1. Introduction

§1 Key steps in data science

1. Programming: collect the past data (fundamental part, automating decision)
2. Probability: describle data pattern/model (quantify uncertainty)
3. Statistics: 选择最佳pattern (test pattern's credit (whether reliable))
4. Sampling: prediction for given input (calculate complicated objectives)
5. Optimization: optimize input/objectives

其他steps: Simulation: 给出pattern, calculate objective for different stratagies.
Machine learning: 当step 2 & 3过于复杂时.
- Neural network: propose black boxes
- Deep learning: optimize parameters

2. Probability

§3.1 Defintions

1. Probability: 给出model, 预测outcome
2. 一些相关词: random experiment, sample space (discrete/continuous), event

§3.2 集合运算

1. union (or): $\cup$; intersection (and): $\cap$; complement (not): A'/A^c
2. 性质: $(A\cup B)' = A'\cap B'$; $(A\cap B)' = A'\cup B'$

§3.3 Axioms of probabilities

1° $P(S)=1$ 2° $0\le P(A)\le 1$ 3° $A\cap B=\varnothing \Rightarrow P(A\cup B)=P(A)+P(B)$

推论: $P(\varnothing)=0$, $P(E')=1-P(E)$

§3.4 Events间的关系

1. joint and disjoint: joint: $P(A+B)<P(A)+P(B)$
   disjoint/mutually exclusive: $P(A\cap B)=0$, $P(A+B)=P(A)+P(B)$
2. 加法法则: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$
   *拓展: $P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$
3. independence (独立性)
   independent: $P(A\cap B)=P(A)\cdot P(B)$ / $P(A|B)=P(A)$ / $P(B|A)=P(B)$
   dependent: $P(A\cap B)\ne P(A)\cdot P(B)$ (A对B发生有影响)
4. 一些结论:
   1° 若$P(A),P(B)\ne 0$, 则A与B互斥不独立, 独立不互斥
   2° $P(A)=0 \not\rightleftarrows A=\varnothing$ (A为不可能事件) (考虑连续情况)
   3° 零概率事件与任何事件独立.
   不可能事件与任何事件既独立又互斥
   4° 不独立可能不互斥, 不互斥可能不独立

§4 Discrete random variables

1. Probability mass function (PMF): 表示确切的概率
2. Cumulative distribution function (CDF):
   $F(x)=P(X\le x)=\sum_{x_i\le x} f(x_i)$, x定义域为R
3. Mean: $E[X]=\sum x\cdot f(x)$
4. Variance: $Var[X]=\sum (x-E[X])^2\cdot f(x)=E[X^2]-(E[X])^2$
5. Useful formulas
   1° linearity: $E[\sum_i X_i]=\sum_i E[X_i]$
   2° $E[g(x)]=\sum g(x)\cdot f(x)$

§5 Continuous random variables

1. Probability density function (PDF): 不表示确切的概率,
   对任意x, f(x)可能大于1, $P(X=x)=0$. 若$x\notin$ Sample space, $f(x)=0$
2. Cumulative distribution function (CDF):
   $F(x)=\int_{-\infty}^{x} f(u)\,du$, PDF与CDF间为积分/求导关系
3. Mean: $E[X]=\int_{-\infty}^{\infty} x f(x)\,dx$
4. Variance: $Var[X]=\int_{-\infty}^{\infty}(x-E[X])^2\cdot f(x)\,dx$

§6.1 Common discrete R.V. distributions

1. Bernoulli distribution: Ber(p) (1次实验成功的概率)
   仅进行一次实验, 对于随机变量X, 有概率p取1为值, 否则取0
   Mean = p Var = p(1-p) = pq
2. Binomial distribution: Bin(n,p) (n次实验成功k次的概率)
   ① n次伯努利实验 ② n次实验相互独立(有放回)
   ③ 系数n(实验次数), p(每次成功概率) ④ 变量X: 成功/失败 次数
   $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$, $\binom{n}{k}=\frac{n!}{k!(n-k)!}$
   Mean = np Var = np(1-p) = npq
   *最可能值
   ① (n+1)p不为整数, 则 $x=[(n+1)\cdot p]$是最可能值
   ② (n+1)p为整数, 则 $x=(n+1)p$, $x=(n+1)p-1$是最可能值
3. Geometric distribution: Geo(p) (首次成功需要k次的概率)
   ① 连续进行伯努利实验, 直至成功 ② 实验间相互独立(有放回)
   ③ 含有系数p(每次成功概率) ④ 变量X: 第X次实验首次成功
   $P(X=k)=(1-p)^{k-1}\cdot p$, $k=1,2,3,\dots$
   Mean = $\frac{1}{p}$ Var = $(1-p)/p^2$
4. Poisson distribution: $Poi(\lambda)$ (某段时间内随机事件发生k次的概率)
   ① 泊松分布是二项分布在n很大p很小时的一种极限形式
   ② 用于估计在某段连续的时间内, 某件事发生的次数
   ③ 系数λ (历史上单位时间内随机事件的平均发生次数)
   ④ 变量X: 某段时间内随机事件的发生次数
   $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$, $k=0,1,2\dots$
   Mean = Var = λ

§6.2 Common continuous R.V. distributions

1. Uniform distribution: Unif[a,b]
   ① X在[a,b]上取值有着相同的概率
   ② 系数: a(区间下界), b(区间上界)
   $f(x)=\begin{cases}\frac{1}{b-a}, & a\le x\le b\\ 0, & \text{otherwise}\end{cases}$ $F(x)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\le x<b\\ 1 & b<x\end{cases}$
   Mean = $\frac{1}{2}(a+b)$ Var = $\frac{1}{12}(b-a)^2$
2. Normal distribution: $N(\mu,\sigma^2)$
   ① 系数: μ(均值), σ(标准差)
   $f(x,\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, $-\infty<x<+\infty$
   *性质: ① $y=\varphi(x)$以$x=\mu$为对称轴, $\varphi(\mu)$取最大值 $\frac{1}{\sqrt{2\pi}\sigma}$
   ② $y=\varphi(x)$以x轴为渐近线. $x=\mu\pm\sigma$为拐点
   ③ σ固定, μ变化. 左右移动
   μ固定, σ变小(大), 最高点上(下)移, 变陡(缓)
   *标准正态分布: $\mu=0$, $\sigma=1$, $\varphi_0(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$
   化为标准正态分布: $\varphi(x)=\frac{1}{\sigma}\varphi_0(\frac{x-\mu}{\sigma})$, $\Phi(x)=\Phi_0(\frac{x-\mu}{\sigma})$
3. Exponential distribution: $Exp(\lambda)$
   ① 与泊松分布相对
   ② 用于估计同一事件相邻两次发生的时间间隔
   ③ 系数: λ (历史上单位时间内随机事件的平均发生次数)
   ④ 变量X: 随机事件相邻两次发生的时间间隔
   $f(x,\lambda)=\begin{cases}\lambda e^{-\lambda x} & x\ge 0\\ 0 & x<0\end{cases}$, $F(x,\lambda)=\begin{cases}1-e^{-\lambda x} & x\ge 0\\ 0 & x<0\end{cases}$
   *无记忆性: $X\sim exp$, $s>0$, $t>0$, 则$P\{X>s+t\mid X>s\}=P(X>t)$

§7 Correlation & Conditional probability

1. 相关系数: $\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\cdot\sqrt{Var(Y)}}$
   分母用来ensure $\rho\in[-1,1]$, $|\rho|$越大, 线性相关越强
2. Covariance
   $Cov(X,Y)=E[(X-E[X])(Y-E[Y])]=E[XY]-E[X]\cdot E[Y]$
   注: 1° Correlation 不会揭示 causality (因果关系)
   2° 若X与Y独立, 则 $E[XY]=E[X]\cdot E[Y]$
   3° 方差可理解为两变量相同的协方差
3. Conditional probability
   $f(x,y)=f(x)\cdot f(Y=y|X=x)$, $f(Y=y|X=x)=f(x,y)/f(x)$
4. 1° Uncorrelated: $E[XY]=E[X]\cdot E[Y]$
   2° Disjoint $\overset{\times}{\underset{\times}{\rightleftarrows}}$ independent $\underset{\times}{\rightleftarrows}$ uncorrelated

3. Statistics

§8 Statistics Estimation: point estimation (MLE)

1. Likehood function
   1° 对于 discrete model: Likehood: $L(\theta)=\prod_i P(X_i|\theta)$
   Log-Likehood: $l(\theta)=\sum_i \log(P(X_i|\theta))$ (P: PMF)
   2° 对于 Continuous model: Likehood: $L(\theta)=\prod_i f(X_i|\theta)$
   Log-Likehood: $l(\theta)=\sum_i \log(f(X_i|\theta))$ (f: PDF)
2. MLE步骤
   ① 求PDF/PMF ② 求每个model的(log-)likelihood ③ 比较
   对于continuous distribution, 还要check SOC: $f''(\theta)\le 0$,
   或证明Log-likelihood function为concave function.

