\ne j} E(Y_iY_j) + E\left[(Y_i^2)\right]$
   $= \sum E[Y_i]^2 = \sum Var(X_i)$

3. Bernoulli Distribution $E(X)=P$ $Var(X)=P(1-P)$
   $\uparrow f(x)=p^x(1-p)^{1-x}$
   Geometric Distribution $f(k)=(1-p)^{k-1}p$ $E(X)=\frac{1}{p}$ $Var(X)=\frac{1}{p^2}-\frac{1}{p}$
   Bionomial Distribution $f(k)=\binom{n}{k}p^k(1-p)^{n-k}$ $E(X)=nP$
   $Var(X)=nP(1-P)$

$\frac{P(X=x)}{P(X=Y)}=\frac{f(y)}{f(x)}$ 密度↑, 取到某值P↓

Monte carlo [Uniform sampling & Law of large numbers]

approximate $\int_a^b h(x)dx$ $f(x)=\frac{1}{b-a}$ $x\in[a,b]$

$E[(b-a)h(X)]=\int_a^b(b-a)h(x)f(x)dx=\int_a^b h(x)dx$

- X, Z 分布相同, 则 $E(g(X))=E(g(Z))$

离散性 点概率为0 (但可能发生)

$X \sim Uniform(a,b)$
$E(X)=\frac{a+b}{2}$ $Var(X)=\frac{(b-a)^2}{12}$

4. normal random varibles
   $f(x,u,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-u)^2}{2\sigma^2}}$

中心极限定理 CLT
$\overline{X}\sim Normal(E[X], \sqrt{\frac{Var(X)}{n}})$

$\sqrt{n}(\bar{x}_n-u)\xrightarrow{d} N(0,\sigma)^2$ (converges)

5. correlation 相关性
$\rho(X,Y)=\frac{cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}\in[-1,1]$

$cov(X,Y)=E[(X-E(X))(Y-E(Y))]$

- Statistic

1. likehood function 似然函数

$\mathcal{L}(\theta|x) = P(x|\theta)$

$P(x|\theta)$ ↓ 在已知θ情况下，发生x的可能性大小

$\mathcal{L}$ ↓ 结果为x，分布函数的参数为θ的可能性大小

2. 最大似然函数 Maximum Likehood Estimation (MLE)

$\mathcal{L}(\theta, X) = P(X_1, X_2 \cdots X_n|\theta) = \prod_{i=1}^{n} P(X_i|\theta)$ [独立分布]

$\Rightarrow \log(\mathcal{L}) = \sum_{i=1}^{n} \log(P_i)$ $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta, X)$ θ̂是所有可能θ中的最大值

↓最大参数 (argmax)　↑参数空间 (Θ)

[e.g.] Uniform (a,b). find MLE of (a,b)

$\log(\mathcal{L}) = \log\left[\left(\frac{1}{b-a}\right)^n\right] = -n\log(b-a)$ $\hat{a} = \max\{X_i\}$, $\hat{b} = \min\{X_i\}$

3. 最小二乘估计 Least Squares Estimation LSE

(1) $\sum_i (\text{Mean} - X_i)^2$

$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$

[e.g.] $f(X;\theta) = \frac{1}{\theta} I\{0 \le x \le \theta\}$

MLE:

$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I\{0 \le x \le \theta\} = \begin{cases} \frac{1}{\theta^n}, & \theta \ge \max\{X_i\} \\ 0, & \theta < \max\{X_i\} \end{cases}$

$\hat{\theta} = \max\{X_i\}$

LSE: $\sum_{i=1}^{n} \left(\frac{\theta}{2} - X_i\right)^2 = \sum_{i=1}^{n} \left(\frac{\theta^2}{4} + X_i^2 - \theta X_i\right)$

$\frac{\partial}{\partial\theta} = -\sum_{i=1}^{n} X_i + n\frac{\theta}{2} = 0 \Rightarrow \hat{\theta} = \frac{2}{n}\sum_{i=1}^{n} X_i = 2\bar{X}$

(2) $\min_{\alpha,\beta} S(\alpha,\beta) = \sum_i (Y_i - \alpha - \beta X_i)^2$ ← X & Y is linear

$\frac{\partial S}{\partial \alpha} = \sum_{i=1}^{n} 2\,(Y_i - \alpha - \beta X_i)(-1) = 0$

$\Rightarrow \sum_{i=1}^{n} (Y_i - \beta X_i - \alpha) = 0$ $\Rightarrow \alpha = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta X_i) = \bar{Y} - \beta\bar{X}$

$\frac{\partial S}{\partial \beta} = \sum_{i=1}^{n} 2(Y_i - \alpha - \beta X_i)\cdot(-X_i) = 0$

