§2. 方程求解与矩阵.

0-范数：向量中非零元素个数

范数 度量性质

衡量矢量函数的长度或大小

1. 向量范数.

- $1-$范数: $\|x\|_1=\sum_{i=1}^{N}|x_i|$ 绝对值之和
- $2-$范数: $\|x\|_2=\left(\sum_{i=1}^{N}|x_i|^2\right)^{\frac{1}{2}}$, Euclidean Norm,欧几里德范数 → 向量的模长.
- $p-$范数:$\|x\|_p=\left(\sum_{i=1}^{N}|x_i|^p\right)^{1/p}$
- $\infty-$范数: $\|x\|_\infty=max_i|x_i|$
- $-\infty-$范数: $\|x\|_{-\infty}=min_i|x_i|$

2. 矩阵范数

矩阵的迹：矩阵A的主对角线上各元素的总和.

顺序无关

$tr(AB)=tr(BA)=\sum_{i=1}^{n}a_{ij}b_{ji}$

区分行列

- $1-$范数:$\|A\|_1=max_j\sum_{i=1}^{m}|a_{i,j}|$,列和范数 最大的列和.
- $2-$范数:$\|A\|_2=\sqrt{\lambda_1},\lambda_1$是$A^TA$的最大特征值，也称为谱范数
- $\infty-$范数:$\|A\|_\infty=max_i\sum_{j=1}^{n}|a_{i,j}|$,行和范数
- $F-$范数:$\|A\|_F=\left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{i,j}|^2\right)^{\frac{1}{2}}$,Frobenius范数 → 衡量矩阵的大小.

柯西不等式

$\sum_{k=1}^{n}(a_kb_k)^2\le\sum_{k=1}^{n}a_k^2\sum_{k=1}^{n}b_k^2$

① $\|A\|_F^2=\sum_{ij}|a_{ij}|^2=\sum_i\|A_{i*}\|_2^2=\sum_j\|A_{*j}\|_2^2=trace(A^TA)=trace(AA^T)=Energy(A)$

② 三角不等式：$||A+B||_F\le||A||_F+||B||_F$ (A.B形状相同)

证明按定义展开.

$\sqrt{\sum(x+y)^2}=\sqrt{\sum(x^2+y^2+2xy)}\le\sqrt{\sum(x^2+y^2)+2\sqrt{\sum x^2\sum y^2}}=\sqrt{\sum x^2}+\sqrt{\sum y^2}$.(柯西不等式)

③ $||AB||_F\le||A||_F||B||_F$

证明：令A的行为$\bar{a}_i$,$\bar{b}_j$为B的转置的列，则AB的第(i,j)处的元素为$\bar{a}_i\cdot\bar{b}_j$,矩阵AB的平方Frobenius范数为$\sum_{i=1}^{n}\sum_{j=1}^{d}(\bar{a}_i\cdot\bar{b}_j)^2$,而$(\bar{a}_i\cdot\bar{b}_j)^2\le||\bar{a}_i||^2||\bar{b}_j||^2$,因此

$||AB||_F^2=\sum_{i=1}^{n}\sum_{j=1}^{d}(\bar{a}_i\cdot\bar{b}_j)^2\le\sum_{i=1}^{n}\sum_{j=1}^{d}||\bar{a}_i||^2||\bar{b}_j||^2\le(\sum_{i=1}^{n}||\bar{a}_i||^2)(\sum_{j=1}^{d}||\bar{b}_j||^2)\le||A||_F^2||B||_F^2$

④ 一个$n\times n$的F范数为ϵ的矩阵的逆矩阵的F范数至少为$\sqrt{n}/\epsilon$。

证明：由③. $||\underbrace{AA^{-1}}_{E_n}||_F\le||A||_F||A^{-1}||_F\Rightarrow||A^{-1}||_F\ge\frac{||AA^{-1}||_F}{||A||_F}=\frac{\sqrt{n}}{\epsilon}$

3. 诱导矩阵范数.

(将矩阵转化到向量上)

max & min 选择满足

$\|Ax\|\le\|A\|\|x\|$,

eg. $\|A\|=max_{\|x\|=1}\|Ax\|,A\in C^{m\times n},x\in C^{n\times 1}$ $\quad min_{\|x\|=1}\|Ax\|=\frac{1}{\|A^{-1}\|}$

$\|A\|_2=max_{\|x\|_2=1}\|Ax\|_2=\sqrt{\lambda_{max}}$ $\quad \|A^{-1}\|_2=\frac{1}{min_{\|x\|_2=1}\|Ax\|_2}=\frac{1}{\sqrt{\lambda_{min}}}$

证明：$maximize\ f(x)=\|Ax\|_2^2=x^TA^TAx$, $s.t.\ g(x)=x^Tx=1$,采用Lagrange乘子法，引入新的变量λ,考虑函数$h(x,\lambda)=f(x)-\lambda g(x)$,从而得$(A^TA-\lambda I)x=0$,表明函数$f$在满足上述方程的向量$x$处取得最大值，并且$\|x\|_2=1$.从而$\lambda$必须是使得$A^TA-\lambda I$奇异的值。

又由于$x^TA^TAx=\lambda x^Tx=\lambda$,可得$\|A\|_2=max_{\|x\|=1}\|Ax\|=max_{\|x\|^2=1}\|Ax\|=(\sum_{x^Tx=1}x^TA^TA\,x)^{1/2}=\sqrt{\lambda_{max}}$

§2. 方程求解与矩阵

4. 线性空间与子空间

向量 ⟶ 物体，矩阵 ⟶ 运动。

k个向量$a_i \in R^m$张成的空间：k个向量的任意线性组合形成空间，

$$span\{a_1, \cdots, a_k\} = \{y \in R^m | y = x_1 a_1 + x_2 a_2 + \cdots + x_k a_k, x_i \in R\}$$

矩阵$A \in R^{m \times n}$的列空间：由矩阵A的列向量所张成的空间

$$R(A) = span\{a_1, a_2, \cdots, a_n\} = \{y \in R^m | y = Ax, x \in R^n\}$$

- $m \le n$. A行满秩 ⇒ Ax可构成R^m中任一向量。所有线性系统$Ax=b$是兼容的。($b \in R(A)$)
  - ↑ 线性相关
- $m \ge n$. A列满秩 ⇒ Ax无法表示0向量，即A零空间为$N(A)=\{0\}$（满足$Ax=0$的x）
  - ↑ 线性无关

行满秩

$$\begin{bmatrix} 1 & & & & 0 & 0 \\ & 1 & & & 0 & 0 \\ & & 1 & & \vdots & \vdots \\ & & & \ddots & & \\ & & & 1 & 0 & 0 \\ & & & & 1 \; 0 & 0 \end{bmatrix}$$

列满秩

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1 \\ & & & & & 1 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

基本矩阵列操作：相加、交换、伸缩。（矩阵相乘实现）

5. 拟合（实验一）

① 思想：多项式：$y(x,w) = w_0 + w_1 x + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$

最小二乘 ⇒ 最优化问题，$\min E(w), E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2$

② 评价：RMS均方差。

③ 正则项：对权系数进行限定，对近似模型进行约束

④ 优化算法：梯度下降.（分块/批…）⟶ SGD

拓展：移动最小二乘$f_{mls}(w_x) = \sum_{i=1}^{k} \theta(x, x_i)||\phi(x_i) - y_i||^2$

计算某个节点附近的拟合曲线（局部性）

方阵A, $\det(A) \neq 0$的等价表述：

① A可逆（A^{-1}存在，$AA^{-1} = A^{-1}A = I$）

② $Ax=b$有唯一解。

③ $Ax=0$的解是$x=0$

6. $Ax=b$求解*

① 高斯消元法（计算题）　PPT例题 P2-k2.

$A\bar{x} = \bar{b} \Rightarrow A'\bar{x} = \bar{b}', A'$是上三角矩阵　$\underbrace{E_k E_{k-1} \ldots E_1 A}_{A'} \bar{x} = \underbrace{E_k E_{k-1} \ldots E_1 \bar{b}}_{\bar{b}'}$

- 按照前面的介绍，基本的高斯消元法，将矩阵A用左乘基本矩阵后将矩阵A转化为上三角矩阵

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2n}x_n = b_2 \\ \vdots \qquad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nn}x_n = b_n \end{array} \Longrightarrow \begin{array}{r} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n = b_1 \\ a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n = b'_2 \\ a''_{33}x_3 + \ldots + a''_{3n}x_n = b''_3 \\ \vdots \qquad \vdots \\ a_{nn}^{(n-1)}x_n = b_n^{(n-1)} \end{array}$$

① 前向消元：逐渐将第i到n行中的所有$x_j, j < i$都消掉，这时矩阵A成为上三角矩阵

② 后向替换：先求出x_n，代入第$n-1$个方程，求出x_{n-1}，最后求出x_1

缺陷：可能主元过小　　主元法

改进：在第k步前向消除时，选择$|a_{kk}|, |a_{k+1,k}|, \ldots, |a_{nk}|$中最大的那个方程，与第$k$个进行交换

非齐次线性方程组的解

若$rank(A) < rank(A\ b)$，不相容的系统

若$rank(A) = rank(A\ b)$，相容系统

- 若$rank(A)$ =未知变量的个数，唯一解
- 若$rank(A)$ <未知变量的个数，无穷多解

线性无关.　　增广矩阵

关于行列式的求值.

- 定理：方阵A，如果矩阵B是由A的行或列增加或减去一行或一列的倍数得到的矩阵，则$\det(A) = \det(B)$
- 定理：方阵A是上三角、下三角或对角矩阵，则$\det(A) = a_{11} \times a_{22} \times \cdots \times a_{nn}$
- 定理：如果矩阵B是交换了A的一行或一列后得到的矩阵，则$\det(B) = -\det(A)$

§2. 方程求解与矩阵

7. 矩阵分解 $A=LU$. 其中L是下三角矩阵，U是上三角矩阵

类比高斯消元的过程.

解$Ax=b$,则$LUx=b \Rightarrow Ux=L^{-1}b$

令$L^{-1}b=y$,则得$Ly=b, Ux=y$

注意：$Ly=b$用前向替换解得y；然后$Ux=y$用后向替换解得x

具体求解：

① $\begin{cases} L_{(n)}\cdots L_{(1)}A=U & \text{通过初等行变换将}A\text{变为上三角矩阵.} \\ L_{(1)}^{-1}\cdot L_{(2)}^{-1}\cdots L_{(n)}^{-1}=L. \end{cases}$

LU 分解的好处在于：①节省存储空间。下三角阵主对角元素均为 1，故可省略。所以，可将一个矩阵分解为两个同等大小的矩阵，且不会引入新的存储需求。②若将 n 阶方阵 $\boldsymbol{A}$ 视作线性变换，则给定任意变换后的 n 维向量 y，求解变换前的 n 维向量 x 时，关键变换信息已存储在 $\boldsymbol{L}$ 矩阵中，无须重新计算。

8. 矩阵的稳定性

$Ax=b$稳定性：

病态的：右端项稍有扰动，解会发生很大的变化.

设近似解$\hat{x}$，残差 $r=b-A\hat{x}=Ax-A\hat{x}=Ae$ 对解的好坏指示作用差.

绝对误差

$$||e||=||x-\hat{x}||=||A^{-1}b-A^{-1}\hat{b}||=||A^{-1}(b-\hat{b})||\le||A^{-1}||\cdot||b-\hat{b}||=||A^{-1}||\cdot||r||$$

相对误差

$$||e||=||A^{-1}||\cdot||r||\cdot\frac{||Ax||}{||b||}\le||A^{-1}||\cdot||A||\cdot||x||\cdot\frac{||r||}{||b||} \Rightarrow \frac{||e||}{||x||}\le \underline{||A^{-1}||\cdot||A||}\cdot\frac{||r||}{||b||}$$

条件数：$\kappa(A)=||A^{-1}||\cdot||A||$

进一步细化.

- 假设$Ax=b$,且$\hat{A}\hat{x}=\hat{b}$，这里$\hat{A}=A+\delta A$，$\hat{b}=b+\delta b$，$\hat{x}=x+\delta x$,则有：
  $$\delta Ax+A\delta x+\delta A\delta x=\delta b$$
- 假设δ项都比较小，则有如下线性逼近
  $$\delta Ax+A\delta x\approx\delta b$$
- 从而求得$\delta x\approx A^{-1}(\delta b-\delta Ax)$
- 取范数：$||\delta x||\lesssim||A^{-1}||(||\delta b||+||\delta A||||x||)$
- 除$||x||$得关于x的相对误差：
- $\frac{||\delta x||}{||x||}\lesssim||A^{-1}||\,||A||\left(\frac{||\delta b||}{||A||||x||}+\frac{||\delta A||}{||A||}\right)\lesssim\kappa(A)\left(\frac{||\delta b||}{||b||}+\frac{||\delta A||}{||A||}\right)$
- 表明：关于解x的相对误差是以A和b的相对误差的条件数倍数为界！

迭代优化

- 1.得到一个近似解$A\hat{x}_1\approx b$
- 2.计算残差$r=b-A\hat{x}_1$(到某个足够好的精度)
- 3.近似求解$A\delta x_1\approx r$
- 4.得到一个新的近似解$\hat{x}_2=\hat{x}_1+\delta x_1$；重复上述过程！

考虑$\hat{A}=A+E$ →

(1)：$Ax=\hat{A}x-Ex=b$,假设x是一个迭代的不动点

(2)：$\hat{A}x_{k+1}-Ex_k=b \Longrightarrow \hat{A}x_{k+1}-(\hat{A}-A)x_k=b$

$\Longrightarrow x_{k+1}=x_k+\hat{A}^{-1}(b-Ax_k)$

残差估计为：

$||\hat{x}-x||=||A^{-1}||||r||$ →

条件数的计算

Hager算法.