例: Bernoulli R.V. N个样本成功M个, 求p
$L(p)=p^M(1-p)^{N-M}$ ($\binom{N}{M}$可略去) $l(p)=M\log p+(N-M)\log(1-p)$
$l'(p)=\frac{M}{p}-\frac{N-M}{1-p}=0 \Rightarrow p=\frac{M}{N}$

例: Uniform distribution [a,b], 样本$\{x_1,\dots,x_n\}$, 求a, b
$f(a,b)=\begin{cases}\frac{1}{a-b} & a\le x\le b\\ 0 & \text{else}\end{cases} \Rightarrow L(a,b)=\begin{cases}\frac{1}{(b-a)^n} & a\le X_i\le b\\ 0 & \text{else}\end{cases}$
$X_{(0)}=\min\{X_1,\dots,X_n\}$, $X_{(n)}=\max\{X_1,\dots,X_n\}$
$L(a,b)=\frac{1}{(b-a)^n}\le\frac{1}{(X_{(n)}-X_{(0)})^n} \Rightarrow a=X_{(0)}, b=X_{(n)}$

例: • Suppose we aim to use normal distribution to fit the data.
• Given parameter mean μ, variance σ^2, the model is $N(\mu,\sigma^2)$

$f(x|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$

• Samples: $X_1, X_2 \dots, X_n$
• $L(\mu,\sigma^2)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_i(X_i-\mu)^2}$
• $l(\mu,\sigma^2)=\text{Log } L(\mu,\sigma^2)=-\frac{1}{2\sigma^2}\sum_i(X_i-\mu)^2-\frac{n}{2}\log 2\pi-\frac{n}{2}\log\sigma^2$
• $l'(\mu)=\frac{1}{\sigma^2}\sum_i(X_i-\mu)=0$ $\hat\mu=\bar X$.
• $l'(\sigma^2)=\frac{1}{2\sigma^4}\sum_i(X_i-\mu)^2-\frac{n}{2}\frac{1}{\sigma^2}=0$ $\widehat{\sigma^2}=\frac{1}{n}\sum_i(X_i-\bar X)^2$
• $l(\mu,\sigma^2)=-\frac{1}{2\sigma^2}\sum_i(X_i-\mu)^2-\frac{n}{2}\log 2\pi-\frac{n}{2}\log\sigma^2$ is concave in (μ,σ^2)
• Proof:

Note that $l(\mu,\sigma^2)=-\frac{1}{2}\sum_i[\frac{(X_i-\mu)^2}{\sigma^2}+\log\sigma^2]-\frac{n}{2}\log 2\pi$

§9 Linear regression

1. 模型: x与Y呈linear dependent $\Rightarrow Y\sim N(\beta_0+\beta_1 x,\sigma^2)$
2. 确定β_0,β_1 (MLE)
   $L(\beta_0,\beta_1)=\frac{1}{(\sqrt{2\pi})^n\sigma^n}\cdot\exp\left[-\frac{1}{2}\cdot\frac{\sum_i(Y_i-\beta_1X_i-\beta_0)^2}{\sigma^2}\right]$
   仅需要minimize $g(\beta_0,\beta_1)=\sum_i(Y_i-\beta_1X_i-\beta_0)^2$
   $\frac{\partial g}{\partial\beta_0}=\sum_i(Y_i-\beta_1X_i-\beta_0)=0$
   $\frac{\partial g}{\partial\beta_1}=\sum_i(Y_i-\beta_1X_i-\beta_0)\cdot X_i=0$
   $\hat\beta_1=\frac{\sum_i(X_i-\bar X)(Y_i-\bar Y)}{\sum_i(X_i-\bar X)^2}$
   $\hat\beta_0=\bar Y-\hat\beta_1\bar X$
3. Residual analysis
   Residual: $e_i=Y_i-\beta_0-\beta_1X_i\sim N(0,\sigma^2)$
   检验两个假设:
   ① X与Y为linear relationship (e_i doesn't depend on X_i)
   ② 对∀X, $Y-\beta_0-\beta_1X$的σ^2均相同 ($Var(e_i)$ doesn't depend on X_i)

拓展: MLE of unknown parameters: σ, given $\hat\beta_1$, $\hat\beta_0$

■ **Likelihood function**:

$$L(\sigma)=\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2}\frac{\sum_{i=1}^n(Y_i-\hat\beta_1X_i-\hat\beta_0)^2}{\sigma^2}\right],$$

$$\log[L(\sigma)]=\underbrace{-\frac{n}{2}\log(2\pi)}_{constant}-n\log(\sigma)-\frac{1}{2\sigma^2}\sum_{i=1}^n(Y_i-\hat\beta_1X_i-\hat\beta_0)^2$$

■ **MLE of parameter**:

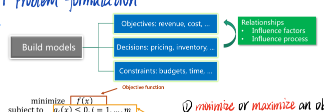
$$\hat\sigma^2=\frac{\sum_{i=1}^n(Y_i-\hat\beta_1X_i-\hat\beta_0)^2}{n}.$$

§10. Confidence interval

1. Central limit theorem
   1° 对任意概率分布, 只要每轮取样的样本空间足够大, 轮数足够多, 则各轮取样的sample mean服从正态分布.
   2° 样本空间越大, 正态分布曲线就越高, 越窄.
   3° 若true mean为μ, true variance为σ^2, 样本空间为n, 则
   $\bar X\sim N(\mu,\frac{\sigma^2}{n})$ 或写作 $\frac{\sqrt{n}(\bar X-\mu)}{\sigma}\sim N(0,1)$
2. Interval estimation (confidence level = $1-\alpha$)
   $1-\alpha$ Confidence interval: $[\bar X-z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt n}, \bar X+z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt n}]$
3. 影响因素
   ① 样本容量↑, CI长度↓ ② 方差σ^2↑, CI长度↑
   ③ confidence level $(1-\alpha)$↑, CI长度↑

4. Optimization

§11 Problem formulation

1.


minimize $f(x)$ (Objective function)
subject to $g_i(x)\le 0, i=1,\dots,m$; $h_i(x)=0, i=1,\dots,n$ (Constraints)

① minimize or maximize an objective
② subject to constraints

2. convex optimization的前提
   ① feasible set为convex set ② objective function为convex ..

§12 Convex set

1. 定义: $\forall x_1,x_2\in C$, $0\le\theta\le 1 \Rightarrow \theta x_1+(1-\theta)x_2\in C$, 即
   对于集合C, C内任意两点形成的线段均落在C内
2. 性质: 若S_1与S_2均为convex sets, 则$S_1\cap S_2$也为convex set.
3. 常见的凸集: ① ∅, 单个点集, R ② 直线 ③ halfspace ④ 球

例: 证明set $\{x=(x_1,\dots,x_n)\in R^n: \alpha\le a^Tx\le\beta\}$为凸集
• For any $x=(x_1,\dots,x_n)\in R^n$, $y=(y_1,\dots,y_n)\in R^n$, $\alpha\le a^Tx, a^Ty\le\beta$.
• Then for any $\theta\in[0,1]$,
$\theta\alpha\le\theta a^Tx\le\theta\beta$, $(1-\theta)\alpha\le(1-\theta)\alpha\le(1-\theta)a^Ty\le(1-\theta)\beta$
• $a^T(\theta x+(1-\theta)y)=\sum a_i\cdot[\theta x+(1-\theta)y]$
$=\sum a_i\cdot\theta x+\sum a_i\cdot(1-\theta)y$
$=\theta a^Tx+(1-\theta)a^Ty$
• Thus $\alpha\le a^T(\theta x+(1-\theta)y)\le\beta$, the set is convex.

§13 Convex function (凸函数/下凸函数)

1. 定义: 一个函数$f(x): R^n\to R$被称作convex, 若
   1° f的定义域为凸集 2° 对$\forall\vec x,\vec y\in dom(f)$, 且$0\le\lambda\le 1$, 有
   $f(\lambda\vec x+(1-\lambda)\vec y)\le\lambda f(\vec x)+(1-\lambda)f(\vec y)$
2. 凸函数的性质一(关于sublevel set)
   若f(x)为凸函数, 则满足$f(x)\le r$的所有定义域内的点$\vec x$构成的集合C为凸集
   证: 在集合C中任选两点$\vec x_1,\vec x_2$, 有$f(\vec x_1)\le r$, $f(\vec x_2)\le r$
   · 在点$\vec x_1,\vec x_2$形成的线段上任取一点, $\lambda\vec x_1+(1-\lambda)\vec x_2$
   · 由f(x)为凸函数得:
   $f(\lambda\vec x_1+(1-\lambda)\vec x_2)\le\lambda f(\vec x_1)+(1-\lambda)f(\vec x_2)\le r$
3. A function f with a convex domain is convex 当且仅当
   ① Line segments are always above the graph (定义)
   ② Epigraph $\{(\vec x,y): y\ge f(\vec x)\}$ is a convex set
   证: 在集合C中任选两点$(\vec x_1,y_1),(\vec x_2,y_2)$, 有$y_1\ge f(\vec x_1)$, $y_2\ge f(\vec x_2)$
   · 在点$(\vec x_1,y_1),(\vec x_2,y_2)$形成的线段上任取一点, $\lambda(\vec x_1,y_1)+(1-\lambda)(\vec x_2,y_2)$
   即 $<\lambda\vec x_1+(1-\lambda)\vec x_2, \lambda y_1+(1-\lambda)y_2>$
   · 由f(x)为凸函数得:
   $\lambda y_1+(1-\lambda)y_2\ge\lambda f(\vec x_1)+(1-\lambda)f(\vec x_2)\ge f(\lambda\vec x_1+(1-\lambda)\vec x_2)$
   证: 任取两点$(\vec x_1,f(\vec x_1))$与$(\vec x_2,f(\vec x_2))$
   · 在两点连线上取一点, $\lambda(\vec x_1,f(\vec x_1))+(1-\lambda)(\vec x_2,f(\vec x_2))$
   $=(\lambda\vec x_1+(1-\lambda)\vec x_2, \lambda f(\vec x_1)+(1-\lambda)f(\vec x_2))$
   · 由凸集定义, 有$\lambda f(\vec x_1)+(1-\lambda)f(\vec x_2)\ge f(\lambda\vec x_1+(1-\lambda)\vec x_2)$
   ③ FOC: 图像在任一切线上方
   $f(\vec x+s\vec e)\ge f(\vec x)+s\cdot\frac{df(\vec x+\theta\vec e)}{d\theta}\Big|_{\theta=0}$
   一维时: $f(x_2)\ge f(x_1)+(x_2-x_1)f'(x_1)$
   另一种表示形式: $f(\vec y)\ge f(\vec x)+\nabla f(\vec x)^T(\vec y-\vec x)$