$\Rightarrow \sum_{i=1}^{n} (Y_i - \beta X_i - \alpha) X_i = 0$

$$\Rightarrow \sum_{i=1}^{n} x_i Y_i - \beta \sum_{i=1}^{n} x_i^2 - \alpha \sum_{i=1}^{n} x_i$$

$$= \sum_{i=1}^{n} x_i Y_i - \beta \sum_{i=1}^{n} x_i^2 - (\bar{Y} - \beta \bar{X} \cdot N\bar{X}) = 0$$

$$\Rightarrow \beta \left(\sum_{i=1}^{n} x_i^2 - N\bar{X}^2 \right) = \sum_{i=1}^{n} x_i Y_i - N\bar{X}\bar{Y}$$

$$\beta \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right] = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})$$

$$\Rightarrow \beta = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(\sum x_i)^2 - n\bar{x}^2}$$

- $Y - \alpha - \beta X \sim N(0, \sigma^2)$

$$L(\alpha, \beta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2} \frac{\sum (Y_i - \beta x_i - \alpha)^2}{\sigma^2} \right] \Rightarrow \text{Minimizes } \underbrace{\sum (Y_i - \beta X_i - \alpha)^2}_{\Uparrow \text{ LSE}}$$

(3) residual 残差 $e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$

examine
- e_i independ on X_i ?
- variance of e_i independ on x_i ?

- KL - Divergence (KL散度 用分布Q来估计数据的真实分布P的编码损失)

$$D_{KL}(P \| q) = \int P(x) \log \left(\frac{P(x)}{q(x)} \right) dx$$

$$\theta_{\min KL} = D_{KL}(P(x) \| P(x|\theta)) = E_{x \sim P(x)} \log \frac{P(x)}{P(x|\theta)}$$

$$\Rightarrow \theta_{\min KL} = \arg\min_{\theta} E_{x \sim P(x)} [-\log P(x|\theta)] = \arg\max_{\theta} E_{x \sim P(x)} \log P(x|\theta) = \hat{\theta}$$

[e.g.] ① Binomal

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$l(p:k) = \log \binom{n}{k} + k \log(p) + (n-k) \log(1-p)$$

$$\hat{p} = \frac{k}{n}$$

② Normal

$$l(u, \sigma) = \sum -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i - u)^2}{2\sigma^2}$$

$$\frac{\partial l}{\partial u} = \sum \frac{x_i - u}{\sigma^2} = 0 \Rightarrow u = \frac{1}{n} \sum x_i$$

$$\frac{\partial l}{\partial \sigma} = \sum \left[-\frac{1}{2\sigma^2} + \frac{(x_i - u)^2}{2(\sigma^2)^2} \right] = 0$$

$$\sigma^2 = \frac{1}{n} \sum (x_i - u)^2$$

(4) confidence interval (CI) 置信区间

$$\bar{x}-\frac{b\sigma}{\sqrt{n}} \le u \le \bar{x}+\frac{a\sigma}{\sqrt{n}} \Rightarrow \frac{\sqrt{n}(\bar{x}-u)}{\sigma} \sim N(0,1)$$

$$\Rightarrow P\left(\bar{x}-\frac{b\sigma}{\sqrt{n}} \le u \le \bar{x}+\frac{a\sigma}{\sqrt{n}}\right) = P\left(-a \le \frac{\sqrt{n}(\bar{x}-u)}{\sigma} \le b\right) = \phi(b)-\phi(-a)$$

CDF

$$= \underbrace{\phi(b)+\phi(a)-1}_{\text{置信水平}}$$

· Bernoulli $\left[\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$ $\left(\left[\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]\right)$

$$Var(\hat{p}) = Var\left(\frac{Y}{n}\right) = \frac{1}{n^2}Var(Y) = \frac{1}{n^2}\cdot np(1-p) = \frac{p(1-p)}{n}$$

[e.g.] find a $1-\alpha$ CI. T?

w.p. (with probability) α, it is not in T

$\Rightarrow \forall a, b \quad \phi(b)+\phi(a)-1 = 1-\alpha$

let $a=b$. $\quad 2\phi(a)-1 = 1-\alpha$

$\alpha\downarrow$, length of CI $\uparrow$

$n\uparrow$, $\downarrow$

σ^2, $\downarrow$

· optimization

1. feasible set Ω (可行集)
2. $B(x,\varepsilon) = \{y : \|y-x\| \le \varepsilon\}$ $\quad \forall y \in \Omega \cap B(x,\varepsilon),\ f(x) \le f(y)$
   - for any $\varepsilon > 0$, x: global minimizer
   - for some $\varepsilon > 0$, x: local minimizer