①$||\cdot||_1$几乎处处可微. 若y无非0元素, 则 $\xi^Ty=||y||_1, \xi=sign(y)$

②若δy足够小不改变y符号.

$\xi^T(y+\delta y)=||y+\delta y||_1$

一般地, $\xi^Tu\le||\xi||_\infty||u||_1=||u||_1$ → 下界.

③对于$y=A^{-1}x$.

$|\xi^TA^{-1}(x+\delta x)|\le||A^{-1}(x+\delta x)||$

当且仅当δx充分小...

④令$x=x_{new}$, $||x_{new}||\le1$.

最大化$|\xi^TA^{-1}x_{new}|$,

实际上$x_{new}=e_j$, j选为使得$z^T=\xi^TA^{-1}$的第j个元素有最大幅度

- 这实际上来自于$Ax_k-b=0$的近似雅各比矩阵$\hat{A}$的非精确牛顿迭代
  - 将 **(2)** 式减去 **(1)** 式：$\hat{A}(x_{k+1}-x)-E(x_k-x)=0$,
  - 取范数$||x_{k+1}-x||\le||\hat{A}^{-1}E||||x_k-x||$ 收敛条件：$||\hat{A}^{-1}E||<1$，则当$k\to\infty, x_k\to x$
  - $x_{k+1}=x_k+\hat{A}^{-1}(b-Ax_k+\delta_k)+\mu_k$
  - 其中δ_k是与计算残差相关的误差，μ_k是与更新相关的误差。如果$||\delta_k||<\alpha, ||\mu_k||<\beta$对所有$k$都成立，且如果$||A^{-1}E||$距离1不是太近（例如，$||A^{-1}E||<\frac{1}{2}$），则有类似 **(3)**：$||x_k-x||\le||A^{-1}E||^k||x_0-x||+C(\alpha||A^{-1}||+\beta)$ 这里C是一个不太大的常数
  - 如果估计一下残差，通常有$\alpha\le c_1\epsilon_{mach}||A||||x||, \beta\le c_2\epsilon_{mach}||x||$，代入 **(3)** 得 $\frac{||x_k-x||}{||x||}\le C_1\epsilon_{mach}\kappa(A)+C_2\epsilon_{math}$

§2. 方程求解与矩阵

造成矩阵病态的原因：在对问题进行建模的时候，单位不一致

改写 → $D_1AD_2y = D_1b$，s.t. $\kappa(D_1AD_2) \ll \kappa(A)$ 其中D_1, D_2都是对角缩放矩阵。

范数同时改变

物理：无量纲处理

预条件子：通过变换求解一些条件数比较小的线性方程组来求原始问题

- **前向类目标$M \approx A$**
- Ⅰ（左）：$M^{-1}Ax = M^{-1}b$
- Ⅱ（右）：$AM^{-1}y = b, x = M^{-1}y$
- Ⅲ（混合）：$M_2^{-1}AM_1^{-1}y = M_2^{-1}b, x = M_1^{-1}y$
- **逆向类目标$M \approx A^{-1}$**:
- Ⅰ（左）：$MAx = Mb$
- Ⅱ（右）：$AMy = b, x = My$
- Ⅲ（混合）：$M_2AM_1y = M_2b, x = M_1y$

$MAx = Mb$，当 $M \approx A^{-1}$ 时才有意义

范数不改变，书式基本不改变

9. 矩阵的逆

① 矩阵存在逆矩阵，则逆矩阵唯一。如果$B_1 = B_2$,且$AB_1 = AB_2 = I$,则必有$B_1 = B_2$

证明：$AB_1 = AB_2 \Rightarrow A(B_1 - B_2) = 0 \Rightarrow B_1A(B_1 - B_2) = 0 \Rightarrow B_1 = B_2$

逆存在即唯一

② 两个矩阵的和的逆矩阵不容易表示为单个矩阵逆或多项式表示.

1). 两个矩阵可逆，其和未必可逆.

2) $\lim_{n\to\infty} A^n = 0$，则:

- $(I + A)^{-1} = I - A + A^2 - A^3 + \cdots + \cdots$
- $(I - A)^{-1} = I + A - A^2 + A^3 + \cdots + \cdots$

$\frac{1}{1+a} = 1 - a + a^2 - a^3 + a^4 + \cdots + \cdots$

类比泰勒展开.

③ 矩阵逆引理：令A是一个可逆的$d \times d$矩阵，$\bar{u}, \bar{v}$是非零的d维列向量。则$A + \bar{u}\bar{v}^T$是可逆的，当且仅当$\bar{v}^T A^{-1}\bar{u} \neq -1$。此时，其逆矩阵为：

$$(A + \bar{u}\bar{v}^T)^{-1} = A^{-1} - \frac{A^{-1}\bar{u}\bar{v}^T A^{-1}}{1 + \bar{v}^T A^{-1}\bar{u}}$$

证明：如果矩阵$(A + \bar{u}\bar{v}^T)$可逆，则其和A^{-1}的乘积也可逆。因此$(A + \bar{u}\bar{v}^T)\,A^{-1}\bar{u}$是一个非零向量。否则可得$\bar{u} = 0$,矛盾。因此有：

$(A + \bar{u}\bar{v}^T)\,A^{-1}\bar{u} \neq 0 \Rightarrow \bar{u} + \bar{u}\bar{v}^T A^{-1}\bar{u} \neq 0 \Rightarrow \bar{u}(1 + \bar{v}^T A^{-1}\bar{u}) \neq 0 \Rightarrow 1 + \bar{v}^T A^{-1}\bar{u} \neq 0$，因此前置可逆得证。

反向，如果$1 + \bar{v}^T A^{-1}\bar{u} \neq 0$成立，我们证明矩阵$P = A^{-1} - \frac{A^{-1}\bar{u}\bar{v}^T A^{-1}}{1+\bar{v}^T A^{-1}\bar{u}}$时，$Q = (A + \bar{u}\bar{v}^T)$。注意，矩阵$P$仅当预条件成立时才有效。展开$PQ$得：

$$\begin{aligned} PQ &= I + A^{-1}\bar{u}\bar{v}^T - \frac{A^{-1}\bar{u}\bar{v}^T + A^{-1}\bar{u}\,[\bar{v}^T A^{-1}\bar{u}]\,\bar{v}^T}{1 + \bar{v}^T A^{-1}\bar{u}} \\ &= I + A^{-1}\bar{u}\bar{v}^T - \frac{A^{-1}\bar{u}\bar{v}^T(1 + [\bar{v}^T A^{-1}\bar{u}])}{1 + \bar{v}^T A^{-1}\bar{u}} \\ &= I + A^{-1}\bar{u}\bar{v}^T - A^{-1}\bar{u}\bar{v}^T = I \end{aligned}$$

证明使用标量$\bar{v}^T A^{-1}\bar{u}$可以在矩阵乘法中移动！

⇒ 具有非零对角元素$L_{d\times d}$矩阵可表示为$(\Delta + A)$的形式，Δ表示可逆对角矩阵，A是一个严格三角矩阵。证明如何仅使用对角矩阵逆和乘法或加法来计算矩阵L的逆矩阵。（严格三角矩阵$A^d = 0$）

逆引理变种的应用：

- 例如在增量线性回归中，需要求逆矩阵：$C = D^T D$，D是$n \times d$的数据矩阵，当接收到新的d维数据点$\bar{v}$时，数据矩阵成为$(n+1) \times d$维，相当于矩阵D增加了一行，矩阵C更新为：$D^T D + \bar{v}\bar{v}^T$，此时利用上述引理，用$O(d^2)$次更新逆矩阵！
- 上述结果可以推广到向量$\bar{u}, \bar{v}$都用扁矩阵U, V替换（U, V只包含k列，k较小）

Sherman-Morrison恒等式

& Sherman-Morrison-Woodbury恒等式

§2. 方程求解与矩阵

⑥ Sherman-Morrison恒等式.

- 用来描述当A的分解已知时，求解$(A+UV^T)x=b$。
- 可通过分块高斯消元法来推导该公式:
- 为了计算$(A+UV^T)x$，首先计算$\xi=V^Tx$，然后计算$(A+UV^T)x=Ax+U\xi$，因此可将$(A+UV^T)x=b$写为如下的形式:
- $\begin{bmatrix}A & U\\ V^T & -1\end{bmatrix}\begin{bmatrix}x\\ \xi\end{bmatrix}=\begin{bmatrix}b\\ 0\end{bmatrix}$
- 将其做因子分解$\begin{bmatrix}A & U\\ V^T & -1\end{bmatrix}=\begin{bmatrix}I & 0\\ V^TA^{-1} & 1\end{bmatrix}\begin{bmatrix}A & U\\ 0 & -1-V^TA^{-1}U\end{bmatrix}$,应用块下三角因子做前向替代:
- $y=b,\eta=-V^TA^{-1}y$，然后块状上三角因子做后向替代:
- $\xi=(-1-V^TA^{-1}U)^{-1}\eta, x=A^{-1}(y-U\xi)$,结合上述过程得:
- $x=\left[A^{-1}-\frac{A^{-1}UV^TA^{-1}}{1+V^TA^{-1}U}\right]b$，等价于
- $(A+UV^T)^{-1}=\left[A^{-1}-\frac{A^{-1}UV^TA^{-1}}{1+V^TA^{-1}U}\right]$,**将其推广到秩为$k$的矩阵$A$上** $(\overset{d\times d}{A}+\underbrace{U}_{d\times k}\underbrace{V^T}_{k\times d})^{-1}=A^{-1}-A^{-1}U(I+V^TA^{-1}U)^{-1}V^TA^{-1}$
- $(I+UV^T)^{-1}=I-U(I+V^TU)^{-1}V^T$ (A=I)

- 这可以看出恒等式可用来化简求解$n\times n$矩阵求逆为求解一个$k\times k$矩阵的逆再加上一些矩阵向量乘法。如果$k\ll n$，则计算量从$O(n^3)$化为$O(k^3+kn)$,显著减少!

- 这种类型的更新称为低秩更新

$P=PI_k$.

问题：假设I,P都是$k\times k$的矩阵，证明：$(I+P)^{-1}=I-(I+P)^{-1}P$

问题：如果U,V都是$n\times d$的矩阵，证明：$U^T(I_n+VU^T)^{-1}=(I_d+U^TV)^{-1}U^T$,使用该结论，证明：对任意$n\times d$矩阵$D$和标量$\lambda>0$，有$D^T(\lambda I_n+DD^T)^{-1}=(\lambda I_d+D^TD)^{-1}D^T$

① $U^T(I_n+VU^T)^{-1}=U^T(I_n-V(I_d+U^TV)^{-1}U^T)$

$(I_d+U^TV)(I_d+U^TV)^{-1}=I_d$

$\Leftrightarrow\quad U^TV(I_d-U^TV)^{-1}=I_d-(I_d+U^TV)^{-1}$

$\therefore$ 左$=U^T-[I_d-(I_d+U^TV)^{-1}]U^T$

$=(I_d+U^TV)^{-1}U^T$.

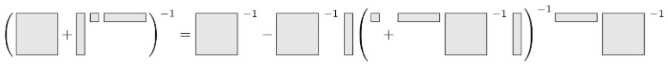
② 令$\lambda U=D$. 代换回原式即可.

⇒ Sherman-Morrison-Woodbury恒等式.

- 扩展到更一般的形式，逆由下式给出:

$$(A+USV^T)^{-1}=A^{-1}-A^{-1}U(S^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1}$$

- 证明：令$B=A+USV^T$,则$B^{-1}=(I+A^{-1}USV^T)^{-1}A^{-1}$.设$W=A^{-1}U, Z^T=SV^T$使得
- $B^{-1}=(I+WZ^T)^{-1}A^{-1}$ $\qquad (I+UV^T)^{-1}=I-U(I+V^TU)^{-1}V^T$
- 则有$B^{-1}=(I-W(I+Z^TW)^{-1}Z^T)A^{-1}=(I-A^{-1}U(I+SV^TA^{-1}U)^{-1}SV^T)A^{-1}=A^{-1}-A^{-1}U(S^{-1}+V^TA^{-1}U)^{-1}V^TA^{-1}$



- 如果已知矩阵A^{-1}，则可加速其计算，这是**BFGS**算法的关键！将计算量从$O(n^3)$转化为$O(k^3+k(n^2+n))$

10. 线性映射.