证: For any $\boldsymbol{x},\boldsymbol{y}\in$ **dom f**,
• denote $\boldsymbol{e}$ as the unit vector from $\boldsymbol{x}$ to $\boldsymbol{y}$
• let $\boldsymbol{y}=\boldsymbol{x}+s\,\boldsymbol{e}$

Let $\boldsymbol{z}=(1-\theta)\boldsymbol{x}+\theta\boldsymbol{y}=\boldsymbol{x}+\theta\, s\,\boldsymbol{e}$, by convexity of f, we have

$$(1-\theta)f(\boldsymbol{x})+\theta f(\boldsymbol{y})\ge f(\boldsymbol{z})$$

$$f(\boldsymbol{y})\ge f(\boldsymbol{x})+\frac{f(\boldsymbol{Z})-f(\boldsymbol{x})}{\theta}=f(\boldsymbol{x})+\frac{f(x+\theta\, s\, e)-f(x)}{\theta}$$

When $\theta\to 0$, $\frac{f(x+\theta s e)-f(x)}{\theta}\to s\frac{d\,f(\boldsymbol{x}+\theta\boldsymbol{e})}{d\theta}\Big|_{\theta=0}$

证: • Denote $\boldsymbol{e}$ as the unit vector from $\boldsymbol{x}$ to $\boldsymbol{y}$
• Let $\boldsymbol{y}=\boldsymbol{x}+s\,\boldsymbol{e}$
• Let $\boldsymbol{z}=(1-t)\boldsymbol{x}+t\,\boldsymbol{y}=\boldsymbol{x}+t\,s\,\boldsymbol{e}$,

As a result, $\boldsymbol{x}=\boldsymbol{z}-t\,s\,\boldsymbol{e}$, $\boldsymbol{y}=\boldsymbol{z}+(s-t\,s)\,\boldsymbol{e}$,

$f(\boldsymbol{x})\ge f(\boldsymbol{z})-ts\frac{d\,f(z+\theta e)}{d\theta}\Big|_{\theta=0}$ and

$f(\boldsymbol{y})\ge f(\boldsymbol{z})+s(1-t)\frac{d\,f(z+\theta e)}{d\theta}\Big|_{\theta=0}$

Multiplying the first inequality by 1-t, the second by t,

and adding them yields $(1-t)f(\boldsymbol{x})+t\,f(\boldsymbol{y})\ge f(\boldsymbol{z})$

④ SOC: $\frac{d^2f(\vec x+\theta\vec e)}{d\theta\, d\theta}\ge 0$
一维时: $f''(x)\ge 0$ for all x
另一种表示形式: $\nabla^2 f(x)\ge 0$ (Hessian)

证: • Denote $\boldsymbol{e}$ as the unit vector from $\boldsymbol{x}$ to $\boldsymbol{y}$
• Let $\boldsymbol{y}=\boldsymbol{x}+s\,\boldsymbol{e}$, then

$0\le\int_0^s(s-t)\,g''(t;\boldsymbol{x},\boldsymbol{e})\,dt$
$=-s\,g'(0;\boldsymbol{x},\boldsymbol{e})+\int_0^s g'(t;\boldsymbol{x},\boldsymbol{e})\,dt$
$=-s\,g'(0;\boldsymbol{x},\boldsymbol{e})+g(s;\boldsymbol{x},\boldsymbol{e})-g(0;\boldsymbol{x},\boldsymbol{e})$
$=-s\,g'(0;\boldsymbol{x},\boldsymbol{e})+f(\boldsymbol{y})-f(\boldsymbol{x})$

• Then $f(\boldsymbol{y})\ge f(\boldsymbol{x})+s\,g'(0;\boldsymbol{x},\boldsymbol{e})$

证: For any $\boldsymbol{x}$ and θ,
• let $\boldsymbol{y}=\boldsymbol{x}+(s+\theta)\boldsymbol{e}$ and $\boldsymbol{z}=\boldsymbol{x}+\theta\,\boldsymbol{e}$ with $s>0$
• $\boldsymbol{y}=\boldsymbol{z}+s\,\boldsymbol{e}$ and $\boldsymbol{z}=\boldsymbol{y}-s\,\boldsymbol{e}$

By FOC of f, we have

$$f(\boldsymbol{y})\ge f(\boldsymbol{z})+s\,g'(0;\boldsymbol{z},\boldsymbol{e}),\ f(\boldsymbol{z})\ge f(\boldsymbol{y})-sg'(0;\boldsymbol{y},\boldsymbol{e})$$

Sum both sides together,

$$g'(0;\boldsymbol{y},\boldsymbol{e})-g'(0;\boldsymbol{z},\boldsymbol{e})\ge 0\text{, then }\frac{g'(0;\,y,\,e)-g'(0;\,z,\,e)}{s}\ge 0$$

As $g(t;\boldsymbol{y},\boldsymbol{e})=f(\boldsymbol{y}+t\boldsymbol{e})=f(\boldsymbol{x}+(t+s+\theta)\boldsymbol{e})=g(t+s+\theta;\boldsymbol{x},\boldsymbol{e})$, $g'(t;\boldsymbol{y},\boldsymbol{e})=g'(t+s+\theta;\boldsymbol{x},\boldsymbol{e})$. Similarly, $g'(t;\boldsymbol{z},\boldsymbol{e})=g'(t+\theta;\boldsymbol{x},\boldsymbol{e})$. Accordingly,

$$0\le\frac{g'(0;\,y,\,e)-g'(0;\,z,\,e)}{s}=\frac{g'(s+\theta;\,x,\,e)-g'(\theta;\,x,\,e)}{s}.$$

When s is approaching 0, $g''(\theta;\boldsymbol{x},\boldsymbol{e})\ge 0$

4. 凸优化性质: local minimizer为global minimizer
   证: · Let $\bar x$ be a local minimum.
   $\Rightarrow\bar x\in\Omega$ and $\exists\epsilon>0$ s.t. $f(\bar x)\le f(x)$, $\forall x\in B(\bar x,\epsilon)$
   · Suppose $\exists z\in\Omega$ with $f(z)<f(\bar x)$
   · Because of convexity of Ω, we have
   $\lambda\bar x+(1-\lambda)z\in\Omega$, $\forall\lambda\in[0,1]$
   · By convexity of f, we have
   $f(\lambda\bar x+(1-\lambda)z)\le\lambda f(\bar x)+(1-\lambda)f(z)$
   $<\lambda f(\bar x)+(1-\lambda)f(\bar x)=f(\bar x)$
   · But, as $\lambda\to 1$, $\lambda\bar x+(1-\lambda)z\to\bar x$ and
   the previous inequality contradicts local optimality of $\bar x$.
5. 凸函数的性质二(保凸运算)
   ① 若f凸, 则$g(\vec x)=f(a\vec x+b)-c$也凸.
   证: $g(\lambda\vec x_1+(1-\lambda)\vec x_2)$
   $=f[a(\lambda\vec x_1+(1-\lambda)\vec x_2)+b]-c$
   $=f[\lambda(a\vec x_1+b)+(1-\lambda)(a\vec x_2+b)]-c$
   $\le\lambda f(a\vec x_1+b)+(1-\lambda)f(a\vec x_2+b)-c$
   $=\lambda\cdot(f(a\vec x_1+b)-c)+(1-\lambda)\cdot(f(a\vec x_2+b)-c)$
   $=\lambda\cdot g(\vec x_1)+(1-\lambda)g(\vec x_2)$
   ② 若$f_1,\dots,f_m$凸, 则$f(\vec x)=\max\{f_1(\vec x),\dots,f_m(\vec x)\}$也凸.
   证: $epi\ f=\bigcap_{i=1}^m(epi\ f_i)$
   由$epi\ f_i$为凸集, 得$epi\ f$为凸集. 因此f为凸函数
   ③ 若$f_1,\dots,f_n$凸, $w_1,\dots,w_n>0$, 则$f=w_1f_1+\dots+w_nf_n$凸.
   证: 对任意$\vec x,\vec y$, $0\le\lambda\le 1$和m.
   $f_m(\lambda\vec x+(1-\lambda)\vec y)\le\lambda f_m(x)+(1-\lambda)f_m(y)$
   · 在不等式两侧同乘w_m, 再对m求和, 得
   $w_1f_1(\lambda\vec x+(1-\lambda)\vec y)+\dots+w_nf_n(\lambda\vec x+(1-\lambda)\vec y)$
   $\le\lambda(w_1f(\vec x)+\dots+w_nf(\vec x))+(1-\lambda)\cdot(w_1f(\vec y)+\dots+w_nf(\vec y))$
   · 得: $f(\lambda\vec x+(1-\lambda)\vec y)\le\lambda f(\vec x)+(1-\lambda)f(\vec y)$
   推广: 若f(x,y)对任意$y\in A$都convex in x, 且对每个$y\in A$, 都有$w(y)\ge 0$, 则 $g(x)=\int_A w(y)f(x,y)\,dy$ 也convex in x
   注: 可将g(x)理解为$\sum_{y\in A}w(y)\cdot dy\cdot f_y(x)$,
   其中将$w(y)\cdot dy$视作整体, 其大于0
   例: 证明: 若f为凸函数(可能不可微), $x>0$, 则
   $F(x)=\frac{1}{x}\int_0^x f(t)\,dt$ 也为凸函数
   set $t=xz$ and let z be the integration variable, then the integration domain is changed to [0, 1]
   $F(x)=\frac{1}{x}\int_0^x f(t)\,dt=\frac{1}{x}\int_0^1 f(xz)\,d(xz)$
   $=\frac{1}{x}\int_0^1 f(xz)\cdot x\,dz=\int_0^1 f(xz)\,dz$
   ④ 若f关于(x,y)(所有变量)凸, 且C为非空凸集, 则函数
   $g(x)=\inf_{y\in C}f(x,y)$ 也凸
   注: 注意与maximization的区别, 不是所有minimization form均成立