3. convex set $x = \theta x_1 + (1-\theta)x_2$ $\quad \theta \in [0,1]$

   convex function $\forall\, 0 \le \lambda \le 1 \quad f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$

· $f(x)$ is a convex function and Ω is a convex set. $\quad \min f(x)$ s.t. $x \in \Omega$

Then any local minimum is also a global minimum.

proof: let $\bar{x}$ be a local minimum

$\Rightarrow \bar{x} \in \Omega$. $\exists \epsilon > 0$ s.t. $f(\bar{x}) \le f(x)$ for any $x \in \Omega \cap \{x : \|x-\bar{x}\| \le \epsilon\}$ ①

assume $\exists z \in \Omega$. $f(z) < f(\bar{x})$

we have $\lambda\bar{x}+(1-\lambda)z \in \Omega$ $\forall \lambda \in [0,1]$, then $f(\lambda\bar{x}+(1-\lambda)z) \le \lambda f(\bar{x}) + (1-\lambda) f(z)$

$< \lambda f(\bar{x}) + (1-\lambda) f(\bar{x}) = f(\bar{x})$ $\quad$ as $\lambda \to 1$ $\lambda\bar{x}+(1-\lambda)z \to \bar{x}$, which contradicts with ①

· Prove $H: \{(x,y): y \le ax+b\}$ is a convex set

$(x,y) = \theta(x_1,y_1) + (1-\theta)(x_2,y_2)$

$\begin{cases} x = \theta x_1 + (1-\theta)x_2 \\ y = \theta y_1 + (1-\theta) y_2 \end{cases}$ $\theta \in [0,1]$

$\begin{matrix} y_1 \le ax_1 + b \\ y_2 \le ax_2+b \end{matrix} \Rightarrow \theta y_1 + (1-\theta) y_2 \le \theta(ax_1+b) + (1-\theta)(ax_2+b)$

$= a(\theta x_1 + (1-\theta)x_2) + b$

$\Rightarrow y \le ax+b$

· S_1, S_2 are convex sets $\Rightarrow$ $S_1 \cap S_2$ is a convex set

Proof:

let $x_1, x_2 \in S_1 \cap S_2$

$x = \theta x_1 + (1-\theta)x_2$ $(\theta \in [0,1])$

S_1, S_2 are convex sets

$\Rightarrow x \in S_1$, $x \in S_2$

$\Rightarrow x \in S_1 \cap S_2$

4. second order condition (SOC) [f is convex]

f is convex iff dom(f) is a $\begin{cases} \text{convex} \\ \text{concave} \end{cases}$ set and $\begin{cases} \boxed{f''(x) \ge 0} \\ f''(x) \le 0 \end{cases}$ for all $x \in$ dom(f)

☆ let $g(\theta) = f(y+\theta e)$ [y 起始点. e 方向向量. θ 步长]

choose any y / e / θ. prove $g''(\theta=0) \ge 0$ for any y and e.

[e.g.] prove $f(x) = \sum_i a_i (x_i - c_i)^2$ is a convex function in x with $a_i > 0$

$g(\theta) = f(y+\theta e) = \sum_i a_i (y_i + \theta e_i - c_i)^2$

$g''(\theta) = \sum_i a_i e_i^2 \ge 0$ $\Rightarrow$ f is a convex function $(a_i > 0)$

$\left[\begin{matrix} \text{concave up} & \text{convex} \\ \text{concave down} & \text{concave} \end{matrix} \right.$

5. the epigraph of a function 函数上图

$\text{epi } C = \{(x, y) : y \geq f(x),\ x \in \Omega\}$

① $C = \{(x, y) : y \geq f(x),\ x \in \Omega\}$ is a convex set if $f(x)$ is a convex function.

new point $\lambda(x_1, y_1) + (1-\lambda)(x_2, y_2) = (\lambda x_1 + (1-\lambda)x_2,\ \lambda y_1 + (1-\lambda)y_2)$

$\Rightarrow \lambda y_1 + (1-\lambda)y_2 \geq \lambda f(x_1) + (1-\lambda)f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)$

$\Rightarrow$ C is a convex set

② $f(x)$ is a convex function if $C = \{(x, y) : y \geq f(x),\ x \in \Omega\}$ is a convex set

$x_1 \in \Omega,\ x_2 \in \Omega$ then $(x_1, f(x_1)) \in C$ $(x_2, f(x_2)) \in C$

凸集中任意两点连线仍属于凸集 $(\lambda x_1 + (1-\lambda)x_2,\ \lambda f(x_1) + (1-\lambda)f(x_2)) \in C$

$\Rightarrow \lambda f(x_1) + (1-\lambda)f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)$

$\Rightarrow f$ is a convex function

6. Suppose f is differentiable and convex. If x^* is feasible such that $\nabla f(x^*) = 0$, then x^* is the global minimizer.