线性映射是一个函数$T:V\to W,V,W$都是向量空间，满足如下的齐次和可加性　　线性映射如果$W=V$，称之为线性算子

- $T(x+y)=Tx+Ty,\forall x,y\in V$
- $T(\alpha x)=\alpha Tx,\forall x\in V,\alpha\in R$

同态 $\xrightarrow{\text{可逆}}$ 同构.

- 如果V,W是实向量空间且$dimV=dimW=n$，则有天然的同构:
- $\phi:V\to W,\alpha_1v_1+\cdots+\alpha_nv_n\mapsto\alpha_1w_1+\cdots+\alpha_nw_n$,这里$v_1,\cdots,v_n$和$w_1,\cdots,w_n$是$V$和$W$的任意一组基。每个$n$维的向量空间都同构于$R^n$。

- 定义线性映射$T:V\to W$,则映射T的零空间（Nullspace）
  - $N(T)=\{v\in V|Tv=0\}$
- 定义其值域空间(Range Space，Columnspace)表示为
  - $R(T)=\{w\in W|\exists v\in V,Tv=w\}$
- 集合S上的度量是一个函数$d:S\times S\to R$,满足:
  - （1）非负性，$d(x,y)\geq 0$,等号当且仅当$x=y$时成立
  - （2）对称性，$d(x,y)=d(y,x)$
  - （3）三角不等式：$d(x,z)\leq d(x,y)+d(y,z),\forall x,y,z\in S$
- 具有度量的空间就是度量空间。一个度量空间中的所有柯西序列都收敛在该空间中的一点，则称该空间为完备空间。

在数学中，柯西序列、柯西列、柯西数列或基本列是指这样一个数列，它的元素随着序数的增加而愈发靠近。任何收敛数列必然是柯西列，任何柯西列必然是有界序列。
度量空间中的元素序列$\{x_n\}$，使得对于任意$\epsilon>0$，序列中存在自然数N，使得对任意$m,n>N$，都有$|x_m-x_n|<\epsilon$

- 范数满足：非负，齐次，和三角不等式，向量空间赋予范数称为赋范空间，完备的赋范空间称为巴拿赫空间
- 内积$<\cdot,\cdot>:V\times V\to R$满足
  - 非负性：$<x,x>\geq 0$,等号当且仅当$x=0$时成立
  - 线性性：$<x+y,z>=<x,z>+<y,z>,<\alpha x,y>\alpha<x,y>$
  - 自反性：$<x,y>=<y,x>$

赋予了内积的向量空间就是内积空间

- 勾股定理
  - 如果$x\perp y$，则$||x+y||^2=||x||^2+||y||^2=<x+y,x+y>$

内积空间的性质：

- 从内积和诱导出范数，$||x||=\sqrt{<x,x>}$
- 显然，内积空间肯定也是赋范空间
- 若$<x,y>=0$,称两个向量正交，可写为$x\perp y$,正交推广了欧几里德空间里面的垂直概念，如果两个正交向量都是单位长度，$||x||=||y||=1$,则称为标准正交（orthonormal）
- $<x,y>=\sum_{i=1}^n x_iy_i=x^Ty$
- 若存在对称正定矩阵H，则存在由H引导的内积：$<x,y>_H=x^*Hy$
- 若$x^*Hy=0$则称向量x,y关于矩阵H共轭
- 完备的内积空间就是希尔伯特(Hilbert)空间，完备赋范空间就是巴拿赫(Banach)空间:具有封闭性!

§2. 方程求解与矩阵

11. 正交补和投影

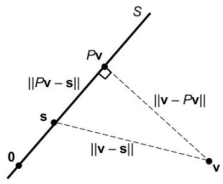
- 如果$S \subseteq V$，V为内积空间，则S的正交补表示为$S^\perp$
  - $S^\perp = \{v \in V | v \perp s, \forall s \in S\}$
  - 显然，对任意$S \subseteq V$， $S^\perp$都是V的子空间。
  - 如果S是一个V中的有限维子空间，则有如下的分解成立（直和分解）

直和分解命题：$\forall v \in V, v = v_S + v_\perp, v_S \in S, v_\perp \in S^\perp$,且表达式唯一确定!

证明：令$u_1, \cdots, u_m$是S的标准正交基，假设$v \in V$,定义$v_S =< v, u_1 > u_1 + \cdots + < v, u_m > u_m$，且$v_\perp = v - v_S$,对$\forall i = 1, \cdots, m$， $< v_\perp, u_i > = < v - v_S, u_i > = 0$；下面证明表达式唯一，不依赖于基的选择。假设有另外一组标准正交基$u'_1, \cdots, u'_m$，类似定义$v'_S, v'_\perp$，

下面证明$v_S = v'_S, v_\perp = v'_\perp$,由$v_S + v_\perp = v = v'_S + v'_\perp$,因此，$v_S - v'_S = v'_\perp - v_\perp$，显然，左边$\in S$,右边$\in S^\perp$,根据这两个空间正交，因此，其内积为0

$0 =< v_S - v'_S, v'_\perp - v_\perp > = \; < v_S - v'_S, v_S - v'_S > = ||v_S - v'_S||^2 \Rightarrow v_S = v'_S$,从而$v'_\perp = v - v'_S = v - v_S = v_\perp$



上述分解的存在性和唯一性表明，只要S是一个空间，则有：$V = S \oplus S^\perp$

投影原理：在空间S中找与v最近的一个点。

↓ 内积空间中的点v映射到v_S总是存在且唯一，定义如下函数：$\boldsymbol{P_S: V \to S}, v \to v_S$ 为到S上的正交投影

正交投影的性质

(1) 对$\forall v \in V$,和一组S的标准正交基$u_1, \cdots, u_m$,$P_S v =< v, u_1 > u_1 + \cdots + < v, u_m > u_m$

P_S还可写成矩阵形式：

$P_S = \sum_{i=1}^m u_i u_i^T = UU^T$，这里$U$的列为$u_1, u_2, \cdots, u_m$

(2) $\forall v \in V, v - P_S v \perp S$

(3) P_S是线性映射

证：假设$x, y \in V, \alpha \in R$.根据分解定理，$x = x_S + x_\perp, y = y_S + y_\perp, x_S, y_S \in S, x_\perp, y_\perp \in S^\perp$，则有 $x + y = x_S + y_S + x_\perp + y_\perp$，因此$P_S(x + y) = x_S + y_S = P_S x + P_S y$,且$\alpha x = \alpha x_S + \alpha x_\perp$，因此 $P_S(\alpha x) = \alpha x_S = \alpha P_S x$,因此$P_S$是线性的

(4) $P_S s = s, \forall s \in S$，当限定在$S$内时，$P_S$是恒等映射

证：如果$s \in S$,则s可写为$s = s + 0, s \in S, 0 \in S^\perp$,因此$P_S = s$

(5) $Range(P_S) = S, Null(P_S) = S^\perp$

证：$Range(P_S) \subseteq S$:有定义可知/ $Range(P_S) \supseteq S$：$\forall s \in S, s = P_S v, \exists v \in V$| $Null(P_S) \subseteq S^\perp$：假设$v \in Null(P)$.将$v$写为：$v = v_S + v_\perp, v_S \in S, v_\perp \in S^\perp$,则$0 = P_S v = v_S$,因此，$v = v_\perp \in S^\perp$/ $Null(P_S) \supseteq S^\perp$:若$v \in S^\perp$,则$v = 0 + v, 0 \in S, v \in S^\perp$,因此$P_S v = 0$

(6) $P_S^2 = P_S$

证：$\forall v \in V, P_S^2 v = P_S(P_S v)) = P_S v$，因此$P_S^2 = P_S$

(7) $\forall v \in V, ||P_S v|| \leq ||v||$

证：$||v||^2 = ||P_S v + (v - P_S v)||^2 = ||P_S v||^2 + ||v - P_S v||^2 \geq ||P_S v||^2$

(8) $\forall v \in V, s \in S, ||v - P_S v|| \leq ||v - s||$,等号当且仅当$s = P_S v$时成立。也就是 $P_S v = argmin_{s \in S} ||v - s||$

证：假设$v \in V, s \in S$. $||v - s||^2 = ||(v - P_S v) + (P_S v - s)||^2 = ||(v - P_S v)||^2 + ||(P_S v - s)||^2 \geq ||(v - P_S v)||^2 \Rightarrow ||v - s|| \geq ||v - P_S v||$,**等号成立当且仅当**$||(P_S v - s)||^2$**=0，即**$s = P_S v$

- 所有信号和数据处理中，都需要对问题进行抽象和刻画
  - 前提：连续性，周期性，平滑性
  - 目的：从已知的数据去预测未知的数据！
  - 连续函数：$C[a,b]$:定义在某闭区间$[a,b]$上的一切连续函数所成的集合
  - 周期函数：$C_{2\pi}$,是定义在整个实轴$(-\infty,\infty)$上的以2π为周期的连续实函数全体所构成的集合
  - 平滑函数（李普希兹连续函数）？

§3.矩阵的稀疏表达

Lipschitz连续
任意斜率小于同一个常数
Lipschitz常数
可导且导数有界
在无限区间上不能有超过线性的增长.

$C_L^{1,1}(R^n)$函数举例.

1.线性函数 $f(x)=a+\langle a,x\rangle$
因为$\nabla f(x)=a, \nabla^2 f(x)=0$

2.二次函数$f(x)=a+\langle a,x\rangle+\frac{1}{2}\langle Ax,x\rangle$，满足$A=A^T$.
$\nabla f(x)=a+Ax, \nabla^2 f(x)=A$

3.单变量函数$f(x)=\sqrt{1+x^2}$
$\nabla f(x)=\frac{x}{\sqrt{1+x^2}}, \nabla^2 f(x)=\frac{1}{(1+x^2)^{\frac{3}{2}}}\le 1$

1.逼近的思想(函数类)

重要积分形式：$\int_0^1 \nabla^2 f(x+\tau(y-x))(y-x)d\tau$.

- 函数类$C_L^{k,p}(Q), Q\subseteq R^n$
  - 任意函数$f\in C_L^{k,p}(Q)$在集合Q上是k次连续可微的
  - 它的p阶导数在Q上是关于常数L是Lipschitz连续的，即对所有$x,y\in Q$，满足$||\nabla^p f(x)-\nabla^p f(y)||\le L||x-y||$
- 函数类具有如下性质：
  - $p\le k$,如果$q\ge k$,则$C_L^{q,p}(Q)\subseteq C_L^{k,p}(Q)$,
  - 如果$f_1\in C_{L_1}^{k,p}(Q), f_2\in C_{L_2}^{k,p}(Q)$,且$\alpha_1,\alpha_2\in R$，则对$L_3=|\alpha_1|L_1+|\alpha_2|L_2$,有$\alpha_1 f_1+\alpha_2 f_2\in C_{L_3}^{k,p}(Q)$,
  - 最常用的函数类：$\boxed{C_L^{1,1}(Q)}$即具有$Lipschitz$连续梯度的函数类，即满足对所有的$x,y\in Q$,有$||\nabla f(x)-\nabla f(y)||\le L||x-y||$

①**引理**：函数$f(x)$属于类$C_L^{2,1}(R^n)\subset C_L^{1,1}(R^n)$,当且仅当对所有$x\in R^n$，我们有：$||\nabla^2 f(x)||\le L$

证明：$\Rightarrow$:对任意的$x,y\in R^n$,我们有：

$\nabla f(y)-\nabla f(x)=\int_x^y \nabla^2 f(t)dt \leftarrow$ 令$t=x+\tau(y-x)$

$\nabla f(y)=\nabla f(x)+\int_0^1 \nabla^2 f(x+\tau(y-x))(y-x)\,d\tau=\nabla f(x)+\left(\int_0^1 \nabla^2 f(x+\tau(y-x))d\tau\right)\cdot(y-x)$

因此有$||\nabla f(y)-\nabla f(x)||=||\left(\int_0^1 \nabla^2 f(x+\tau(y-x))d\tau\right)\cdot(y-x)||\le||\left(\int_0^1 \nabla^2 f(x+\tau(y-x))d\tau\right)||\cdot||(y-x)||\le L||(y-x)||$

$\Leftarrow$：如果$f\in C_L^{2,1}(R^n)$,则对于任意$s\in R^n$和$\alpha>0$，我们有：

$$||\left(\int_0^\alpha \nabla^2 f(x+\tau s)d\tau\right)\cdot s||=||\nabla f(x+\alpha s)-\nabla f(x)||\le \alpha L||s||$$

两边同时除以α,并令$\alpha\downarrow 0$,即得证。

α单调递减收敛于0.