证: We prove this by verifying Jensen's inequality for $x_1, x_2\in\mathbf{dom}\,g$. Let $\epsilon>0$. Then there are $y_1, y_2\in C$ such that $f(x_i,y_i)\le g(x_i)+\epsilon$ for $i=1,2$. Now let $\theta\in[0,1]$. We have

$$\begin{aligned} g(\theta x_1+(1-\theta)x_2) &= \inf_{y\in C} f(\theta x_1+(1-\theta)x_2,y)\\ &\le f(\theta x_1+(1-\theta)x_2,\theta y_1+(1-\theta)y_2)\\ &\le \theta f(x_1,y_1)+(1-\theta)f(x_2,y_2)\\ &\le \theta g(x_1)+(1-\theta)g(x_2)+\epsilon.\end{aligned}$$

Since this holds for any $\epsilon>0$, we have

$$g(\theta x_1+(1-\theta)x_2)\le\theta g(x_1)+(1-\theta)g(x_2).$$

例: A newsboy needs to decide how many newspapers to buy each morning in order to generate the most expected sales profit. The demand D for newspapers in a day is random, with f(·) being its pdf and F(·) being its CDF. He needs to buy each newspaper at price c, and can sell it at price p.

Objective: maximize $E[profit]=E[p\min(q,D)]-cq$
$E[\min(q,D)]=\int_0^q xf(x)\,dx+q(1-F(q))$.
$E[profit]=p\int_0^q xf(x)\,dx+pq(1-F(q))-cq$
SDC: $\frac{\partial}{\partial q}E[profit]=pqf(q)+pq(-F'(q))+p[1-F(q)]-c$
$=p[1-F(q)]-c$
$\frac{\partial^2}{\partial q^2}E[profit]=p[-F'(q)]\le 0$ (F递增)
$\frac{\partial}{\partial q}E[profit]=0\Rightarrow q^*=F^{-1}(\frac{p-c}{p})$

§14 Quasiconvex (拟凸函数)

1. 定义: $S_\alpha=\{x|f(x)\le\alpha\}$ is convex for each $\alpha\in R$, 即
   A function $f: R^n\to R$ is quasiconvex if all its sublevel sets are convex.
   注: 所有的凸函数均为拟凸函数, 拟凸函数can be concave
   Epigraph与凸函数能建立双向联系
   Sublevel set与凸函数不能建立双向联系
2. f is quasiconvex if and only if
   ① f减/f增/f先不增, 后不减

证: First suppose that f satisfies one of the conditions and suppose that for some number α, both x_1 and x_2 are members of the sublevel set $S_\alpha=\{x|\ f(x)\le\alpha\}$. Then $f(x_1)\le\alpha$ and $f(x_2)\le\alpha$. Therefore $f(x)\le\alpha$ for every point x between x_1 and x_2. Therefore S_α is convex.

证: Suppose f is quasiconvex. If f does not satisfy any of the conditions then we can find x_1 and x_2 and x_3 such that $x_1<x_2<x_3$ and $f(x_2)>\max(f(x_1), f(x_3))$. Then the sublevel set S_α for $\alpha=\max(f(x_1), f(x_3))$ includes x_1 and x_3, but not x_2, this contradicts with the assumption that f is quasiconvex. Therefore, if f is quasiconvex, then f must satisfy one of the conditions.

② $f(\theta\vec x+(1-\theta)\vec y)\le\max\{f(\vec x),f(\vec y)\}$ for any $\vec x,\vec y$ and $0\le\theta\le 1$.

证: Suppose f is quasiconvex. Pick any **x** and **y** in the domain of f. Let α = max(f(**x**), f(**y**)), then **x** and **y** belongs to the sublevel set S_α= {x| f(x) ≤ α}. Since f is quasiconvex, then S_α is convex, and θ**x** + (1 − θ)**y** ∈ S_α. Therefore f(θ**x** + (1 − θ)**y**) ≤ α = max {f(**x**), f(**y**)}.

证: Pick any sublevel set S_α and **x**, **y** ∈ S_α. Then max {f(**x**), f(**y**)} ≤ α. Since f(θ**x** + (1 − θ)**y**) ≤ max {f(**x**), f(**y**)} for 0 ≤ θ ≤1, therefore f(θ**x** + (1 − θ)**y**) ≤ max {f(**x**), f(**y**)} ≤ α, which means θ**x** + (1 − θ)**y** ∈ S_α. Therefore S_α is convex, and f is quasiconvex.

③ FOC: 对$\forall f(\vec y)\le f(\vec x)$, 令$\vec y=\vec x+s\vec e$, 有 $s\frac{df(\vec x+t\vec e)}{dt}\Big|_{t=0}\le 0$
一维时: $f(y)\le f(x)\Rightarrow f'(x)(y-x)\le 0$

证: First suppose f is a differentiable function on **R** and satisfies

$$f(y)\le f(x)\Longrightarrow f'(x)(y-x)\le 0. \qquad (3.43.A)$$

Suppose $f(x_1)\ge f(x_2)$ where $x_1\ne x_2$. We assume $x_2>x_1$ (the other case can be handled similarly), and show that $f(z)\le f(x_1)$ for $z\in[x_1,x_2]$. Suppose this is false, *i.e.*, there exists a $z\in[x_1,x_2]$ with $f(z)>f(x_1)$. Since f is differentiable, we can choose a z that also satisfies $f'(z)<0$. By (3.43.A), however, $f(x_1)<f(z)$ implies $f'(z)(x_1-z)\le 0$, which contradicts $f'(z)<0$.

证: To prove sufficiency, assume f is quasiconvex. Suppose $f(x)\ge f(y)$. By the definition of quasiconvexity $f(x+t(y-x))\le f(x)$ for $0<t\le 1$. Dividing both sides by t, and taking the limit for $t\to 0$, we obtain

$$\lim_{t\to 0}\frac{f(x+t(y-x))-f(x)}{t}=f'(x)(y-x)\le 0,$$

which proves (3.43.A).

④ SOC:
· 若f quasiconvex, 则对$\forall\vec x$与$\vec e$,
若 $\frac{df(\vec x+t\vec e)}{dt}\Big|_{t=0}=0$, 则 $\frac{d^2f(\vec x+t\vec e)}{dt^2}\Big|_{t=0}\ge 0$
· 若f满足以下条件, 则f quasiconvex
对$\forall\vec x$与$\vec e$, 若 $\frac{df(\vec x+t\vec e)}{dt}\Big|_{t=0}=0$, 则始终有 $\frac{d^2f(\vec x+t\vec e)}{dt^2}\Big|_{t=0}>0$
一维时, $f'(x)=0\Rightarrow f''(x)\ge 0$

证: We first show that if $f:\mathbf{R}\to\mathbf{R}$ is quasiconvex on an interval (a,b), then it must satisfy (3.21), *i.e.*, if $f'(c)=0$ with $c\in(a,b)$, then we must have $f''(c)\ge 0$. If $f'(c)=0$ with $c\in(a,b)$, $f''(c)<0$, then for small positive ϵ we have $f(c-\epsilon)<f(c)$ and $f(c+\epsilon)<f(c)$. It follows that the sublevel set $\{x\mid f(x)\le f(c)-\epsilon\}$ is disconnected for small positive ϵ, and therefore not convex, which contradicts our assumption that f is quasiconvex.