7. convex function can be discontinous at the boundary [但需高于单侧极限]

8. $f(x) = h(g(x))$

- h convex & non-decreasing, g convex $\Rightarrow f$ convex
- h convex & non-increasing, g concave $\Rightarrow f$ convex
- h concave & non-decreasing, g concave $\Rightarrow f$ concave
- h concave & non-increasing, g convex $\Rightarrow$ concave

9. exercises.

(1.) ① S is a convex set

<1> choose $x, y \longrightarrow$ $\lambda x + (1-\lambda) y$ is in S?

<2> S_1, S_2 convex $\rightarrow$ $S_1 \cap S_2$ convex

<3> $f(x)$ is a convex function $\begin{cases} \{x : f(x) \le y,\ x \in \Omega\} \text{ for any } y \\ \{(x,y) : y \ge f(x),\ x \in \Omega\} \end{cases}$ are convex set

② f is a convex function

<1> by definition $\begin{cases} f(\lambda x_1 + (1-\lambda) x_2) \le \lambda f(x_1) + (1-\lambda) x_2 \\ \text{domain } \Omega \text{ is a convex set } \& \ \dfrac{d^2 f(x+\theta e)}{d\theta^2} \ge 0 \quad \forall\ x+\theta e \in \Omega \end{cases}$

(2) composition

[e.g.] $g(x) = f(Ax+b) - c$ is convex if f is convex

$$\begin{aligned} g(\lambda x + (1-\lambda) y) &= f[A(\lambda x + (1-\lambda) y) + b] - c \\ &= f[\lambda (Ax+b) + (1-\lambda)(Ay+b)] - c \\ &\le \lambda f(Ax+b) + (1-\lambda) f(Ay+b) - c \\ &= \lambda (f(Ax+b) - c) + (1-\lambda)(f(Ay+b) - c) \\ &= \lambda g(x) + (1-\lambda) g(y) \end{aligned}$$

(3) elementwise maximum

$f_1, f_2 \cdots f_n$ convex $\Rightarrow$ $f(x) = \max\{f_1(x), f_2(x) \cdots f_n(x)\}$ convex

proof:

pick any $x, y \in \text{dom}(f)$, $\lambda \in [0,1]$. then

$$\begin{aligned} f(\lambda x + (1-\lambda) y) &= f_j(\lambda x + (1-\lambda) y) \quad (\text{for some } j \in \{1, 2, \cdots n\}) \\ &\le \lambda f_j(x) + (1-\lambda) f_j(y) \\ &\le \lambda \max\{f_1(x), \cdots f_n(x)\} + (1-\lambda) \max\{f_1(y), \cdots f_n(y)\} \\ &= \lambda f(x) + (1-\lambda) f(y) \end{aligned}$$

(4) nonnegative weighted sums

$f_1, f_2 \cdots f_n$ convex, $w_i \ge 0 \Rightarrow f = \sum_{i=1}^{n} w_i f_i$ convex

$\forall\ x, y$, $0 \le \lambda \le 1$ $\quad f_m(\lambda x + (1-\lambda) y) \le \lambda f_m(x) + (1-\lambda) f_m(y)$

$\sum_{i=1}^{n} f_i(\lambda x + (1-\lambda) y) w_i = f \le \sum_{i=1}^{n} w_i f_i$ $\quad [\ g(x) = \int_A w(y) f_y(x)\, dy \ \text{ convex}\ ]$

(5) Show $S=\{(x,y,z): z\geq x^2+y^2\}\subset\mathbb{R}^3$ is convex 【同知识点 5】

$a_1=(\lambda x_1, \lambda y_1, \lambda z_1)$, $a_2=((1-\lambda)x_2, (1-\lambda)y_2, (1-\lambda)z_2)$

$\Rightarrow$ show $[\lambda x_1+(1-\lambda)x_2]^2+[\lambda y_1+(1-\lambda)y_2]^2\leq \lambda z_1+(1-\lambda)z_2$ ①

$[\lambda x_1+(1-\lambda)x_2]^2\leq \lambda x_1^2+(1-\lambda)x_2^2$ (x^2 is convex)

① $\Rightarrow \lambda x_1^2+(1-\lambda)x_2^2+\lambda y_1^2+(1-\lambda)y_2^2$

$=\lambda(x_1^2+y_1^2)+(1-\lambda)(x_2^2+y_2^2)\leq\lambda z_1+(1-\lambda)z_2$

(6) Show that if S_1 and S_2 are convex sets in $\mathbb{R}^{m\times n}$, then so is their partial sum