②**定理**：令$f\in C_L^{1,1}(R^n)$,则$\forall x,y\in R^n$，我们有$|f(y)-f(x)-\langle\nabla f(x),y-x\rangle|\le\frac{L}{2}||y-x||^2$

证明：对任意$x,y\in R^n$，我们有：$f(y)=f(x)+\int_0^1\langle\nabla f(x+\tau(y-x)),y-x\rangle d\tau=f(x)+\langle\nabla f(x),y-x\rangle+\int_0^1\langle\nabla f(x+\tau(y-x))-\nabla f(x),y-x\rangle d\tau$

因此，$|f(y)-f(x)-\langle\nabla f(x),y-x\rangle|=|\int_0^1\langle\nabla f(x+\tau(y-x))-\nabla f(x),y-x\rangle d\tau|\le\int_0^1|\langle\nabla f(x+\tau(y-x))-\nabla f(x),y-x\rangle|d\tau\le\int_0^1||\nabla f(x+\tau(y-x))-\nabla f(x)||\cdot||y-x||\,d\tau\le\int_0^1||L\tau(y-x)||\cdot||y-x||\,d\tau=\frac{L}{2}||y-x||^2$

$\rightarrow$由此，可知，对于任意$f\in C_l^{1,1}(R^n)$,取定点$x_0\in R^n$，则可构造两个二次函数

$$\begin{cases}\phi_1(x)=f(x_0)+\langle\nabla f(x_0),x-x_0\rangle-\frac{L}{2}||x-x_0||^2\\ \phi_2(x)=f(x_0)+\langle\nabla f(x_0),x-x_0\rangle+\frac{L}{2}||x-x_0||^2\end{cases}$$

则函数f的图像位于ϕ_1和ϕ_2之间，即$\phi_1(x)\le f(x)\le\phi_2(x),\forall x\in R^n$

$C_L^{1,1}(R^n)$
$\downarrow$
$C_L^{2,2}(R^n)$

定理：令$f\in C_L^{2,2}(R^n),\forall x,y\in R^n$,我们有

- $||\nabla f(y)-\nabla f(x)-\nabla^2 f(x)(y-x)||\le\frac{L}{2}||y-x||^2$
- $|f(y)-f(x)-\langle\nabla f(x),y-x\rangle-\frac{1}{2}\langle\nabla^2 f(x)(y-x),y-x\rangle|\le\frac{L}{6}||y-x||^3$

推论：令$f\in C_L^{2,2}(R^n)$和$x,y\in R^n$,满足$||y-x||=r$,则有 $\nabla^2 f(x)-LrI_n\le\nabla^2 f(y)\le\nabla^2 f(x)+LrI_n$

证明：令$G=\nabla^2 f(y)-\nabla^2 f(x)$,因为$f\in C_L^{2,2}(R^n)$,从而$||G||<Lr$,因此矩阵$G$的特征值$|\lambda_i|\le Lr, i=1,\cdots,n$,因此$-LrI_n\le G\le LrI_n$

$\lambda||X||=||GX||\le||G||\cdot||X||<Lr||X||\Rightarrow|\lambda_i|\le Lr$

2.Weierstrass定理.

定理（Weierstrass）：设函数$f(x)\in C[a,b]$,那么对于任意给定的$\epsilon>0$，都存在这样的多项式$P(x)$，使得：

$$\max_{a\le x\le b}|P(x)-f(x)|<\epsilon$$

这个称为Weierstrass第一定理！

(不要求)证明：$C[a,b]\Leftrightarrow C[0,1]$ 线性变换 $t=(b-a)x+a\ (x\in(0,1))$.

构造n次多项式：$B_n^f(x)=\sum_{k=0}^n f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k}\Rightarrow$ 只须证 $\lim_{n\to\infty}B_n^f(x)=f(x)$在$[0,1]$一致成立.

$\binom{n}{k}$——组合数

多项式根数逼近函数..

§7. 矩阵的稀疏表达

尽可能少的资源表示尽可能多的知识. 提高计算速度.

定理（Weierstrass第二定理）：设$f(x) \in C_{2\pi}$，则对任意给定的$\epsilon > 0$,都有三角多项式$T(x)$存在，使得:

$$\max_{-\pi \le x \le \pi} |f(x) - T(x)| < \epsilon$$

证明（略）：但采用$Vallee - Poussin$算子即可，$V_n[f;x] = \frac{1}{2\pi} \cdot \frac{(2n)!!}{(2n-1)!!} \int_{-\pi}^{\pi} f(t) \cos^{2n} \frac{t-x}{2} dt$,这里$(2n)!! = (2n)(2n-2)\cdots 4 \cdot 2$；$(2n-1)!! = (2n-1)(2n-3)\cdots 3 \cdot 1$

eg. 周期连续函数（不妨假定周期为2π）的最简单逼近工具是如下的三角多项式

$$T(x) = A + \sum_{k=1}^{n} (a_k coskx + b_k sinkx)$$

如果其中的系数a_n, b_n不全为0，则称$T(x)$为n阶三角多项式。

7. 矩阵的特征值与特征向量.

特征值和特征向量的性质

- 矩阵的迹(trace):特征值之和=对角元素之和
- 行列式=特征值之积
- 对称矩阵具有实特征值
- 不同特征值对应的特征向量正交

- 矩阵A可以分解为特征值和特征向量的线性组合，实际上对于$Ax = \lambda x$，如果存在n个特征值$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$,则有$Ax_i = \lambda_i x_i$
- 因此$A[x_1\ x_2 \cdots x_n] = [x_1\ x_2 \cdots x_n] \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \cdots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix}$ → 特征向量相互正交.
- 从而可得$A = [x_1\ x_2 \cdots x_n] \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \cdots & 0 \\ 0 & 0 & \lambda_n \end{pmatrix} [x_1\ x_2 \cdots x_n]^{-1} = \Sigma_{i=1}^{n} \lambda_i x_i x_i^T$
  - 注意：这等价于将矩阵A分解为多个秩为1的矩阵的和　特征分解
  - 若矩阵A的特征值和特征向量为λ_i, x_i，若向量$v = c_1 x_1 + \cdots + c_n x_n$，则
  - $Av = c_1 \lambda_1 x_1 + \cdots + c_n \lambda_n x_n$
  - $A^k v = c_1 \lambda_1^k x_1 + \cdots + c_n \lambda_n^k x_n$
  - 因此，如果$|\lambda_1| > 1$，则$c_1 \lambda_1^k x_1$随k增加而增长，反之，则该成分会成份逐渐消失!

行和为1

随机矩阵(左&右)

列和为1

状态转移矩阵 / 马尔科夫矩阵.

• 矩阵变换与特征值(平移. 相似. 对角化. 转置)

- 如果矩阵A平移为矩阵$A + sI$,则其特征值和特征向量如何变化?
  $$(A + sI)x = \lambda x + sx = (\lambda + s)x$$
- 相似矩阵,
  - 可逆矩阵B，则BAB^{-1}具有相同的特征值:
  - 若$Ax = \lambda x$,则$(BAB^{-1})(Bx) = BAx = B\lambda x = \lambda(Bx)$
  - BAB^{-1}称为矩阵A的相似矩阵! 这可用来计算大矩阵的特征值，逐步化为对角矩阵!
  - 由于特征值满足：$\det(A - \lambda I) = 0$,对于三角矩阵来说，对角线上的元素都是特征值!
- 矩阵的对角化,假设矩阵有n个独立的特征向量$x_1, \cdots, x_n$（大部分矩阵都有）
  - $A[x_1 \cdots x_n] = [Ax_1 \cdots Ax_n] = [\lambda_1 x_1\ \cdots \lambda_n x_n] = [x_1\ \cdots x_n] diag(\lambda_1, \cdots, \lambda_n)$
  - 从而有$A = X\Lambda X^{-1} \Rightarrow A^2 = (X\Lambda X^{-1})(X\Lambda X^{-1}) = X\Lambda^2 X^{-1}$
- 如果矩阵$S = S^T$，则所有特征值λ都是实数，且所有特征向量q可选为相互正交的!
  - $Q = [q_1 \cdots q_n], QQ^T = I, q_i$是标准正交的
  - 此时$Q^{-1} = Q$
  - $A = X\Lambda X^{-1}$变为$S = Q\Lambda Q^T$(谱定理)
  - 对于复数对称矩阵（此时为复共轭）：$S = \begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix} = \bar{S}^T = S^*$其特征值为8和-1.（注意对于复数向量：$< u, v > = u^* v$）　取共轭转置.

• 矩阵的能量.

- 对称正定矩阵的特征值都非负，根据表达形式$x^T S x = \lambda x^T x > 0$，因此只要$\lambda > 0$则$x^T S x > 0$，表示在各种信号处理中的能量!

- 从向量x的角度来看
  - 这是一种有矩阵S引导的范数，如果矩阵S为单位阵，就对应$\sum_{i=1}^{n} x_i^2$，表示能量各个量贡献一致
  - 如果S为非负特征值形成的对角阵，这时对应$\sum_{i=1}^{n} \lambda_i x_i^2$，表示能量对各个成份进行了**权**求和，考虑了成份的重要性!
  - 如果S为对称正定矩阵，对应更复杂的加权求和，可能考虑更多的实际问题的复杂关的描述，例如社交网络中关联矩阵、连接度矩阵等特点

矩阵S的角度：F范数 - tr迹.

矩阵A的一个特征值为λ,则有:
- 特征向量（几何意义上）：$Ax = \lambda x$有非零解
- 特征值（代数意义上）：行列式$A - \lambda I$为0

特征值可能有多重，也就是$A - \lambda I$的行列式中包含有幂次的分解形式
- 几何重数GM：对与λ对应的独立特征向量的个数，对应$A - \lambda I$的零空间维数
- 代数重数AM：特征值λ对应的重复次数，对应$\det(A - \lambda I) = 0$的根
  当$GM < AM$的时候，意味着缺少特征向量，表明A是不能对角化的，没有可逆的特征向量矩阵

§3. 矩阵的稀疏表达

4. 矩阵的特征分解

给定方阵 $A \in R^{n\times n}$，$\lambda \in C$ 是 A 的特征值（$\lambda_1 \dots \lambda_n$），$x \in C^n$ 是相应的特征向量（$x_1 \dots x_n$）

$A[x_1 \dots x_n] = [\lambda_1 x_1 \dots \lambda_n x_n] \Leftrightarrow AX = X\Lambda$

$$\Rightarrow A = [x_1 \dots x_n]\begin{bmatrix}\lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_n\end{bmatrix}[x_1 \dots x_n]^{-1}. \qquad A = X\Lambda X^{-1}$$

特殊地，当矩阵A为对称矩阵时，$A = X\Lambda X^T$，Λ 是从大到小排列的特征值

注：ppt中例子D不是特征值.

$LU \longrightarrow LDL^T$

5. 矩阵的奇异值分解(SVD)*

①矩阵对角化的推广

对方阵 B 而言，矩阵 $A = B^TB$ 是对称矩阵，如果 $Ax = \lambda x$（BB^T），也就是存在特征值和特征向量，并假设 $||x||_2 = 1$，是单位特征向量，则 Bx 也是矩阵 A（BB^T）的特征向量，且特征值也是 λ，并且 Bx 的范数是 $\sqrt{\lambda}$

$BB^TBx = B\lambda x$

$\|Bx\|_2^2 = (Bx)^T(Bx) = x^TAx = x^T\lambda x = \lambda\|x\|_2^2$

如果 y 满足：$BB^Ty = \lambda y, ||y||_2 = 1$，则 B^Ty 也是矩阵 A 的特征值为 λ 的特征向量，而且 B^Ty 的范数是 $\sqrt{\lambda}$

$BB^TBv_i = \lambda_i Bv_i$

②方阵的SVD分解

$Bv_i = u_i\sqrt{\lambda_i}$　　证明

- 方阵 B，则 B^TB 与 BB^T 具有相同特征值。假设 B^TB 的标准正交的特征向量和特征值为 v_i, λ_i，则可找到 BB^T 的所有标准正交特征向量 u_i，使得：$u_i\sqrt{\lambda_i} = Bv_i$，也就是在 B^TB 与 BB^T 的特征向量之间存在关系 (u_i, v_i) 且满足关系：$u_i\sqrt{\lambda_i} = Bv_i$
- （**SVD存在性**）：B^TB 的有序特征向量 $v_1, \cdots, v_n$，以及 BB^T 的有序特征向量 $u_1, \cdots, u_n$ 存在，且 $u_i\sqrt{(\lambda_i)} = Bv_i$，因而有 $[u_1, \cdots, u_n]\Sigma = B[v_1, \cdots, v_n] \Rightarrow U\Sigma = BV$，且 U, V 都是单位正交矩阵，$UU^T = VV^T = I$，从而有 $B = U\Sigma V^T$，从而方阵的SVD分解总是存在

特征值…

③非方阵的SVD分解

- 假设矩阵 B 是由矩阵 $D_{n\times d}$ 填充0而成的 $m \times m$ 的方阵($m = \max\{n, d\}$)，这时按照分块矩阵的处理可以获得非方阵的SVD分解
- $B = [D\ 0] = U\Sigma V^T = U\begin{bmatrix}\Sigma_1 & 0 \\ 0 & 0\end{bmatrix}\begin{bmatrix}V_1 & 0 \\ 0 & \boxed{V_2}\end{bmatrix}^T, d < n$
- $B = \begin{bmatrix}D \\ 0\end{bmatrix} = U\Sigma V^T = \begin{bmatrix}U_1 & 0 \\ 0 & \boxed{U_2}\end{bmatrix}\begin{bmatrix}\Sigma_1 & 0 \\ 0 & 0\end{bmatrix}V^T, n < d$

无意义.