证: Now we show that if the condition (3.22) holds, then f is quasiconvex. Assume that (3.22) holds, *i.e.*, for each $c\in(a,b)$ with $f'(c)=0$, we have $f''(c)>0$. This means that whenever the function f' crosses the value 0, it is strictly increasing. Therefore it can cross the value 0 at most once. If f' does not cross the value 0 at all, then f is either nonincreasing or nondecreasing on (a,b), and therefore quasiconvex. Otherwise it must cross the value 0 exactly once, say at $c\in(a,b)$. Since $f''(c)>0$, it follows that $f'(t)\le 0$ for $a<t\le c$, and $f'(t)\ge 0$ for $c\le t<b$. This shows that f is quasiconvex.

3. 拟凸函数的性质(保拟凸运算)
   ① 若f拟凸, $w>0$, 则wf也拟凸.
   下水平集相同: $wf(y)\le\alpha$ & $f(x)\le\alpha/w$
   ② 若$f: R^n\to R$拟凸, h不减, 则$h\circ f$拟凸
   下水平集相同: $h(f(y))\le\alpha$ & $f(x)\le h^{-1}\alpha$
   ③ 若f拟凸, 则$g(\vec x)=f(a\vec x+b)$拟凸
   证: · Want to show g(x) is quasi convex
   That is equivalent to showing, $\forall\alpha$, $S_\alpha^{(g)}$ is convex
   · Pick any $x_1,x_2\in S_\alpha^{(g)}$.
   Want to show $\theta x_1+(1-\theta)x_2\in S_\alpha^{(g)}$
   That is equavalent to showing, $\forall\alpha$, $g(\theta x_1+(1-\theta)x_2)\le\alpha$
   · $g(\theta x_1+(1-\theta)x_2)=f(a[\theta x_1+(1-\theta)x_2]+b)$
   $=f(\theta(ax_1+b)+(1-\theta)(ax_2+b))$
   Since $f(ax_1+b)=g(x_1)\le\alpha$; $f(ax_2+b)=g(x_2)\le\alpha$,
   This means $ax_1+b, ax_2+b\in S_\alpha^{(f)}$
   Since f is quasiconvex, $\theta(ax_1+b)+(1-\theta)(ax_2+b)\in S_\alpha^{(f)}$
   ④ 若f_1,f_2为拟凸, 则$g(x)=\max\{f_1(x),f_2(x)\}$也拟凸
   证: $S_\alpha^{(f)}=\bigcap_{i=1}^m(S_\alpha^{(f_i)})$
   由$S_\alpha^{(f_i)}$为凸集, 得$S_\alpha^{(f)}$为凸集. 因此f为拟凸函数
   ⑤ 若f关于(x,y)(所有变量)拟凸, 且C为非空凸集, 则函数
   $g(x)=\inf_{y\in C}f(x,y)$ 也拟凸
   注: 若$C\in R$, 可理解为将f(x,y)投影到一垂直于定义域平面的平面上,
   所构成图形的下边缘构成的函数为拟凸.
   证: · Pick x_1,x_2,α such that $g(x_1)\le\alpha$, $g(x_2)\le\alpha$
   Want to show $g(\theta x_1+(1-\theta)x_2)\le\alpha$
   · $g(\theta x_1+(1-\theta)x_2)=\min_{y\in C}f(\theta x_1+(1-\theta)x_2,y)\le f(\theta x_1+(1-\theta)x_2, \theta y_1+(1-\theta)y_2)$
   where y_1,y_2 satisfy $g(x_1)=f(x_1,y_1)\le\alpha$, $g(x_2)=f(x_2,y_2)\le\alpha$
   · Then $(\theta x_1+(1-\theta)x_2, \theta y_1+(1-\theta)y_2)=\theta(x_1,y_1)+(1-\theta)(x_2,y_2)$
   $g(\theta x_1+(1-\theta)x_2)\le f(\theta x_1+(1-\theta)x_2, \theta y_1+(1-\theta)y_2)$
   $=f(\theta(x_1,y_1)+(1-\theta)(x_2,y_2))\le\alpha$
   ⑥ f_1,f_2拟凸, 则f_1+f_2不一定拟凸 (例: x^3+x^2)
4. Why learn quasi-convex
   · 若f拟凸, 则optimal minimizers form a convex set.
   · 若x^*为一个strict local minimizer, 则$\vec x^*$同时也是global ...

5. Machine Learning

15.1 Machine learning

1. 流程与术语：① input data的质量 ② model choice itself
2. 运行步骤：①选择并准备一个 training data set ②选择一个 algorithm 来应用于 training data set ③训练 algorithm 来建立模型 ④使用并提升 model
3. Supervised learning
   ① 使用 labeled data ② training data: $(\vec{x}_i, y_i)$ ③算法目的：a new $\vec{x}$, predict the output $\vec{y}$ ④ Not all x of samples impact y of new data; All y of samples impact y of new data.
4. Unsupervised learning
   ① 使用 data that does not have structured of objective answer
   ②算法目的：examine the information and identify structure within it.

15.2 KNN

1. 步骤：①找K个最靠近 X 的 training points $X_1 \cdots X_K$
   ②若 $X_1 \sim X_K$ 中有较多的属于 classifier C，则 label X 为 C
   注：K 较小： noise 对结果的影响较大 (overfitting)
   K 较大：破坏 KNN 的基本原则 (underfitting)
   若 K > 各data点数，任一新 input 都有相同的 label.
   注：KNN 的决策边界为非线性的

15.3 Logistic regression model

1. Logistic regression (二分类：$y \in \{0, 1\}$)

$$P(y=1 \mid x, \theta, b) = \frac{1}{1+\exp(-(\theta^T x+b))}$$

$$P(y=0 \mid x, \theta, b) = \frac{\exp(-(\theta^T x+b))}{1+\exp(-(\theta^T x+b))} = 1 - P(y=1 \mid x, \theta, b)$$

注：对于多分类问题：$y \in \{0, 1, \cdots, k\}$

$$\frac{P(y=i)}{P(y=0)} = e^{\theta_i^T x + b_i} \qquad P(y=0 \mid x, \theta, b) = \frac{\exp(-(\theta_i^T x+b_i))}{1+\sum_i \exp(-(\theta_i^T x+b_i))}$$

注：逻辑回归得出的决策边界为线性的

2. MLE for logistic regression

找出 θ, b，使得 likelihood of the labels 最大

$$\max_{\theta,b} l(\theta, b) := \log \prod_i P(y^i \mid x^i, \theta, b) = \sum_i \log P(y^i \mid x^i, \theta, b)$$

通常，我们只需将 averaged likelihood 最大化

$$\max_{\theta,b} l(\theta, b) := \frac{1}{n} \sum_i \log P(y^i \mid x^i, \theta, b)$$

1° $l(\theta, b)$ is concave in (θ, b)

$\log P(y^i \mid x^i, \theta, b) = (y^i - 1)(\theta x^i + b) - \log(1 + \exp(-\theta x^i - b))$

也就要证明 $l(\theta, b) := \sum_i \log P(y^i \mid x^i, \theta, b)$ concave，

仅需要证明 $\log(1 + \exp(-\theta x^i - b))$ convex.

引理：若 $f(\vec{x}+t\vec{u})$ 对任意 $\vec{x}, \vec{u}$ 皆为 convex in t，则 f is convex in $\vec{x}$

证明：令 $g(t) = f(\vec{x}+t\vec{u})$，取 $\vec{x}_1, \vec{x}_2$，使 $\vec{x}_3 = \vec{x}_1 + t_3\vec{u}$

若 $f(\vec{x}+t\vec{u})$ is convex in t，则 $\lambda g(t_1) + (1-\lambda) g(t_2) \geq g(\lambda t_1 + (1-\lambda) t_2)$

即 $\lambda f(\vec{x}_1 + t_1\vec{u}) + (1-\lambda) f(\vec{x}_2 + t_2\vec{u}) \geq f(\vec{x}_3 + t_3\vec{u})$

即 $\lambda f(\vec{x}_1) + (1-\lambda) f(\vec{x}_2) \geq f(\lambda\vec{x}_1 + (1-\lambda)\vec{x}_2) = f(\lambda(\vec{x}_1 + t_1\vec{u}) + (1-\lambda)(\vec{x}_2 + t_2\vec{u}))$

$= f(\vec{x}_3 + (1-\lambda) t_2\vec{u})$

令 $(\theta, b) = (\theta_0, b_0) + t\vec{e}$，$(\theta_1, b_1)$ 为 any point，$\vec{e} = (\theta, b)$ 为任意单位向量

则 $h(t) := \log(1 + \exp(-\theta x - b))$

$= \log\{1 + \exp[(-\theta_0 x - b_0) + t(-\theta_1 x - b_1)]\}$

$= \log\{1 + \exp[C_1 + tC_2]\}$

则 $h''(t) = \frac{C_2^2}{(1+\exp[C_1+tC_2])^2} \cdot \exp[C_1 + tC_2] \geq 0$

2° No closed form solution maximizing $l(\theta, b)$

3. Gradient descent

$x^{(t+1)} = x^{(t)} - \alpha^{(t)} f'(x^{(t)})$ (x 为参数)