$S=\{(x, y_1+y_2)\mid x\in\mathbb{R}^m, y_1, y_2\in\mathbb{R}^n, (x,y_1)\in S_1, (x,y_2)\in S_2\}$

Proof:

consider $(\bar{x}, \bar{y}_1+\bar{y}_2), (\tilde{x}, \tilde{y}_1+\tilde{y}_2)\in S$.

for $0\leq\theta\leq1$

$\theta(\bar{x}, \bar{y}_1+\bar{y}_2)+(1-\theta)(\tilde{x}, \tilde{y}_1+\tilde{y}_2)$

$=(\theta\bar{x}+(1-\theta)\tilde{x}, (\theta\bar{y}_1+(1-\theta)\tilde{y}_1)+(\theta\bar{y}_2+(1-\theta)\tilde{y}_2))$

Since S_1, S_2 is convex. $(\bar{x},\bar{y}_1)\in S_1$, $(\bar{x},\bar{y}_2)\in S_2$, $(\tilde{x},\tilde{y}_1)\in S_1$ $(\tilde{x},\tilde{y}_2)\in S_2$

$\Rightarrow(\theta\bar{x}+(1-\theta)\tilde{x}, \theta\bar{y}_1+(1-\theta)\tilde{y}_1)\in S_1$, $(\theta\bar{x}+(1-\theta)\tilde{x}, \theta\bar{y}_2+(1-\theta)\tilde{y}_2)\in S_2$

$\Rightarrow(\theta\bar{x}+(1-\theta)\tilde{x}, (\theta\bar{y}_1+(1-\theta)\tilde{y}_1)+(\theta\bar{y}_2+(1-\theta)\tilde{y}_2))\in S$

(7) C is a convex set $x_1, x_2, x_3\in C$, $\theta_1+\theta_2+\theta_3=1$ ($\theta_i>0$) show $\theta_1x_1+\theta_2x_2+\theta_3x_3\in C$

Proof:

$\theta_1x_1+(1-\theta_1)y\Rightarrow\theta_1x_1+(\theta_2+\theta_3)\left(\frac{\theta_2}{\theta_2+\theta_3}x_2+\frac{\theta_3}{\theta_2+\theta_3}x_3\right)$

$\frac{\theta_2}{\theta_2+\theta_3}+\frac{\theta_3}{\theta_2+\theta_3}=1\Rightarrow y\in C\Rightarrow\theta_1x_1+(1-\theta_1)y=\theta_1x_1+\theta_2x_2+\theta_3x_3\in C$

$\Rightarrow$① $\theta_i>0$, $\sum_i\theta_i=1$ then $\sum_i\theta_ix_i\in C$

$\Rightarrow$② $f(\theta_1x_1+\theta_2x_2+\theta_3x_3)=f(\theta_1x_1+(1-\theta)y)\leq\theta_1f(x_1)+(1-\theta)f(y)$

$\leq\theta_1f(x_1)+\theta_2f(x_2)+\theta_3f(x_3)$ [f(x) is convex in x for the real space]

$\Rightarrow$③ $\theta_i>0$ $\sum_i\theta_i=1$ $f(\sum_i\theta_ix_i)\leq\sum_i\theta_if(x_i)$

$\Rightarrow$④ x be the random variable $f(E(x))\leq E(f(x))$

(8) show $f(x)=\sum_{i=1}^{r}\alpha_i x_{[i]}$ is a convex function of x. $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r \geq 0$

$x_{[i]}$ denotes the i-th largest compenent of x.

$$f(x)=\sum_{i=1}^{r}\alpha_i x_{[i]} = (\alpha_1-\alpha_2)x_{[1]} + (\alpha_2-\alpha_3)(x_{[1]}+x_{[2]}) + \cdots$$
$$+ (\alpha_{r-1}-\alpha_r)(x_{[1]}+\cdots+x_{[r-1]}) + \alpha_r(x_{[1]}+\cdots+x_{[r]})$$