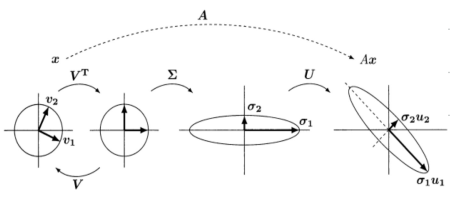
- 其中 U, V, Σ 都是 $m \times m$ 大小，V_1, U_1 为 $d \times d$ 大小，V_2, U_2 为 $(m-d) \times (m-d)$ 和 $(m-n) \times (m-n)$，Σ_1 为 $\min\{n, d\} \times \min\{n, d\}$
- 矩阵 A 的维数为 $m \times n$，其$\boxed{\text{秩为}r}$。例如数据矩阵中列表示属性，行表示样本，此时一般情况下不是方阵，期望
  - $Av_1 = \sigma_1 u_1, \cdots, Av_r = \sigma_r u_r, Av_{r+1} = 0, \cdots, Av_n = 0$
  - 特征值降序排列：$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$
  - 后 $n - r$ 个 v 在 A 的零空间中，最后 $m - r$ 个 u 在 A^T 的零空间中
  - $V^T = V^{-1}, U^T = U^{-1}$

$$\boldsymbol{AV = U\Sigma} \qquad A\underbrace{\begin{bmatrix} v_1 \,..\, v_r \,..\, v_n \end{bmatrix}}_{n\times n} = \underbrace{\begin{bmatrix} u_1 \,..\, u_r \,..\, u_m \end{bmatrix}}_{m\times m}\underbrace{\left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]}_{m\times n}$$

- $A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T$
- 其中 $\sigma_1 u_1 v_1^T$ 是与矩阵 A 最接近的秩一的矩阵，而前 k 项之和则是最好的秩 k 的矩阵！$A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T$

Eckart-Young：如果 B 秩为 k，则 $||A - A_k|| \leq ||A - B||$

PCA降维



P0

§3. 矩阵的稀疏表达

6. 矩阵的稀疏表示 $\min_x ||x||_0, s.t. Ax=b$

对于 $A\in R^{m\times n}$，$m<n$，且矩阵A为行满秩矩阵

①稀疏解的唯一性

假设$A=[I,F]$，I为单位矩阵，F为傅立叶变换矩阵。此时线性方程组$Ax=b$是欠定方程组，因而信号b可以表示为单位阵中的一列和傅立叶矩阵中的列所形成的三角函数的叠加

方程个数小于变元个数

此时解为稀疏的，则表示信号可以由一些尖刺函数和三角函数的叠加

②不确定性原理

- 一般的不确定原理是指由傅立叶变换耦合的一对共轭变量不能都具有任意高精度
- 即对函数$f(x)$，以及其傅立叶变换$F(\omega)$必定满足不等式：
- $\int_{-\infty}^{\infty} x^2|f(x)|^2dx\cdot\int_{-\infty}^{\infty}\omega^2|F(\omega)|^2d\omega\geq\frac{1}{2}$,假设$\int_{-\infty}^{\infty}|f(x)|^2dx=1$
- 用时频理论意味着不能同时在时域和频域都能高精度确定信号！
- 在我们的视角，就是信号不能同时在时域和频域都能稀疏表示！在时频分析中称为测不准原理！

互相关性 $\mu(A)$

列向量内积最大值

$Proximity(\Phi,\Psi)=\mu(A)=\max_{i\leq n,j\leq n}|\psi_i^T\phi_j|$

且$\frac{1}{\sqrt{n}}\leq\mu(A)\leq 1$

下界值当取某些正交基时达到

例如单位阵和Fourier变换阵，

或单位阵和Hadamard矩阵等

- 假设非零向量或信号$b\in\mathbb{R}^n$,以及两个正交基Ψ,Φ，则b要么能由Ψ或者Φ的列的线性组合来表示，即：
- $b=\Psi\alpha=\Phi\beta$
- 显然，α,β唯一确定！尤其重要的一类情况是，Ψ就是简单的单位矩阵，而Ψ是傅立叶变换矩阵，此时α相当于b的时域表达，而β相当于b的频域表达！
- 而对于任意一组基Ψ,Φ，一个有趣的现象就是信号b在由α,β表达是，最多只能有一个是稀疏的，而不能两个都是稀疏的！（测不准原理）

△ 定理：互相关性$\mu(A)$的一对正交基Ψ,Φ，任意非零向量$b\in\mathbb{R}^n$在基上的表示分别为α,β,则下列不等式成立： $||\alpha||_0+||\beta||_0\geq\frac{2}{\mu(A)}$

证明：1.不确定性原理(均值不等式) ⟶ 1.不确定性原理(拉格朗日乘子法)

$1=b^Tb=\alpha^T\Psi^T\Phi\beta=\sum_{i=1}^n\sum_{j=1}^n\alpha_i\beta_j\psi_i^T\phi_j\leq\mu(A)\sum_{i=1}^n\sum_{j=1}^n|\alpha_i||\beta_j|\leq\mu(A)||\alpha||_1||\beta||_1$

- 由此可得，如果两组基的互相关性小，则α,β不能都很稀疏
  - 例如，如果Ψ为恒等单位阵，Φ为傅里叶变换矩阵，则$\mu([\Psi,\Phi])=\frac{1}{\sqrt{n}}$
  - 表明信号不能有在时域和频语都有小于$2\sqrt{n}$的非零系数。
  - 极限情况就是都有$\sqrt{n}$个值，如均匀间隔$\sqrt{n}$的尖峰信号，由傅里叶变换转换为同一个信号，累积$2\sqrt{n}$个非零值。

△ **定理2**：线性系统 $[\Psi,\Phi]x=b$的任意两个不同解不能都是稀疏的，有以下的测不准原理：$||x_1||_0+||x_2||_0\geq\frac{2}{\mu(A)}$

证明：考虑$Ax=[\Psi,\Phi]x=b$在不确定性条件下，假设两个解x_1,x_2，其中一个是稀疏的，则另一个肯定不能是非常稀疏的。注意：$e=x_1-x_2$肯定在A的零空间中，将其分割为$e=[e_\psi\ e_\phi]^T$，则$Ae=0\Rightarrow\Psi e_\psi=-\Phi e_\phi=y\neq0$

显然y非0，因为e非0，并且Ψ,Φ非奇异：

注意：$||e||_0=||e_\Psi||_0+||e_\Phi||_0\geq\frac{2}{\mu(A)}$

因此有：$||x_1||_0+||x_2||_0\geq||e||_0\geq\frac{2}{\mu(A)}$

唯一性

定理3：线性系统 $[\Psi,\Phi]x=b$的 一个候选解有小于$\frac{1}{\mu(A)}$个 非零值，则该解就是最稀疏的解，并且任何其他解都必定是稠密的。

全局最优！ ⟵ 一般非凸优化，只有局部最优。

§7. 矩阵的稀疏表达

⑤ 矩阵的Spark

Spark用来使用l_0范数来刻画矩阵的零空间!

定义：给定矩阵 A的 spark是指矩阵A中线性相关的最小列的数目!

如果$Ax=0$，则$||x||_0 \geq spark(A)$

$\Longleftrightarrow$ 秩是指矩阵A中线性独立的最大列数

Spark的计算更困难，需要搜索矩阵A的所有可能列数的组合

由spark判定稀疏解唯一性的简单准则：

定理：唯一性：如果线性方程组$Ax=b$的解x满足$||x||_0 < spark(A)/2$,则这个解必定是最稀疏的!

证明：考虑另一个解$y, Ay=b \Rightarrow A(x-y)=0 \Rightarrow ||x||_0+||y||_0 \geq ||x-y||_0 \geq spark(A)$,由于$||x||_0 < spark(A)/2$，因此必定由别的解中非零值的个数大于$spark(A)/2$。

⑥ 贪婪算法 ⟶ 局部最优的单项更新

$$\min_x ||x||_0, s.t. Ax=b$$

$||x||_0 = k_0 < \frac{1}{2} spark(A)$

x为稀疏解

假设$spark(A) > 2k_0$，最优化问题已知有值$val(P_0)=k_0$.

$\Rightarrow$ b是矩阵A最多k_0列的线性组合.

(底层搜索)

常规算法：if $k_0=1$ ⟶ $O(m^{k_0} n k_0^2)$浮点操作.

因此采用m次测试，来测试矩阵A的每一列就可以识别该列!第j次测试可以通过最小化$\epsilon(j)=||a_j z_j - b||_2$来完成，从而可得

$b = z_j^* a_j$ 常数

$$z_j^* = \frac{a_j^T b}{||a_j||_2^2} \Rightarrow \epsilon(j) = \min_{z_j} ||\, a_j z_j - b||_2^2 = ||\frac{a_j^T b}{||a_j||_2^2} a_j - b||_2^2 = ||b||_2^2 - 2\frac{\left(a_j^T b\right)^2}{||a_j||_2^2} + \frac{\left(a_j^T b\right)^2}{||a_j||_2^2} = ||b||_2^2 - \frac{\left(a_j^T b\right)^2}{||a_j||_2^2}$$

如果误差为0，则求得正确解。因此测试完成就意味着：

$$||a_j||_2^2\, ||b||_2^2 = \left(a_j^T b\right)^2$$

这对应柯西施瓦茨不等式中等式成立的条件，表明b和a_j平行，过程要求$O(mn)$阶的浮点数操作，比较合理!

1′ 正交匹配追击(OMP)

$$\epsilon(j) = \min_{z_j} ||a_j z_j - r^{k-1}||_2^2 = ||\frac{a_j^T r^{k-1}}{||a_j||_2^2} a_j - r^{k-1}||_2^2$$
$$= ||r^{k-1}||_2^2 - 2\frac{\left(a_j^T r^{k-1}\right)^2}{||a_j||_2^2} + \frac{\left(a_j^T r^{k-1}\right)^2}{||a_j||_2^2}$$
$$= = ||r^{k-1}||_2^2 - \frac{\left(a_j^T r^{k-1}\right)^2}{||a_j||_2^2}$$

$O(mnk_0)$次

因此计算最小误差等价于计算残差与矩阵A的剩余列向量之间的内积

对比迭代优化

假设不能得到精确解

任务：获取(P_0): $\min_x ||x||_0, s.t. Ax=b$的近似解

输入参数：给定矩阵A, b,以及误差阈值ϵ_0

输出　　：第k次近似解x^k

初始化：$k=0$,并设定

- 初始解：$x^0=0$
- 初始残差：$r^0 = b - Ax^0 = b$
- 初始解支撑$S^0 = Support\{x^0\} = \emptyset$

迭代过程：$k \leftarrow k+1$

- **Sweep**:计算误差$\epsilon(j) = \min_{z_j} ||a_j z_j - r^{k-1}||_2^2$,其中$z_j$用最优值$\boxed{z_j^* = \frac{a_j^T r^{k-1}}{||a_j||_2^2}}$代入 贪心
- **更新支撑**：求$\epsilon(j)$的最小指标j_0: $\forall j \notin S^{k-1}, \epsilon(j_0) \leq \epsilon(j)$，更新$S^k = S^{k-1} \cup \{j_0\}$
- **更新临时解**：计算$x^k = minize_x ||Ax-b||_2^2, s.t. Support\{x\} = S^k$;
- **更新残差**:计算$r^k = b - Ax^k$;
- **停止准则**：如果$||r^k||_2 < \epsilon_0$则停止，否则继续循环

输出：输出x^k

支撑

x在S^k上不为0.

$\min ||A_{S^k} x_{S^k} - b||_2^2$，导数为0 $\Leftrightarrow A_{S^k}^T(A_{S^k} x_{S^k} - b) = -A_{S^k}^T r^k = 0$

下一次不会被选中.

正交性

§3. 矩阵的稀疏表达

最小二乘

2' 改进LS-OMP (Sweep阶段)

O(mnk₀)次.