注：前几步 $\alpha^{(t)}$ 可以大一些，靠近极小值时小一些

若 $\alpha^{(t)}$ 为常数，可能会找不到极小值点

- At each iteration, we randomly choose a small batch of data points, and update using the stochastic gradient

$$\theta^{t+1} \leftarrow \theta^t + \alpha^{(t)} \frac{1}{|B|} \sum_{i \in B} (y^i - 1) x^i + \frac{\exp(-\theta^t x^i - b^t) x^i}{1 + \exp(-\theta^t x^i - b^t)}$$

$$b^{t+1} \leftarrow b^t + \alpha^{(t)} \frac{1}{|B|} \sum_{i \in B} (y^i - 1) + \frac{\exp(-\theta^t x^i - b^t)}{1 + \exp(-\theta^t x^i - b^t)}$$

- B: the batch we use in each iteration

§16 K-means clustering

1. Steps
   1° 随机选取 k 个 cluster centers $\{C^1, C^2, \cdots, C^k\}$
   2° 分类并调整 cluster centers
   ① cluster assignment
   将每个数据点 x^i 归入最近的 cluster centers
   $\pi(i) = \operatorname{argmin}_{j=1,\cdots,k} d(x^i, C^j)$ (通常为 $\|x^i - C^j\|^2$)
   ② center adjustment
   将 cluster centers 移至该类数据点的平均位置

$$C^j = \frac{1}{|\{i : \pi(i) = j\}|} \cdot \sum_{i : \pi(i)=j} x^i \quad \text{(在 } C^j \text{ 类别内的点的个数)}$$

   3° 重复上一步直到 cluster centers 收敛或 $\pi(i)$ 不改变
   注：1° 不同的 initialization 可能会导致不同的结果
   2° algorithm 并适合在 some iteration 之后停下来 tie breaking rule

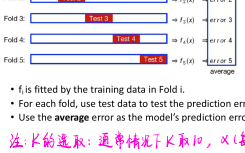
2. Similarity / Dissimilarity function
   ① Symmetry (对称性)：$d(x,y) = d(y,x)$
   ② Positive separability：$d(x,y) = 0$ 当且仅当 $x = y$
   ③ Triangular inequality：$d(x,y) \leq d(x,z) + d(z,y)$

3. Distance functions
   Minkowski distance：$d(x,y) = \sqrt[p]{\sum_{i=1}^n (x_i - y_i)^p}$
   - Euclidian distance：$p = 2$
   - Manhattan distance：$p = 1$，$d(x,y) = \sum_{i=1}^n |x_i - y_i|$
   - "inf" distance：$p = \infty$，$d(x,y) = \max_{i=1}^n |x_i - y_i|$

§17 K-fold validation

1. Validation set approach
   对于每一个可选的 model m：① 用 training data 来 train model
   ② 用 trained model 来预测 test data，得出 prediction error (Errm)
   ③ 选出 test error 最小的 model (min Errm)
   注：此方法选出的模型高度取决于 your separation of the data.
   解决方法：重复多次，每次使用不同的 training and test data.

2. K-fold validation
   - 5-fold cross-validation (blank: training; red: test)



Error

Model 1 $h(\theta, X)$	Model 2 $g(\gamma, X)$
$Err_1(\theta_1)$	$Err_1(\gamma_1)$
$Err_2(\theta_2)$	$Err_2(\gamma_2)$
$Err_3(\theta_3)$	$Err_3(\gamma_3)$
$Err_4(\theta_4)$	$Err_4(\gamma_4)$
$Err_5(\theta_5)$	$Err_5(\gamma_5)$

Model 1 is better iff. $\frac{1}{K}\sum_{i=1}^K Err_{\theta_i}(\theta_i) < \frac{1}{K}\sum_{i=1}^K Err_{\gamma_i}(\gamma_i)$

- f_i is fitted by the training data in Fold i.
- For each fold, use test data to test the prediction error.
- Use the **average** error as the model's prediction error.

注：K的选取：通常情况下 K 取 10，α (每一部分的占比) $= \frac{1}{K} = 0.1$
K 大大会 time-consuming. 若不考虑时间成本，K 越大越好

---

1. MLE: $X_1, \cdots X_n \sim Ber(p)$ max $l(p)$, $0 \leq p \leq 1$

$f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$

$L(p) = \prod_{i=1}^n f(x_i, p) = p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n}$

$l(p) = \sum_{i=1}^n x_i \cdot \log p + \sum_{i=1}^n 1 - x_i \cdot \log(1-p)$ $\quad l'(p) = \sum_{i=1}^n x_i \cdot \frac{1}{p} + \sum_{i=1}^n x_i - 1 \cdot \frac{1}{1-p} = 0$

$\Rightarrow \sum_{i=1}^n x_i (1-p) + (\sum_{i=1}^n x_i - n) p = 0 \quad \Rightarrow \hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$

2. MLE

A geologist studied the composition of rocks in the shoreline area of Lake Michigan. He randomly selected 100 samples from the area, each containing 10 stones, and recorded the number of limestone stones in each sample. The geologist's data are as follows:

number of limestone in a sample	0	1	2	3	4	5	6	7	8	9	10
number of samples	0	1	6	7	23	26	21	12	3	1	0

number of limestone in a sample	0	1	2	3	4	5	6	7	8	9	10
number of samples	0	1	6	7	23	26	21	12	3	1	0

X_i = # limestone in sample i.

Assume that these 100 observations are independent of each other. Find the maximum likelihood estimate (MLE) of the proportion p of limestone in the stones in this area.

$$L(p) = \prod_{i=1}^{100} \left[\binom{10}{x_i} p^{x_i} (1-p)^{10-x_i} \right] \qquad 10 \times 100 \quad M$$

$$= \left[\prod_{i=1}^{100} \binom{10}{x_i} \right] \cdot p^{\sum x_i} \cdot (1-p)^{1000 - \sum x_i} \qquad \hat{p} = \frac{M}{1000}$$

$$l(p) = \log\left[\prod \binom{10}{x_i}\right] + \sum x_i \log p - (1000 - \sum x_i) \log(1-p)$$

3. MLE

- Suppose we ask people to select their favorite ice cream brand: Haagen-Dazs (type 1), Wall's (type 2), Ben & Jerry's (type 3).
- We can view the large population contains 3 different types of individuals, and let $\theta_1, \theta_2, \theta_3$ denote the proportion of individuals of type each type. Clearly, $0 \leq \theta_i \leq 1$ and $\theta_1 + \theta_2 + \theta_3 = 1$. $\theta_1, \theta_2, \theta_3$ unknown.
- Suppose we ask 100 students in CUHK(SZ), and 45 choose HD, 30 choose Walls, and 25 choose BJ. Find the MLE for $\theta_1, \theta_2, \theta_3$.

$L(\theta_1, \theta_2, \theta_3) = \theta_1^{45} \cdot \theta_2^{30} \cdot \theta_3^{25}$ $\quad l(\theta_1, \theta_2, \theta_3) = 45 \log \theta_1 + 30 \log \theta_2 + 25 \log(1 - \theta_1 - \theta_2)$

$$\begin{cases} \frac{45}{\theta_1} - \frac{25}{1-\theta_1-\theta_2} = 0 \\ \frac{30}{\theta_2} - \frac{25}{1-\theta_1-\theta_2} = 0 \end{cases} \Rightarrow \theta_1 = 0.45 \quad \theta_2 = 0.3 \quad \theta_3 = 0.25$$

4. Convex set

Let $C \subseteq \mathbf{R}^n$ be a convex set, with $x_1, \ldots, x_k \in C$, and let $\theta_1, \ldots, \theta_k \in \mathbf{R}$ satisfy $\theta_i \geq 0$, $\theta_1 + \cdots + \theta_k = 1$. Show that $\theta_1 x_1 + \cdots + \theta_k x_k \in C$. (The definition of convexity is that this holds for $k = 2$; you must show it for arbitrary k.) *Hint.* Use induction on k.

For $k=3$, show $y = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \in C$. $\theta_1 + \theta_2 + \theta_3 = 1$.

- At least one $\theta \neq 1$, assume $\theta_1 \neq 1$ $\quad y = \theta_1 x_1 + (1-\theta_1)(\mu_2 x_2 + \mu_3 x_3)$

where $\mu_2 = \frac{\theta_2}{1-\theta_1}$, $\mu_3 = \frac{\theta_3}{1-\theta_1} > 0$, $\mu_2 + \mu_3 = 1$

since C is convex and $x_2, x_3 \in C$, $\mu_2 x_2 + \mu_3 x_3 \in C$, since $x_1 \in C$, $y \in C$

5. Convex set

Midpoint convexity. A set C is *midpoint convex* if whenever two points a, b are in C, the average or midpoint $(a+b)/2$ is in C. Obviously a convex set is midpoint convex. It can be proved that under mild conditions midpoint convexity implies convexity. In a simple case, prove that if C is closed and midpoint convex, then C is convex.

We have to show that $\theta x + (1-\theta) y \in C$ for all $\theta \in [0, 1]$

Let $\theta^{(k)}$ be the binary number of length k, i.e. $\theta^{(k)} = c_1 2^{-1} + c_2 2^{-2} + \cdots + c_k 2^{-k}$ with $c_i \in \{0, 1\}$, closest to θ. By midpoint convexity (applied k times recursively) $\theta^{(k)} x + (1-\theta^{(k)}) y \in C$, because C is closed,

$\lim_{k \to \infty} (\theta^{(k)} x + (1-\theta^{(k)}) y) = \theta x + (1-\theta) y \in C$

6. Convex set

Hyperbolic sets. Show that the *hyperbolic set* $\{x \in \mathbf{R}^2_+ \mid x_1 x_2 \geq 1\}$ is convex. As a generalization, show that $\{x \in \mathbf{R}^n_+ \mid \prod_{i=1}^n x_i \geq 1\}$ is convex. *Hint.* If $a, b \geq 0$ and $0 \leq \theta \leq 1$, then $a^\theta b^{1-\theta} \leq \theta a + (1-\theta) b$; see §3.1.9.