Since $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_r$, $\sum_{i=1}^{k} f_{[i]}$ is convex

by nonnegative weighted sums, we can conclude $f(x)$ is convex

(9) 报童问题 [The Newsvendor's Problem]

c 成本 p 销售价 s 回收价 D 需求 [pdf 为 h(x)] Q 订购量 $\to \int_0^\infty h(x)=1$

$(p>c>s)$, $(Q-D)^+ := \max\{Q-D, 0\}$

$$\pi(Q) = p\,E[\min(Q,D)] + s\,E[(Q-D)^+] - cQ$$
$$= p\int_0^\infty \min(Q,x)h(x)dx + s\,E[(Q-D)^+] - cQ$$
$$= p\Big(\underbrace{\int_0^Q x h(x)dx}_{①} + \underbrace{\int_Q^\infty Q h(x)dx}_{②}\Big) + s\underbrace{\int_0^Q (Q-x)h(x)dx}_{③} - cQ$$

$$\frac{\partial\pi}{\partial Q} = p\Big[Qh(Q) + \int_Q^\infty h(x)dx - Qh(Q)\Big] + s\Big(1-\int_Q^\infty h(x)dx\Big) - c$$
$$= (p-s)\int_Q^\infty h(x)dx + s - c$$

$$\frac{\partial ①}{\partial Q} = Qh(Q)$$

$$\frac{\partial ②}{\partial Q} = \frac{Q\int_Q^\infty h(x)dx}{\partial Q} = \int_Q^\infty h(x)dx - Q\frac{\int_\infty^Q h(x)dx}{\partial Q}$$
$$= \int_Q^\infty h(x)dx - Q[h(Q) - \underbrace{h(\infty)}_{=0}]$$
$$= \int_Q^\infty h(x)dx - Qh(Q)$$

$$\frac{\partial ③}{\partial Q} = \frac{Q\int_0^Q h(x)dx - \int_0^Q x h(x)dx}{\partial Q}$$
$$= Qh(Q) + \int_0^Q h(x)dx - Qh(Q)$$
$$= \int_0^Q h(x)dx = 1 - \int_Q^\infty h(x)dx$$

$$\pi'(Q) = 0 \Rightarrow \int_{Q^*}^\infty h(x)dx = \frac{c-s}{p-s}$$

$$\Rightarrow \int_0^{Q^*} h(x)dx = 1 - \frac{c-s}{p-s} = \frac{p-c}{p-s}$$

$X \sim N(\mu, \sigma)$, $\int_0^{Q^*} h(x)dx = A$

转化为均值为0，方差为1的标准正态分布.

$z = \frac{Q^*-\mu}{\sigma}$, $\phi(z) = A$ (查表得 z)

$$\Rightarrow Q^* = z\sigma + \mu$$

$\pi''(Q) = -(p-s)h(Q) = (s-p)h(Q) \leq 0 \Rightarrow$ concave function

- **(16 points) The Newsvendor's Problem: MLE + Convex Optimization**
A newspaper vendor faces the task of determining the optimal quantity of newspapers to stock, given uncertain demand D. The cost structure for ordering newspapers is as follows: the first twelve copies incur a cost of \$8 per copy, while any additional copies beyond twelve incur a cost of \$5 per copy. Each sold copy yields a revenue of \$10 for the vendor, while each unsold copy has a salvage value of \$2. Remark: for ease of analysis, we allow the vendor to order a non-integer number of copies.

  (a) **(4 points)** Given that the demand D follows a uniform distribution with a support of $[\theta, 2\theta]$, where θ is unknown, the objective is to estimate the maximum likelihood estimator (MLE) of θ. Five samples of D are provided: $\{12, 15, 13, 20, 14\}$. Calculate the MLE of θ. Explain the result in detail.

  (b) **(8 points)** Assuming that D follows a uniform distribution with a support of $[\theta, 2\theta]$, where θ is the MLE obtained in (a), calculate the vendor's optimal order decision. Explain the result in detail.

  (c) **(4 points)** For (b), what's the vendor's optimal order decision, assuming that the vendor can only order an integer number of copies? Explain the result in detail.

(a) $\max\{\frac{x_i}{2}\} = 10$, $\min\{x_i\} = 12$ $\quad L(\theta) = (\frac{1}{\theta})^5 \Rightarrow \hat{\theta} = 10$

(b) order decision Q,

① $Q \le 12$ $\quad C = 8Q$ $\quad$ ② $Q \ge 12$ $\quad C = 12 \cdot 8 + 5(Q-12) = 5Q + 36$

$$W(Q) = 10\min(D,Q) + 2\max(0, Q-D) - C$$

① $Q > 12$ $\quad W(Q) = 10\min(D,Q) + 2\max\{0, Q-D\} - 5Q - 36$

$D \sim U(10, 20)$

$$E[W(Q)] = \frac{1}{10}\left[\int_{10}^{Q} (10D + 2(Q-D) - 5Q - 36)\,dD + \int_{Q}^{20} (10Q - 5Q - 36)\,dD\right]$$

$$= \frac{1}{10}\left[\int_{10}^{Q} (8D - 3Q - 36)\,dD + \int_{Q}^{20} (5Q - 36)\,dD\right]$$

$$= \frac{1}{10}(-4Q^2 + 130Q - 760)$$

$Q = 16.25 \Rightarrow E[W(Q)]$ maximum

② $Q \le 12 \Rightarrow Q \le 10$ $E[W(Q)] = 2Q$, $Q \in [10, 12]$ $\quad E((Q)) = -\frac{2}{5}Q^2 + 10Q - 40$