如在OMP中，在第k步，执行$m-|S^{k-1}|$次LS步。

其中LS步可使用如下更有效的公式：

$$\begin{bmatrix} M & b \\ b^T & c \end{bmatrix}^{-1} = \begin{bmatrix} M^{-1}+pM^{-1}bb^TM^{-1} & -pM^{-1}b \\ -pb^TM^{-1} & p \end{bmatrix} \quad (*)$$

这里$p=\frac{1}{c-b^TM^{-1}b}$.用A_s表示从A中收集的$k-1$列（s^{k-1}）所形成的子矩阵，则待解的LS问题为：

$$\min_{(x_s,z)} ||[A_s\ a_i]\begin{bmatrix} x_s \\ z \end{bmatrix} - b||_2^2$$

其解析解的必要条件（求偏导数为0）：$\begin{bmatrix} A_s^T \\ a_i^T \end{bmatrix}[A_s\ \ a_i]\begin{bmatrix} x_s \\ z \end{bmatrix} - \begin{bmatrix} A_s^T \\ a_i^T \end{bmatrix} b = 0$

从而有：$\begin{bmatrix} x_s \\ z \end{bmatrix}_{opt} = \begin{bmatrix} A_s^TA_s & A_s^Ta_i \\ a_i^TA_s & ||a_i||_2^2 \end{bmatrix}^{-1}\begin{bmatrix} A_s^Tb \\ a_i^Tb \end{bmatrix}$

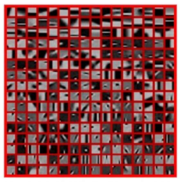
根据(*)式，如果已经存储了矩阵$(A_s^TA_s)^{-1}$,则上式计算很容易，并得到k列的更新矩阵。一旦获得解，也得到残差，从而选择得到最小误差的列a_i。观察到如果a_i与A_s的列正交，则上述矩阵可逆并成为块对角阵，结果为$x_s=A_s^{-1}b, z=\frac{a_i^Tb}{||a_i||_2^2}$

这表明以前的解不变，并引入了一个系数为z的非零值

⑤ 字典

1. 定义

- 给定一幅图像$\boldsymbol{X}$,假设稀疏表示的基本单元为块，图像向量化表示为$\boldsymbol{X}\in\mathbb{R}^N$,其位置$\boldsymbol{i}(\boldsymbol{i}=\boldsymbol{1},\cdots\boldsymbol{n})$处的图像块$\boldsymbol{x_i}\in\mathbb{R}^{\boldsymbol{B_s}}$,图像块大小为$\sqrt{\boldsymbol{B_s}}\times\sqrt{\boldsymbol{B_s}}$,如图所示
- $\boldsymbol{x_i}=\boldsymbol{R_iX},\boldsymbol{R_i}\in\mathbb{R}^{\boldsymbol{B_s}\times\boldsymbol{N}}$表示从图像$\boldsymbol{X}$提取图像块$\boldsymbol{x_i}$的矩阵运算符，假设采用重叠块分解图像
  - 冗余高
  - 质量好
  - 问题：从$\boldsymbol{x_i}$中恢复$\boldsymbol{X}$



- $X=\left(\sum_{i=1}^n R_i^TR_i\right)^{-1}\sum_{i=1}^n\left(R_i^Tx_i\right)$
  - 最小二乘法
- 给定字典$\boldsymbol{D}$,对图像块$\boldsymbol{x_i}$进行稀疏编码，等价于求解一个$\boldsymbol{x_i}$的系数向量$\boldsymbol{\alpha_i}$,使得$\boldsymbol{x_i}\approx\boldsymbol{D\alpha_i}$，因此，图像X表示为一组系数向量$\{\boldsymbol{\alpha_i}\}$。并且具有稀疏性
- 这过程可以表示为
- $\alpha_i=argmin_\alpha\frac{1}{2}\|x_i-D\alpha\|_2^2+\lambda\|\alpha\|_0$, 稀疏性
- 其中λ为常数，$\|\alpha\|_0$表示ℓ_0范数，即α_i中非零元素的个数。由于ℓ_0范数优化问题是非凸的并且是NP难问题，通常采用贪心算法求解次优解
- 比如：正交追踪匹配（Orthogonal matching pursuit, OMP）算法

2. K-Means求解矢量量化问题.

- K均值算法
- 输入：训练数据$\boldsymbol{X}=\{\boldsymbol{x_i}\}_{i=1}^n$
- 输出：字典$\boldsymbol{D}=[\boldsymbol{D_{.1}},\boldsymbol{D_{(.2)}},\cdots,\boldsymbol{D_{(.N)}}]$，码$\boldsymbol{\alpha}$
- 初始化：$\boldsymbol{t=1},\boldsymbol{D^{(0)}}\in\boldsymbol{R^{B_s\times M}}$为$\boldsymbol{X}$中的一个随机样本，$||\boldsymbol{D^j}||_2=1$
- While (不收敛)

  稀疏编码阶段：将训练数据$\boldsymbol{X}$划分为$\boldsymbol{N}$个子集$\left(\boldsymbol{R_1^{(t-1)}},\boldsymbol{R_2^{(t-1)}},\cdots,\boldsymbol{R_N^{(t-1)}}\right)$,每个$\boldsymbol{R_k^{(t-1)}}$都有与列$\boldsymbol{D_{.k}^{(t-1)}}$最相似的样本索引号，即$\boldsymbol{R_k^{(t-1)}}=\left\{\boldsymbol{i}|\forall\boldsymbol{l}\neq\boldsymbol{k},||\boldsymbol{x_i}-\boldsymbol{D_{.k}^{(t-1)}}||_2<||\boldsymbol{x_i}-\boldsymbol{D_{.l}^{(t-1)}}||_2\right\}$

  字典更新阶段：根据$\boldsymbol{D_j^{(t)}}=\frac{1}{|\boldsymbol{R_j^{(t-1)}}|}\boldsymbol{\Sigma}_{\mathrm{i}\in\boldsymbol{R_j^{(t-1)}}}\boldsymbol{x_i}$来更新$\boldsymbol{D^{(t-1)}}$的每一列，$\boldsymbol{j=1,\cdots,N}$

  计数：$\boldsymbol{t=t+1}$
- End while

当$\boldsymbol{D}$给定，则每个信号都可以表示为其最近的码字

$$\boldsymbol{argmin_{\alpha_i}||x_i-D\alpha_i||_2^2, s.t. \alpha_i=e_j}$$

α_i只有一个元素为1，其余均为0!

VQ问题.

$\boldsymbol{MSE}$为$\boldsymbol{E=\Sigma_{i=1}^n||x_i-D\alpha_i||_2^2=||X-D\alpha||_F^2}$

3. K-SVD思路几乎相同

- While (不收敛)
  - 稀疏编码阶段：使用稀疏回复算法来获取x_i的稀疏稀疏向量α_i: 与数据一一对应
  - $\underset{\{\alpha_i\}}{\mathrm{argmin}}\|x_i-D\alpha_i\|_2^2$, s.t. $\|\alpha_i\|_0<T$,
  - 字典更新阶段：根据下述过程来更新$\boldsymbol{D^{(t-1)}}$的每一列，$\boldsymbol{j=1,\cdots,N}$
    - 计算误差矩阵E_m：$E_m=X-\Sigma_{(j\neq m)0}D_{.j}\alpha_{j.}$
    - 从E_m计算$\bar{E}_m$,选择ω_m对应的列:$\omega_m=\{i|1\le i\le n,\alpha_{m.}^{(i)}\neq 0\}$
    - 对$\bar{E}_m$应用SVD分解：$\bar{E}_m=U\Delta V^T$，更新字典为：$D_{.m}=u_1,\alpha_{m.}=\Delta(1,1)\cdot V_1$
  - $t=t+1$

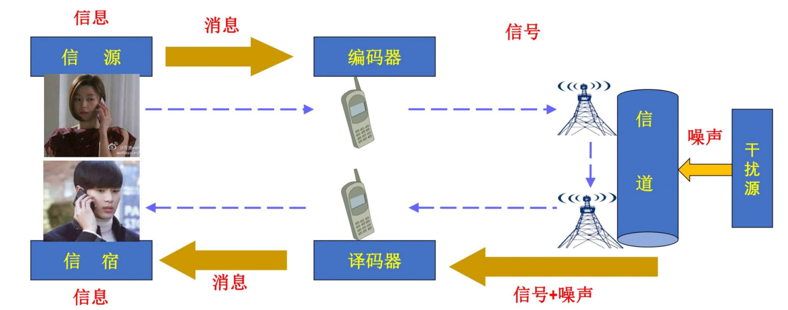
§4. 不确定度量与熵

1. 信息　信息. 消息. 信号.

{ 信号是消息的物理表现形式
{ 消息是信号的具体内容

⇒ 信息 { 广义：对物质存在和运动形式的一般描述
　　　 { 狭义：信息可以用来消除不确定性的东西.

2. 通信系统模型.　在某一点精确或近似地恢复另一点发送的信息.



通信系统模型

有效性
可靠性
安全性

- **香农第一定理**——为了无失真地传输信源信息，信源编码的极限是什么
- **香农第二定理**——在有噪信道中无失真地通信，信道编码的极限是多少
- **香农第三定理**——如果允许一定量的失真，信源编码的极限是什么？

3. 自信息量：该事件发生概率的对数的负值.　$I(x_i)=-\log p(x_i)$

底数 2. e. 10 ⇒ 比特. 奈特. 哈特

事件 { 发生前：先验不确定性
　　 { 发生后：能提供的最大信息量.

1. $f[P(x_i)]$应该是 $P(x_i)$的单调递减函数，若$P(x_1) > P(x_2)$，$f[P(x_1)] < f[P(x_2)]$
2. 当 $P(x_i) = 1$，　$f[P(x_i)] = 0$
3. 当 $P(x_i) = 0$，　$f[P(x_i)] = \infty$
4. 两个独立事件的联合信息量应该等于各自信息量之和。

- **离散无记忆信源**(Discrete Memoryless Source，简记为DMS)输出的是单个符号的消息，不同时刻发出的符号之间彼此统计独立，而且符号集中的符号数目是有限的或可数的。离散无记忆信源的数学模型为离散型的概率空间，即

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_I \\ p(x_1) & p(x_2) & \cdots & p(x_I) \end{bmatrix}$$

- $p(x_i)$：信源输出符号消息x_i的先验概率；满足：$0 \le p(x_i) \le 1$，$1 \le i \le I$
- **离散平稳有记忆信源**用联合概率空间$\{X, P(X)\}$来描述离散有记忆信源的输出。信源在i时刻发出什么符号与i时刻以前信源所发出的符号有关，即由条件概率$p(x_i | x_{i_1} x_{i_2} \ldots)$确定。如果该条件概率分布与时间起点无关，只与关联长度有关，则该信源为平稳信源。

对于离散平稳有记忆信源，有：

$$p(x_1 = a_1) = p(x_2 = a_1) = \ldots$$
$$p(x_2 = a_2 | x_1 = a_1) = p(x_3 = a_2 | x_2 = a_1) = \ldots$$
$$p(x_3 | x_2 x_1) = p(x_4 | x_3 x_2) = \ldots$$
$$\vdots$$
$$p(x_{i+L} | x_{i+L-1} x_{i+L-2} \ldots x_i) = p(x_{j+L} | x_{j+L-1} x_{j+L-2} \ldots x_j) = \ldots$$

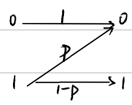
⟹

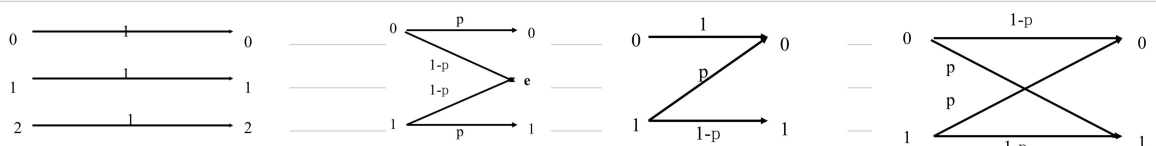
- **离散无记忆信道**
  - 离散无记忆信道的输入和输出消息都是离散无记忆的单个符号，输
号$x_i \in \{a_1, a_2, \ldots, a_k\}, 1 \le i \le I$，输出符号$y_j \in \{b_1, b_2, \ldots, b_D\}, 1 \le j \le$
信道的特性可表示为转移概率矩阵：

$$P = \begin{bmatrix} p(y_1|x_1) & p(y_2|x_1) & \cdots & p(y_J|x_1) \\ p(y_1|x_2) & p(y_2|x_2) & \cdots & p(y_J|x_2) \\ & \vdots & & \\ p(y_1|x_I) & p(y_2|x_I) & \cdots & p(y_J|x_I) \end{bmatrix}$$

  - $p(y_j | x_i)$对应为已知输入符号为x_i，当输出符号为y_j时的信道转移概
满足$0 \le p(y_j | x_i) \le 1$

{ 无干扰信道　输入 = 输出
{ 二元删除信道　引入疑问
{ 二元Z信道
{ 二元对称信道





无干扰信道　　二元删除信道　　二元Z信道　　二元对称信道

§4. 不确定度量与熵

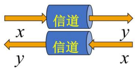
4. 互信息量：随机事件y的出现给出关于事件x的信息量

$$I(x;y)=\log\frac{p(x|y)}{p(x)}$$

即，原有的不确定性－尚存在的不确定性。

- 对称性$I(x;y)=I(y;x)$：由y所提供的关于x的信息量 = 由x所提供的关于y的信息量
- 当事件x、y统计独立时，互信息量为0：p(x|y)=p(x)，x和y之间没有什么关系，无论是否知道y，都对x出现的概率没有影响
- 可正可负：y的出现有利/不利于确定x的发生
  - 正：$I(x;y)=\log\frac{p(x|y)}{p(x)}>0\Rightarrow\frac{p(x|y)}{p(x)}>1\Rightarrow p(x|y)>p(x)$，$x$：张三病了。$y$：张三没来上课。
  - 负：$I(x;y)=\log\frac{p(x|y)}{p(x)}<0\Rightarrow\frac{p(x|y)}{p(x)}<1\Rightarrow p(x|y)<p(x)$，$x$：李四考了全班第一名。$y$：李四没有复习功课。
- 互信息量不大于任一事件的自信息量 $I(x;y)=I(x)-I(x|y)$