(a)

Consider a convex combination z of two points $(x_1, x_2), (y_1, y_2)$ in the set

If $x \geq y$, then $z = \theta x + (1-\theta) y \geq y$ and $z_1 z_2 \geq y_1 y_2 \geq 1$

If $x \leq y$, Similar proof

Suppose $y \not\geq 0$ and $x \not\geq y$, i.e. $(y_1 - x_1)(y_2 - x_2) < 0$. Then

$(\theta x_1 + (1-\theta) y_1)(\theta x_2 + (1-\theta) y_2)$

$= \theta^2 x_1 x_2 + (1-\theta)^2 y_1 y_2 + \theta(1-\theta) x_1 y_2 + \theta(1-\theta) x_2 y_1$

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

≥ 1

(b) Use hint

Assume $\prod x_i > 1$ and $\prod y_i \geq 1$

$\prod (\theta x_i + (1-\theta) y_i) \geq \prod x_i^\theta y_i^{1-\theta} = (\prod x_i)^\theta (\prod y_i)^{1-\theta} \geq 1$

7. Convex set

Voronoi description of halfspace. Let a and b be distinct points in $\mathbf{R}^n$. Show that the set of all points that are closer (in Euclidean norm) to a than b, i.e., $\{x \mid \|x-a\|_2 \leq \|x-b\|_2\}$, is a halfspace. Describe it explicitly as an inequality of the form $c^T x \leq d$. Draw a picture.

Since the norm is always nonnegative, we have $\|x-a\|_2 \leq \|x-b\|_2$ 当且仅当 $\|x-a\|_2^2 \leq \|x-b\|_2^2$

$$\|x-a\|_2^2 \leq \|x-b\|_2^2 \iff (x-a)^T(x-a) \leq (x-b)^T(x-b)$$
$$\iff x^Tx - 2a^Tx + a^Ta \leq x^Tx - 2b^Tx + b^Tx$$
$$\iff 2(b-a)^Tx \leq b^Tb - a^Ta$$

8. Convex set

(d) The set of points closer to a given point than a given set, *i.e.*,

$$\{x \mid \|x - x_0\|_2 \leq \|x - y\|_2 \text{ for all } y \in S\}$$

where $S \subseteq \mathbf{R}^n$.

Convex. Set $= \bigcap_{S} \{x \mid \|x-x_0\|_2 \leq \|x-y\|_2\}$

9. Convex set

(f) [HUL93, volume 1, page 93] The set $\{x \mid x + S_2 \subseteq S_1\}$, where $S_1, S_2 \subseteq \mathbf{R}^n$ with S_1 convex.

Convex. $x + S_2 \subseteq S_1$ if $x + y \in S_1$ for all $y \in S_2$

$\{x \mid x + S_2 \subseteq S_1\} = \bigcap_{y \in S_2} \{x \mid x + y \in S_1\}$

10. Convex set

(g) The set of points whose distance to a does not exceed a fixed fraction θ of the distance to b, *i.e.*, the set $\{x \mid \|x-a\|_2 \leq \theta \|x-b\|_2\}$. You can assume $a \neq b$ and $0 \leq \theta \leq 1$.

$\{x \mid \|x-a\|_2 \leq \theta \|x-b\|_2\}$

$= \{x \mid \|x-a\|_2^2 \leq \theta^2 \|x-b\|_2^2\}$

$= \{x \mid (1-\theta^2) x^Tx - 2(a - \theta^2 b)^T x + a^Ta - \theta^2 b^Tb \leq 0\}$

If $\theta = 1$, this is a halfspace

If $\theta < 1$, it is a ball $\{x \mid (x - x_0)^T (x - x_0) \leq R^2\}$

with center x_0, radius R

$$x_0 = \frac{a - \theta^2 b}{1 - \theta^2} \qquad R = \left(\frac{\theta^2 \|b\|_2^2 - \|a\|_2^2}{1-\theta^2} - \|x_0\|_2^2 \right)^{\frac{1}{2}}$$

11. Convex set

Some sets of probability distributions. Let x be a real-valued random variable with $\mathbf{prob}(x = a_i) = p_i$, $i = 1, \ldots, n$, where $a_1 < a_2 < \cdots < a_n$. Of course $p \in \mathbf{R}^n$ lies in the standard probability simplex $P = \{p \mid \mathbf{1}^T p = 1,\ p \geq 0\}$. Which of the following conditions are convex in p? (That is, for which of the following conditions is the set of $p \in P$ that satisfy the condition convex?)

(a) $\alpha \leq \mathbf{E} f(x) \leq \beta$, where $\mathbf{E} f(x)$ is the expected value of $f(x)$, *i.e.*, $\mathbf{E} f(x) = \sum_{i=1}^n p_i f(a_i)$. (The function $f : \mathbf{R} \to \mathbf{R}$ is given.)

(b) $\mathbf{prob}(x > \alpha) \leq \beta$.

(c) $\mathbf{E}|x^3| \leq \alpha \mathbf{E}|x|$.

(d) $\mathbf{E} x^2 \leq \alpha$.

(e) $\mathbf{E} x^2 \geq \alpha$.

(f) $\mathbf{var}(x) \leq \alpha$, where $\mathbf{var}(x) = \mathbf{E}(x - \mathbf{E} x)^2$ is the variance of x.

(g) $\mathbf{var}(x) \geq \alpha$.

(h) $\mathbf{quartile}(x) \geq \alpha$, where $\mathbf{quartile}(x) = \inf\{\beta \mid \mathbf{prob}(x \leq \beta) \geq 0.25\}$.

(i) $\mathbf{quartile}(x) \leq \alpha$.

(a) 化为 $\alpha \leq \sum_{i=1}^n p_i f(a_i) \leq \beta$ two linear inequalities，为凸

(b) 化为 $\sum_{a_i > \alpha} p_i \leq \beta$ a linear inequality，为凸

(c) 化为 $\sum_{i=1}^n p_i (|a_i^3| - \alpha |a_i|) \leq 0$ a linear inequality，为凸

(d) 化为 $\sum_{i=1}^n p_i a_i^2 \leq \alpha$ a linear inequality，为凸

(e) 化为 $\sum_{i=1}^n p_i a_i^2 \geq \alpha$ a linear inequality，为凸

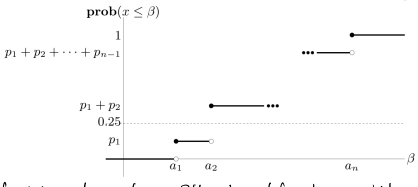
(f) 化为 $\sum_{i=1}^n p_i a_i^2 - (\sum_{i=1}^n p_i a_i)^2 \leq \alpha$，非凸，反例：

取 $n=2$，$a_1=0$，$a_2=1$，$\alpha = \frac{1}{5}$，$p_1 = (1,0)$ 与 $p_2 = (0,1)$ 满足，但 $p = (\frac{1}{2}, \frac{1}{2})$ 不满足

(g) 化为 $\sum_{i=1}^n a_i^2 p_i + (\sum_{i=1}^n a_i p_i)^2 = b^T p + p^T A p \leq \alpha$

$b_i = a_i^2$，$A = aa^T$. 为凸

(h) Let us denote $\mathbf{quartile}(x) = f(p)$ to emphasize it is a function of p. The figure illustrates the definition. It shows the cumulative distribution for a distribution p with $f(p) = a_2$.



if $\alpha \leq a_1$ always true. Otherwise, define $k = \max\{i \mid a_i < \alpha\}$.

$f(p) \geq \alpha$ holds if and only if $prob(x \leq a_k) = \sum_{i=1}^k p_i < 0.25$ 为凸

(i) 化为 $prob(x \leq \beta) \geq 0.25$ for all $\beta \geq \alpha$ 为凸

12. Convex set

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

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

Solution. We consider two points $(\bar{x}, \bar{y}_1 + \bar{y}_2), (\tilde{x}, \tilde{y}_1 + \tilde{y}_2) \in S$, *i.e.*, with

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

For $0 \leq \theta \leq 1$,

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

is in S because, by convexity of S_1 and S_2,

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

13. Convexity

Prove the following statements:

1. $\frac{1}{n}\sum_{i=1}^n x_i \geq (\prod_{i=1}^n x_i)^{1/n}$
2. $f(\mathbf{x}) = (\prod_{i=1}^n x_i)^{1/n}$ is concave where $x_i > 0$ for all i.