$E(W(Q)) \le E(W(12))$

(c) $Q = 16 \Rightarrow 29.6$ $\quad Q = 17 \Rightarrow 29.4$

$\Rightarrow Q = 16$

• Machine Learning

1. K-Nearest Neighbors (KNN . K近邻算法) non-parametric

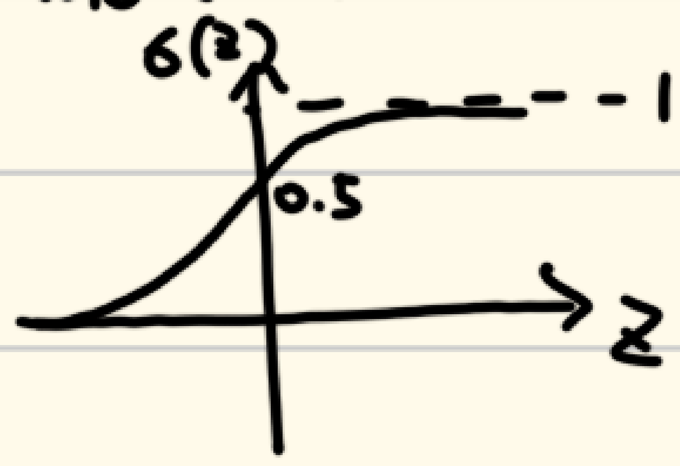
一个新点与其最近的K个点产生关联

record the entry (dis_i, $label_i$) sort by dis_i. pick the first k entries.

Choose the label with the largest frequency

2. logistic regression parametric

$P_{w,b}(C_1|x) = \sigma(z) = \frac{1}{1+e^{-z}}$ ($z = wx+b$. $P_{w,b}(C_1|x) \ge 0.5 \to$ output C_1, otherwise C_2)



• $y \in \{0,1\}$ $\max_{\theta,b} l(\theta,b) = \log \prod_{k=1}^{m} P(y^k|x^k;\theta,b) = \sum_{k=1}^{m} \log P(y^k|x^k,\theta,b)$

(x_k, y_k), $k \in \{1,2,\dots m\}$

$\log P(y^k|x^k,\theta,b) = \log\left[\sigma(\sum_i \theta_i x_i^k + b)\right]^{y^k} \cdot \left[1-\sigma(\sum_i \theta_i x_i^k + b)\right]^{1-y^k}$

$= y^k \log \sigma(\sum_i \theta_i x_i^k + b) + (1-y^k)\log\left[1-\sigma(\sum_i \theta_i x_i^k + b)\right]$ $y^k \log(1+e^{-z})^{-1} + (1-y^k)\log\frac{e^{-z}}{1+e^{-z}}$

$= (y^k-1)(\sum_i \theta_i x_i^k + b) - \boxed{\log\left[1+\exp(-\sum_i \theta_i x_i^k - b)\right]}$ $= -y^k\log(1+e^{-z}) + (1-y^k)\left[-z-\log(1+e^{-z})\right]$

$= (y^k-1)z - \log(1+e^{-z})$

for $\boxed{\quad}$. $\forall (e,c)$ as a direction

$h(t) = \log\left[1+\exp(-\sum_i \theta_i x_i^k - b + t(-\sum_i e_i x_i^k - c))\right]$

$= \log\left[1+\exp(A+tB)\right] = \log\left[1+e^{z(t)}\right]$ ($z(t) = A+tB$)

$h''(t) = \frac{B^2}{[1+e^{z(t)}]^2} e^{z(t)} \ge 0$

$\Rightarrow$ h(t) is convex in (θ,b)

$\Rightarrow$ $l(\theta,b)$ is concave in (θ,b)

☆优化算法因凹性而可稳定收敛

3. Gradient Descent 梯度下降 [$\min_x f(x)$] → $\max_x f(x)$ 需要变号

$x^{(t+1)} = x^{(t)} - \alpha \nabla f(x^{(t)})$ (α 学习率)