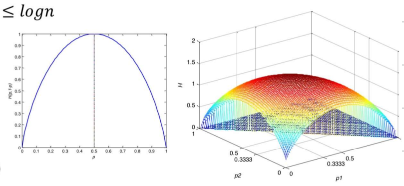
  $I(y;x)=I(y)-I(y|x).$



熵：信源X的不确定程度 $H(x)=-\sum_{i=1}^{n}p_i\log p_i$

也等于各事件的信息量的期望

- 条件熵：$H(X|Y)=\sum_{XY}p(x,y)I(x|y)=-\sum_{XY}p(x,y)\log p(x|y)$
  - 若X表示输入，Y表示输出，$H(X|Y)$表示信道损失
- 联合熵（共熵）：$H(X,Y)=\sum_{XY}p(x,y)I(x,y)=-\sum_{XY}p(xy)\log p(x,y)$
- 熵的性质
- 1. 对称性：与整体有关，个体无关
- 2. 非负性：$H(X)\geq 0$；$H(X)=E[I(x_i)]=\sum_{i=1}^{n}p(x_i)\log\frac{1}{p(x_i)}$
- 3. 扩展性：$\lim_{\varepsilon\to 0}H_{q+1}(p_1,p_2,\cdots,p_q-\varepsilon,\varepsilon)=H_q(p_1,p_2,\cdots,p_q)$
  - **集合X有q个事件，集合Y比X仅仅是多了一个概率接近0的事件，则两个集合的熵值一样**
  - 证明：$x\log x$ 在$[0,\infty)$的连续性，$\lim_{x\to 0}x\log x=0$
  - 含义：集合中，一个事件发生的概率比其它事件发生的概率小得多时，这个事件对于集合的熵值的贡献可以忽略
- 4. 可加性：设X和Y为两个互相关联的随机变量，X的概率分布为$\{p_1,p_2,\ldots,p_m\}$，Y的概率分布为$\{q_1,q_2,\ldots,q_n\}$，则 $H(XY)=H(X)+H(Y|X)$. $\sum_i\sum_j$
  - 当X、Y相互独立时，$H(X,Y)=H(X)+H(Y)$
- **5. 极值性**：$H(p_1,p_2,\ldots,p_n)\leq H(\frac{1}{n},\frac{1}{n},\ldots,\frac{1}{n})=\log n$, $H(X)\leq logn$
  - **离散信源最大熵定理**：各事件等概率发生时，熵最大
  - $\begin{bmatrix}X\\P\end{bmatrix}=\begin{bmatrix}x_1 & x_2\\ p & 1-p\end{bmatrix}$
  - $H(X)=-p\log p-(1-p)\log(1-p)$
  - $\begin{bmatrix}X\\P\end{bmatrix}=\begin{bmatrix}x_1 & x_2 & x_3\\ p_1 & p_2 & 1-p_1-p_2\end{bmatrix}$
  - $H(X)=-p_1\log p_1-p_2\log p_2-(1-p_1-p_2)\log(1-p_1-p_2)$



- 6. 确定性
- 7. 上凸性

最大熵定理证明：

① 首先证明：$\sum p_i\log(p_i)\geqslant\sum p_i\log(q_i)$ $\quad \ln x\leq x-1$

$\Rightarrow \sum p_i\log\frac{q_i}{p_i}\leq\sum p_i(\frac{q_i}{p_i}-1)=0$ 得证.

② 后证函数凹性. $H(\lambda P+(1-\lambda)Q)\geqslant\lambda H(P)+(1-\lambda)H(Q)$

可化为 $-\sum(\lambda p_i+(1-\lambda)q_i)\log(\lambda p_i+(1-\lambda)q_i)+\lambda\sum p_i\log p_i+(1-\lambda)\sum q_i\log q_i$

$=-\lambda\sum p_i\cdot\log\frac{\lambda p_i+(1-\lambda)q_i}{p_i}-(1-\lambda)\sum q_i\cdot\log\frac{\lambda p_i+(1-\lambda)q_i}{q_i}\geqslant 0$

故存在最大值！

③ 找最大值（拉格朗日乘子法）

$G(P)=-\sum p_i\log p_i-\lambda(\sum p_i-1)$

$$\Rightarrow\begin{cases}\frac{\partial G}{\partial p_i}=-\log p_i-1-\lambda\\ \sum p_i-1=0\end{cases}\Rightarrow\begin{cases}p_i=e^{-1-\lambda}\\ ne^{-1-\lambda}=1\end{cases}\Rightarrow p_i=\frac{1}{n}.$$

§4. 不确定度量与熵

(不要求)

结论

- **定理**：设离散随机变量的概率空间为$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ p_1 & p_2 & \cdots & p_N \end{bmatrix}$，函数$f(p_1, p_2, \cdots, p_N)$是随机变量不确定的度量，若此函数满足条件：1）连续性；2）单调性；3）可加性，则此函数必为$f(p_1, p_2, \cdots, p_N) = -C\sum_{i=1}^{N} p_i \log p_i$，其中$C$为常数
- **证明**：（1）考虑随机变量X为等概分布的情况，这时有$p_i = \frac{1}{N}, i = 1, \cdots, N$。由条件3）可知：$g(MN) = f\left(\frac{1}{MN}, \frac{1}{MN}, \cdots, \frac{1}{MN}\right) = g(M) + \sum_{i=1}^{M} \frac{1}{M} g(N) = g(M) + g(N)$，则有$g(s^m) = mg(s), g(t^n) = ng(t)$,其中$s, t, m, n$都是正整数
- 显然，我们可以选择m, n,使之满足$s^m \le t^n < s^{m+1}$,取对数则有：
- $m\log s \le n\log t < (m+1)\log s$，同时除以$n\log s$,则有$\frac{m}{n} \le \frac{\log t}{\log s} < \frac{m+1}{n}$,因此有$\left|\frac{m}{n} - \frac{\log t}{\log s}\right| < \frac{1}{n}$，由条件2），$mg(s) \le ng(t) \le (m+1)g(s)$,两边除以$ng(s)$，得：$\frac{m}{n} \le \frac{g(t)}{g(s)} < \frac{m+1}{n}$,因此有$\left|\frac{m}{n} - \frac{g(t)}{g(s)}\right| < \frac{1}{n}$,因此，根据$|a \pm b| \le |a| + |b|$得：$\left|\frac{g(t)}{g(s)} - \frac{\log t}{\log s}\right| < \frac{2}{n}$，由于$s, t$任意，$n$可任意大，因此有：$\frac{g(t)}{g(s)} = \frac{\log t}{\log s}$，即$g(t) = C\log t$,其中$C$为常数!

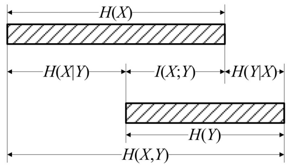
(2)考虑随机变量X为非等概分布，但此时概率为有理数的情况，此时概率可表示为$p_n = \frac{m_n}{\sum_{n=1}^{N} m_n}$,其中$m_n$为整数。令$M = \sum_{n=1}^{N} m_n$,由条件3），$g(M)$可写为$g(M) = f(p_1, p_2 \cdots, p_N) + \sum_{n=1}^{N} p_n g(m_n)$，于是$f(p_1, p_2 \cdots, p_N) = g(M) - \sum_{n=1}^{N} p_n g(m_n) = C\log M - \sum_{n=1}^{N} p_n Cg(m_n) = C\log M(\sum_{n=1}^{N} p_n) - C\sum_{n=1}^{N} p_n g(m_n)$,因此
$f(p_1, p_2 \cdots, p_N) = -C\sum_{n=1}^{N} p_n \log(m_n/M) = \ -C\sum_{n=1}^{N} p_n \log(p_n)$

(3)若X非等概分布，但概率为无理数的情况，根据无理数可以用有理数逼近以及条件1），可以证明此情况下仍然有这个结论!

5. 平均互信息量、相对熵

各种信息量之间的关系：平均互信息量与信源熵、条件熵的关系

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



- **相对熵**：$D(p||q) = \sum_{x\in X} p(x) \log\left(\frac{p(x)}{q(x)}\right)$:相对熵,交叉熵，Kullback熵，K-L散度，K-L距离，方向散度)
  - 相对熵非负，互信息非负
  - **证明**：1)对任意$x > 0$,都有$1 - \frac{1}{x} \le \ln x \le x - 1$成立，等式成立的充要条件是$x = 1$,令$f(x) = x - 1 - \ln x, x > 0$，则$f''(x) = \frac{1}{x^2} > 0$,从而$f(x)$为下凸函数，且最小值$x = 1$取得$f(1) = 0$，从而$x - 1 \ge \ln x$,然后令$x := \frac{1}{x}$，代入得证；
  - 2)$\log\left(\frac{q_i}{p_i}\right) = \log e \cdot \ln\left(\frac{q_i}{p_i}\right) \le \log e \cdot \left(\frac{q_i}{p_i} - 1\right)$,从而：
  $$\sum_i p_i \log\left(\frac{q_i}{p_i}\right) \le \sum_i p_i \log e \cdot \left(\frac{q_i}{p_i} - 1\right) = \log e \cdot \sum_i (q_i - p_i) = 0$$
  - 从而$\sum_i p_i \log\left(\frac{p_i}{q_i}\right) \ge 0$,等号成立当且仅当$\frac{p_i}{q_i} = 1$
  - 对数和不等式
    - 对非负数$a_1, a_2, \cdots, a_n$以及非负数$b_1, b_2, \cdots, b_n$,有：$\sum_{i=1}^{n} a_i \log\left(\frac{a_i}{b_i}\right) \ge (\sum_{i=1}^{n} a_i)\log(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i})$，等式当且仅当$\frac{a_i}{b_i} =$ 常数成立(注意$0\log 0 = 0; a\log\left(\frac{a}{0}\right) = \infty, a > 0; 0\log\left(\frac{0}{0}\right) = 0$)
  - 利用该不等式，可以证明相对熵的凸性：
  - $D(p||q)$关于(p, q)对是凸的，即$\forall\lambda \in [0,1]$,二元组$(p_1, q_1), (p_2, q_2)$,有$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \le \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$
    - 证明：应用对数不等式到上式的左端项，$(\lambda p_1(x) + (1-\lambda)p_2(x)) \log\left(\frac{\lambda p_1(x)+(1-\lambda)p_2(x)}{\lambda q_1(x)+(1-\lambda)q_2(x)}\right) \le \lambda p_1(x) \log\left(\frac{\lambda p_1(x)}{\lambda q_1(x)}\right) + (1-\lambda)p_2(x)\log\frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} = \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$

决策树：信息增益

- 数据处理不等式
  - 如果随机变量X, Y, Z形成马尔可夫链，即$X \to Y \to Z$，$p(x, y, z) = p(x)p(y|x)p(z|y)$，则$I(X;Y) \ge I(X;Z)$
  - 直观：任何对数据进行高级操作都不可能改进由数据得到的推断
  - **证明**：马尔可夫链表明给定Y时,X, Z条件独立，从而$I(X;Z|Y) = 0$.由于$I(X;Y|Z) \ge 0$,由$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y)$,可得$I(X;Y) \ge I(X;Z)$
  - 等式成立当且仅当$I(X;Y|Z) = 0$,也就是$X \to Z \to Y$形成马尔可夫链.类似可以证明$I(Y;Z) \ge I(X;Z)$
  - 且若$Z = g(Y)$,则$I(X;Y) \ge I(X;g(Y))$,表明数据Y的函数不能提高关于X的信息。

条件独立：$p(X|Y,Z) = p(X|Y)$

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

§5. 贝叶斯公式与检测准则

全概率公式

$$P(B)=\sum_{i=1}^{n}P(A_j)P(B|A_j)$$

贝叶斯公式

$$P(A_j|B)=\frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{\infty}P(A_i)P(B|A_i)},$$

$$P(A|B)=\frac{P(A)P(B|A)}{P(A)P(B|A)+P(\bar{A})P(B|\bar{A})}$$

1. 基本概念论

- **定义1（事件）**：设Ω是实验S的样本空间，用F表示Ω的某些子集构成的集合，如果F满足：
  - （1）$\Omega\in F$
  - （2）如果 $A\in F$，则 $\bar{A}\in F$
  - （3）若 $A_j\in F$，则 $\bigcup_{i=1}^{n}A_j\in F$，

就称F是Ω的事件域或σ域(σ-field 或 σ-algebra)，称F中的元素为事件，称(Ω, F)是可测空间。

- **定义2（概率）**：设(Ω, F)是可测空间，P是定义在F上的函数，如果P满足下面的条件：
  - （1）非负性：for any $A\in F$， $P(A)\ge 0$
  - （2）完全性：$P(\Omega)=1$
  - （3）可列可加性：对于F中互不相容的事件 $A_1, A_2, \ldots$，有 $P\left(\bigcup_{j=1}^{\infty}A_j\right)=\sum_{j=1}^{\infty}P(A_j)$

就称P为F上的概率测度，简称概率，称(Ω,F,P)为概率空间。

- **定义3（条件概率）**：设$A\in F, P(A)\ge 0$，若$P(A)>0, B\in F$，定义$P_A(B)=P(B|A)=\frac{P(A,B)}{P(A)}$，则$P_A$也是可测空间(Ω, F)上的概率，我们称$P_A$为条件概率测度，简称条件概率