1. As $\log(t)$ is concave, then $\log(\frac{1}{n}\sum_{i=1}^n x_i) \geq \frac{1}{n}\sum_{i=1}^n \log x_i = \frac{1}{n}\log\prod_{i=1}^n x_i$

因此，$\frac{1}{n}\sum_{i=1}^n x_i \geq (\prod_{i=1}^n x_i)^{\frac{1}{n}}$

2. $\frac{f(x)}{f(\lambda x + (1-\lambda) y)} = (\prod_{i=1}^n \frac{x_i}{\lambda x_i + (1-\lambda) y_i})^{\frac{1}{n}} \leq \frac{1}{n}\sum_{i=1}^n \frac{x_i}{\lambda x_i + (1-\lambda) y_i}$

Similarly $\frac{f(y)}{f(\lambda x + (1-\lambda) y)} \leq \frac{1}{n}\sum_{i=1}^n \frac{y_i}{\lambda x_i + (1-\lambda) y_i}$

Now $\lambda \frac{f(x)}{f(\lambda x + (1-\lambda) y)} + (1-\lambda)\frac{f(y)}{f(\lambda x + (1-\lambda) y)} \leq \frac{1}{n}\sum_{i=1}^n \frac{\lambda x_i + (1-\lambda) y_i}{\lambda x_i + (1-\lambda) y_i} = 1$

As $f(\lambda x + (1-\lambda) y) > 0$,

$\lambda f(x) + (1-\lambda) f(y) \leq f(\lambda x + (1-\lambda) y)$

14. Convexity

证 $f(x,y) = (x - x_0)^2 + (y - y_0)^2$ convex

Let $g(\theta) = (x + \theta e_1 - x_0)^2 + (y + \theta e_2 - y_0)^2$, Then

$g'(\theta) = 2(x + \theta e_1 - x_0) e_1 + 2(y + \theta e_2 - y_0) e_2$

$g''(\theta) = 2e_1^2 + 2e_2^2 = 2 > 0$

where e_i denotes the ith element of the unit vector e.

15. Convexity

Show that $f(x_1, x_2) = (x_1^p + x_2^p)^{1/p}$ for $x_1, x_2 > 0$, $p < 1$, $p \neq 0$ is concave.

Let $g(\theta) = f(x_1 + \theta e_1, x_2 + \theta e_2) = [(x_1 + \theta e_1)^p + (x_2 + \theta e_2)^p]^{1/p}$. Then we have

$$\begin{aligned} g'(\theta) &= [(x_1 + \theta e_1)^p + (x_2 + \theta e_2)^p]^{(1-p)/p} \left[(x_1 + \theta e_1)^{p-1} e_1 + (x_2 + \theta e_2)^{p-1} e_2\right] \\ &= g(\theta)^{1-p} \left[(x_1 + \theta e_1)^{p-1} e_1 + (x_2 + \theta e_2)^{p-1} e_2\right], \\ g''(\theta) &= (1-p) g(\theta)^{-p} g'(\theta) \left[(x_1 + \theta e_1)^{p-1} e_1 + (x_2 + \theta e_2)^{p-1} e_2\right] \\ &\quad - (1-p) g(\theta)^{1-p} \left[(x_1 + \theta e_1)^{p-2} e_1^2 + (x_2 + \theta e_2)^{p-2} e_2^2\right] \\ &= (1-p) g(\theta)^{1-2p} \left[(x_1 + \theta e_1)^{p-1} e_1 + (x_2 + \theta e_2)^{p-1} e_2\right]^2 \\ &\quad - (1-p) g(\theta)^{1-p} \left[(x_1 + \theta e_1)^{p-2} e_1^2 + (x_2 + \theta e_2)^{p-2} e_2^2\right] \\ &= \frac{(1-p)}{g(\theta)} \left[\left(\sum_{i=1}^2 g(\theta)^{1-p} (x_i + \theta e_i)^{p-1} e_i \right)^2 - \sum_{i=1}^2 g(\theta)^{2-p} (x_i + \theta e_i)^{p-2} e_i^2 \right] \\ &\leq 0, \end{aligned}$$

where we apply the Cauchy-Schwarz inequality $(\sum_{i=1}^n a_i b_i)^2 \leq (\sum_{i=1}^n a_i^2)(\sum_{j=1}^n b_j^2)$ with

$$a_i = g(\theta)^{-p/2} (x_i + \theta e_i)^{p/2},\ b_i = g(\theta)^{1-p/2} (x_i + \theta e_i)^{p/2-1} e_i.$$

Note that $\sum_{i=1}^2 a_i^2 = \sum_{i=1}^2 (x_i + \theta e_i)^p / [(x_1 + \theta e_1)^p + (x_2 + \theta e_2)^p] = 1$.

16. Convexity

For $f(x) = ax^3 + bx^2 + cx + d$ with $x \in \mathbb{R}$, find ranges of a, b, c, d to make $f(x)$ quasi-concave.

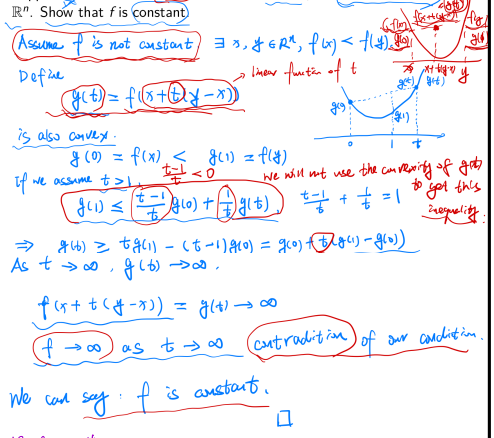
The first-order derivative of $f(x)$ with respect to x is $f'(x) = 3ax^2 + 2bx + c$.

- **Case 1**: $f'(x) \geq 0$ or $f'(x) \leq 0$ for all $x \in \mathbb{R}$. In this case, $f(x)$ is always monotone and thus quasi-concave.
  - If $a = 0$, then $b = 0$.
  - If $a \neq 0$, then $\Delta = 4b^2 - 12ac = 4(b^2 - 3ac) \leq 0$, i.e., $b^2 - 3ac \leq 0$.
- **Case 2**: $f'(x)$ will change its sign for $x \in \mathbb{R}$.
  - If $a = 0$, then only when $b < 0$, $f(x)$ first increases up to x^* and then decreases, and thus $f(x)$ is quasi-concave.
  - If $a \neq 0$ and $\Delta > 0$, i.e., $b^2 - 3ac > 0$, it is impossible that $f(x)$ is quasi-concave.

In summary, $f(x)$ is quasi-concave if and only if $a \neq 0, b^2 - 3ac \leq 0$ or $a = 0, b \leq 0$.

17. Convexity

Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is convex with $\operatorname{dom} f = \mathbb{R}^n$, and bounded above on $\mathbb{R}^n$. Show that f is constant.

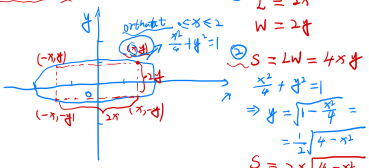


18. Convexity

A rectangle is to be inscribed in the ellipse

$$\frac{x^2}{4} + y^2 = 1$$

What should the dimensions of the rectangle be to maximize its area?
What is the maximum area?



① $L = 2x$, $W = 2y$

② $S = LW = 4xy$

$\frac{x^2}{4} + y^2 = 1$

$\Rightarrow y = \sqrt{1 - \frac{x^2}{4}} = \frac{1}{2}\sqrt{4 - x^2}$

$S = 2x\sqrt{4 - x^2}$

② $S'(x) = 2\sqrt{4 - x^2} + 2x \cdot \frac{1}{2\sqrt{4 - x^2}} \cdot (-2x)$

$= \frac{8 - 4x^2}{\sqrt{4 - x^2}} = 0$

$\Rightarrow x = \sqrt{2}$

Claim: For a continuous $S(x) = 2x\sqrt{4 - x^2}$ with the closed interval $[0, 2]$, then $S(x)$ can only attain its max or min at $x = 0$, $x = 2$, or $x = x^*$ ($S'(x^*) = 0$)

$S(0) = 0$, $S(2) = 0$, $S(x^*) = S(\sqrt{2}) = 2\sqrt{2} \cdot \sqrt{2} = 4$.

In conclusion,

$L = 2x = 2\sqrt{2}$ $W = 2y = \sqrt{2}$

$\operatorname{Max}(S) = 4$.

19. ML

We apply the gradient descent method to minimize the function $f(x) = (x-1)^2$.

1. If the update is stuck in a loop, that is, $f(x^{(t+1)}) = f(x^{(t)})$ yet $|x^{(t+1)} - x^{(t)}| > \varepsilon$, what can we say about the learning rate/step size $\alpha^{(t)}$?
2. If the function $f(x)$ keeps descending as the update goes on, what can we say about the learning rate/step size $\alpha^{(t)}$?

1. The updating scheme is

$$x^{(t+1)} = x^{(t)} - 2\alpha^{(t)}(x^{(t)} - 1).$$

It follows that

$$f(x^{(t+1)}) = f(x^{(t)}),$$

that is

$$\left(x^{(t)} - 2\alpha^{(t)}(x^{(t)} - 1) - 1\right)^2 = (x^{(t)} - 1)^2$$

which yields

$$\alpha^{(t)}(\alpha^{(t)} - 1)(x^{(t)} - 1)^2 = 0.$$

Since the update is stuck and does not converge, $x^{(t)} \neq 1$, we have $\alpha^{(t)} = 1$.

2. If follows that

$$f(x^{(t+1)}) < f(x^{(t)}),$$

that is

$$\left(x^{(t)} - 2\alpha^{(t)}(x^{(t)} - 1) - 1\right)^2 < (x^{(t)} - 1)^2$$

which yields $\alpha^{(t)} < 1$.