→ 沿增长方向相反的方向偏移

stopping criteria by ε

$|x^{(t+1)} - x^{(t)}| \le \varepsilon$ or $|f'(x^{(t)})| \le \varepsilon$

4. Stochastic gradient descent [随机梯度下降]

for logical regression

$$\begin{cases} \theta^{(t+1)} \leftarrow \theta^{(t)} + \alpha^{(t)} \frac{1}{|B|} \sum_{k\in B} \left\{ (y^k - 1) x^k + \frac{\exp(-\theta^{(t)} x^k - b^{(t)}) x^k}{1 + \exp(-\theta^{(t)} x^k - b^{(t)})} \right\} \\ b^{(t+1)} \leftarrow b^{(t)} + \alpha^{(t)} \frac{1}{|B|} \sum_{k\in B} \left\{ (y^k - 1) + \frac{\exp(-\theta^{(t)} x^k - b^{(t)})}{1 + \exp(-\theta^{(t)} x^k - b^{(t)})} \right\} \end{cases}$$

(10 points) Gradient Descent Method
Let's $\boxed{\text{maximize}}$ the function $f(x) = -x^2/2 + x$ through the gradient descent method. We fix the learning rate in the algorithm as a constant, α. Assume that the initial value of x is $x^{(0)} = 2$ and $x^{(t)}$ is the value of x after the t-th iteration of gradient descent. The stoping rule is $|x^{(t+1)} - x^{(t)}| \leq \epsilon$, where ϵ is a non-negative constant.

(a) Let $\epsilon = 1/1000$. In each of the following situations will the final output be the optimal solution? Explain the reasons.
  i. **(1 point)** $\alpha = 0$.
  ii. **(1 point)** $\alpha = 1$.
  iii. **(1 point)** $\alpha = 2$.
  iv. **(2 points)** $\alpha = 1/2$.

(b) **(5 points)** Let $\epsilon = 0$. For what range of α will $\lim_{t\to\infty} x^{(t)}$ converge to the optimal solution? Explain the reasons.

$\nabla f(x) = -x + 1$ maximize

(a) $x^{(t+1)} \leftarrow x^{(t)} + \alpha \nabla f(x^{(t)}) = x^{(t)} + \alpha(-x^{(t)} + 1)$

$x^{(t)} - 1 = (1-\alpha)(x^{(t-1)} - 1) \Rightarrow x^{(t)} = (1-\alpha)^t + 1$ the answer should be 1

① $\alpha = 0$ $x^{(t)} = 2$ No

② $\alpha = 1$ $x^{(t)} = 1$ Yes

③ $\alpha = 2$ $x^{(t)} = -1^t + 1$ No

④ $\alpha = \frac{1}{2}$ $x^{(t)} = (\frac{1}{2})^t + 1 \to 1$ Yes

(b) $(1-\alpha)^t + 1 \to 1 \Rightarrow (1-\alpha)^t \to 0$ $\alpha \in (0, 1]$

5. K-Means (unsupervised machine learning 非监督式)

① initialize k cluster centers $\{c^1, c^2, \cdots c^k\}$ randomly

② 分配每点所在簇、cluster assignment

$\pi(i) = \arg\min_{j=1\cdots k} \|x^i - c^j\|^2$

③ Center adjustment $c^j = \frac{1}{|\{i : \pi(i) = j\}|} \sum_{i:\pi(i)=j} x^i$

☆直至中心点稳定

[**12 points**]Suppose we would like to use the K-means algorithm and L2-norm distance (Euclidian distance) to cluster the 8 data points given in Figure 3 below into K = 3 clusters. The L2-norm distance between points $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ is $d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. The coordinates of the data points are:

$$x^1 = (2,8) \quad x^2 = (2,5) \quad x^3 = (1,2) \quad x^4 = (5,8)$$
$$x^5 = (7,3) \quad x^6 = (6,4) \quad x^7 = (8,4) \quad x^8 = (4,7)$$

Choose x_1, x_3, x_5 as center

① $\begin{cases} 1: x_1, x_2, x_4, x_8 & (3.25, 7) \\ 2: x_3 & (1,2) \\ 3: x_5, x_6, x_7 & (7, 3.67) \end{cases}$

$\xrightarrow{\text{iteration}}$ stop after the centers don't change

6. Cross-validation 交叉验证

K-fold (K-1) for training, 1 for validation

Model 1 is better iff.

$\frac{1}{K}\sum_i err_h(\theta_i) < \frac{1}{K}\sum_i err_g(\gamma_i)$

1 $h(\theta, x)$ 2 $g(\gamma, x)$

for supervised learning

(sol.) Cross-validation for KNN

Small summary

Algorithm	parametric	supervised	underfitting	overfitting
KNN	nonparametric	supervised	large K (oversmooth decison boundaries)	small K (noise-sensitive & complex boundaries)
K-Means	nonparametric	unsupervised	small K (fail to capture clusters)	large K (meaningless tiny clusters)
Logic Regression	parametric	supervised	poor features & overly simple model	too many features & weak regression