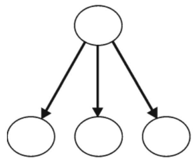
性质：设A, B, A_1，A_2，· · ·，A_n是事件，则：
  - (1)$P(A,B)=P(B|A)P(A)=P(A|B)P(B)$;
  - (2)当$P(A_1A_2\cdots A_n)\ne 0$，有:$P(A_1,A_2,\cdots,A_n)=P(A_1)P(A_2|A_1)\cdots P(A_n|A_1,A_2,\cdots,A_{n-1})$
  - (3)如果$P(A,B)>0$，则$P_A(C|B)=P(C|B|A)=\frac{P(C,B,A)}{P(B,A)}=P(C|A,B)$

2. 贝叶斯公式

$$P(A_j|B)=\frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{\infty}P(A_i)P(B|A_i)}$$

分母不一定为1

- 分类问题
  - 无监督-聚类：将集合分成n组或n类
  - 有监督-分类：已知类别或标签，找到映射关系
- 概率观点的有监督分类
  - 给定属性$A_1,A_2,\cdots,A_n$和类别$C=\{c_1,c_2,\cdots,c_m\}$，使得给定属性下最大化概率MAP：$arg_C[max\,P(C|A_1,A_2,\cdots,A_n)]$
  - 令$A=\{A_1,A_2,\cdots,A_n\}$，则为$arg_C max P(C|A)$



- Bayesian 分类器
  - $P(C|A_1,A_2,\cdots,A_n)=P(C)P(A_1,A_2,\cdots,A_n|C)/P(A_1,\cdots,A_n)$
  - 紧凑表示为：$P(C|A)=P(C)P(A|C)/P(A)$
  - 因此分类器问题表示为：$arg_C max\left[\frac{P(C)(A|C)}{P(A)}\right]$

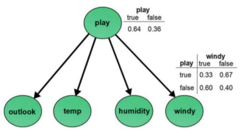
朴素贝叶斯：假定属性之间独立.

进一步，根据单调性，可以表示成很多等价形式(类别属性不会变化，$P(A)$为常数)

$arg_C[max[P(C)P(A|C)]]$

$arg_C\left[max\left[log(P(C)P(A|C))\right]\right]$

$arc_C[max(logP(C)+logP(A|C))]$



图模型 (eg. 贝叶斯网络).

逻辑回归：

二项逻辑回归可用于一些简单的二分类问题，不妨设$Z\in R^n$是问题的输入$X\in\{0,1\}$是输出。则逻辑回归假设$X|Z$服从伯努利分布。具体模型为

$$P(X=1|Z)=\frac{1}{1+e^{-\theta^Tz}};\quad P(X=0|Z)=1-\frac{1}{1+e^{-\theta^Tz}}$$

其中，θ^T为模型的参数，$\theta^Tz=\theta_0+\theta_1z_1+\cdots+\theta_nz_n$。应用时，根据两个条件概率值的大小将实例分到概率值较大的一类。

令观测函数为$h_\theta(z)=1/(1+e^{-\theta^Tz})$，于是，可以整合上述逻辑回归的概率模型得到输入集合的最大似然估计为$L(\theta)=\prod_{i=1}^{m}P(x_i|z_i;\theta)=(h_\theta(z))^x(1-h_\theta(z))^{1-x}$，取对数，乘$-1/m$将最大值问题转化为最小值问题。

$$L(\theta)=\left(-\frac{1}{m}\right)\prod_{i=1}^{m}\log P(x_i|z_i;\theta)=-\frac{1}{m}\sum_{i=1}^{m}[x_i\log h_\theta(z_i)+(1-x_i)\log((1-h_\theta(z_i))]$$

$L(\theta)$是凸的、可微的，因此可以运用梯度下降法、牛顿法及其他方法对此问题进行求解此优化问题。

生成式模型及判别式模型.

	生成式模型	判别式模型
特点	对后验概率建模，从统计的角度表示数据的分布情况，能够反映同类数据本身的相似度	寻找不同类别之间的最优分类面，反映的是异类数据之间的差异
区别	估计的是联合概率分布：$p(x,y)$	估计的是条件概率分布：$p(y\|x)$
联系	由产生式模型可以得到判别式模型，但由判别式模型得不到产生式模型。	
常见模型	朴素贝叶斯，马尔可夫随机场等	逻辑回归，支持向量机（SVM），神经网络，k近邻算法等
优点	①实际上带的信息更丰富 ②研究单类问题灵活性强 ③模型可以通过增量学习得到，能用于数据不完整情况。	①分类边界更灵活，能清晰的分辨出类别间差异特征; ②适用于较多类别的识别; ③模型简单，比较容易学习。
缺点	学习和计算过程比较复杂，且容易产生错误信息	不能反映训练数据本身的特性，变量间的关系不可视。
性能	较差	较好

§5. 贝叶斯公式与检测准则

接收端信号为

$x(t)=s(t)+n(t)$

$\begin{cases} H_0: s=0 \\ H_0: s\neq 0 \end{cases}$ 假设检验

3. 检测准则

当知道先验概率$p(x_0)$和$p(x_1)$，则发送一个信号的平均损失C_M可以表示为：

$C_M = p(x_0)C_0 + p(x_1)C_1 = p(x_0)[(1-\beta)C_{00} + \beta C_{10}] + p(x_1)[(1-\alpha)C_{11} + \alpha C_{01}]$

$= p(x_0)C_{00} + p(x_1)C_{11} + p(x_0)\cdot\beta(C_{10} - C_{00}) + p(x_1)\cdot\alpha(C_{01} - C_{11})$

$= p(x_0)C_{00} + p(x_1)C_{11} + \int_{s_{y_0}} [(C_{01} - C_{11})\, p(x_1)p(y|x_1) - (C_{10} - C_{00})p(x_0)p(y|x_0)]\, dy$

找判决阈值等价于最小化分险，即：$\min\left\{\int_{s_{y_0}}[B(y) - A(y)]\, dy\right\}$，等价于$max\{\int_{-\infty}^{y}[A(y) - B(y)]\, dy\}$，**显然**，当$A(y) = B(y)$是满足条件，即

$$(C_{01} - C_{11})\, p(x_1)p(y|x_1) = (C_{10} - C_{00})p(x_0)p(y|x_0)$$

①**贝叶斯准则**即为损失函数、转移概率分布和事件的先验概率都已知时，

$$y = \begin{cases} 1, \dfrac{p(y|x_1)}{p(y|x_0)} > \dfrac{(C_{10} - C_{00})p(x_0)}{(C_{01} - C_{11})p(x_1)} \\ 0, \dfrac{p(y|x_1)}{p(y|x_0)} \le \dfrac{(C_{10} - C_{00})p(x_0)}{(C_{01} - C_{11})p(x_1)} \end{cases}$$

令$L(y) = \frac{p(y|x_1)}{p(y|x_0)}$为似然比，$\lambda = \frac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)}$为检测阈值，则条件转化为$L(y) \gtreqless \lambda$来判断接收的类别。

②而**最大后验概率准则**为损失函数未知时，简单假设错误判决的损失为1而正确判决的损失为0，即$C_{00} = C_{11} = 0$，$C_{01} = C_{10} = 1$，此时$\lambda = \frac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)} = \frac{p(x_0)}{p(x_1)}$，此时条件转化为$L(y) \gtreqless 1$来判断接收的类别。

③**最大似然准则**则当损失函数和先验概率均未知时，假定两种判决出错的概率和代价基本相同，假定常数$\lambda = \frac{(C_{10}-C_{00})p(x_0)}{(C_{01}-C_{11})p(x_1)} = 1$，此时条件转化为$L(y) \gtreqless 1$来判断接收的类别。

原式 $= p(x_0)C_{00} + p(x_1)C_{11} + p(x_0)\beta(C_{10}-C_{00}) + p(x_1)\alpha(C_{01}-C_{11})$

$= p(x_0)C_{00} + p(x_1)C_{11} + \int_{s_{y_1}}(C_{10}-C_{00})p(y|x_0)p(x_0)dy + \int_{s_{y_0}}(C_{01}-C_{11})p(y|x_1)p(x_1)dy$

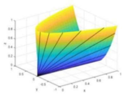
$= p(x_0)C_{00} + p(x_1)C_{11} + (C_{10}-C_{00})p(x_0)\left(1 - \int_{s_{y_0}}p(y|x_0)dy\right) + \int_{s_{y_0}}(C_{01}-C_{11})p(y|x_1)p(x_1)dy$

$= p(x_0)C_{00} + p(x_1)C_{11} + (C_{10}-C_{00})p(x_0) - \int_{s_{y_0}}(C_{01}-C_{11})p(y|x_1)p(x_1) - (C_{10}-C_{00})p(x_0)p(y|x_0)dy$

§6、凸优化

1. 基本概念

- 矩阵A的秩表示为$rank(A)$,非空集合$C \subseteq R^n$的维数$dimC$定义为:
  - $n - max\{rank(A): A \in R^{n\times n}, Ax = Ay$,对所有$x, y \in C\}$
- 凸组合(convex combination):点$x_1, x_2, \cdots, x_n$的凸组合是指点$x = \Sigma_{i=1}^n \alpha_i x_i, \alpha_i \geq 0$, 且$\Sigma\alpha_i = 1$
- 仿射组合：点$x_1, x_2, \cdots, x_n$的仿射组合是指点$x = \Sigma_{i=1}^n \alpha_i x_i$, 且$\Sigma\alpha_i = 1$
- 锥组合：点x_1, x_2的锥组合是指形如$x = \alpha_1 x_1 + \alpha_2 x_2, \alpha_1 > 0, \alpha_2 > 0$
- 凸集(convex set):令$A \subseteq R^n$是凸的，如果$\forall x, y \in A$,则$\alpha x + (1-\alpha)y \in A, \alpha \in (0,1)$
- 凸包（convex Hull)
  - 集合A的凸包$conv(A)$定义为A中点的所有凸组合.
  - 凸包是包含集合的最小凸集
  - 离散点的凸包示例，扇形凸包示例
- 仿射包（affine Hull)
  - 集合$A \subseteq R^n$的仿射包为A中点的组合：affine$A \coloneqq \{x | x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k, x_1, x_2, \cdots, x_k \in A, \sum_i \alpha_i = 1\}$
  - 一般情况下，一个集合的仿射包实际上是包含该集合 的最小的仿射集
- 极点(extreme point):点$x \in A$如果不能表示为非空凸集A中两个任何其它点的凸组合的形式，则称为极点
  - 集合A的极点是这样的点$x \in A$，使得$x \notin conv(A \backslash \{x\})$
  - 集合A为凸，当且仅当A中点的所有凸组合仍在A中
  - 集合A的凸包是包含A的最小凸集
  - 凸集的交是否是凸集？凸集的并集是否是凸集？
- 凸锥(convex cone)
  - 集合$A \subseteq \mathbb{R}^n$中任意元素$x, y \in A, \lambda, \mu \geq 0$,有$\lambda x + \mu y \in A$,则称集合$A$为一个凸锥。若$x_1, \cdots, x_k \in A$,对任意$x \in A$,存在数$\lambda_1, \cdots, \lambda_k \geq 0$,使得$x = \sum_{i=1}^k \lambda_i x_i$,则称凸锥由$x_1, \cdots, x_k$生成
  - 若锥是由有限向量集生成的，则称该锥是有限生成的
- 半正定锥
  - S^n为$n \times n$的对称矩阵集合，$S_+^n = \{x \in S^n | x \geq 0\}$为$n \times n$的半正定矩阵集合，$S_{++}^n = \{x \in S^n | x > 0\}$为$n \times n$的正定矩阵集合
  - 则S_+^n是凸锥，也称为半正定锥
  - 如$(x\ y; y\ z) \in S_+^2$,点(x, y, z)构成一个半正定锥
- 保凸运算：任意多个凸集的交和仿射变换都是保凸的，缩放、平移和投影都是仿射变换，从而证明集合是凸集，或者直接按照定义证明



- 顶点(Vertex):如果点x是两个或多个线段、边缘、表面的交点，则称为顶点
  - 注意，凸集的顶点都是极点
- 如果两个极点由一条边(线，或面)连接，则称为相邻的极点
- 超平面(Hyperplane):集合$H = \{x | Ax = b\}, a \neq 0$:超平面将$R^n$空间分成两个半空间$H^+ = \{x | Ax \geq b\}, H^- = \{x | Ax \leq b\}$
- 多胞形(polyhedron)$P \subseteq R^n$:有限个半空间的交
  - 一个多面锥（polyhedral cone)是形如$\{x: Ax \leq 0\}$的多胞形
  - 多面锥是锥，且是有限生成锥
  - 若$dimP = n$，则称P是满维的；满维等价于它有一内点
- 多面体(polytope):有界的多胞形
- 引理（Minkowski,1896)：设$C = \{x \in R^n, Ax \leq 0\}$是一多面锥，则$C$是肯定由方程组$My = b'$的解集的一个子集生成，这里$M$是由$\binom{A}{I}$的$n$个线性无关行所组成，$b' = \pm e_j$,其中$e_j$是某一单位向量。