1. 高斯分布

1.1 似然函数

MLE: $\theta_{MLE} = \arg\max_\theta P(X|\theta)$ ⇒ 令 $P=1$ 时，$\theta = (\mu, \sigma^2)$

$$\log P(X|\theta) = \log \prod_{i=1}^{N} P(x_i|\theta) = \sum_{i=1}^{N} \log P(x_i|\theta)$$

Data: $X = (x_1, x_2 \dots x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}_{N \times p}$

$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$ (均值)

$\theta = (\mu, \Sigma) = (\mu, \sigma^2)$

均值

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

有偏估计

真正的无偏: $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_{MLE})^2$

期望计算后等于本身 ⇒ 无偏估计

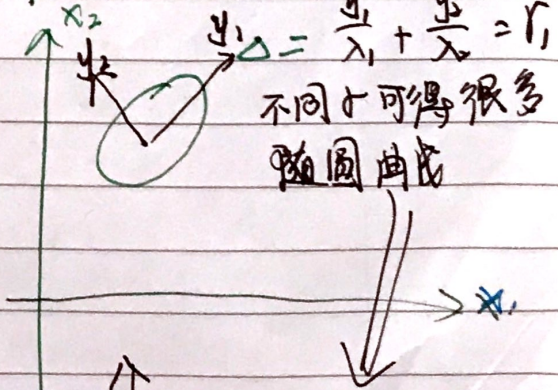
$$E[\mu_{MLE}] = \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu \Rightarrow \text{无偏的}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i\mu_{MLE} + \mu_{MLE}^2) =$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu_{MLE}^2$$

$$E(\sigma^2_{MLE}) = \sigma^2 - \frac{1}{N}\sigma^2 = \boxed{\frac{N-1}{N}\sigma^2}$$ → 所以有偏

$x \sim N(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

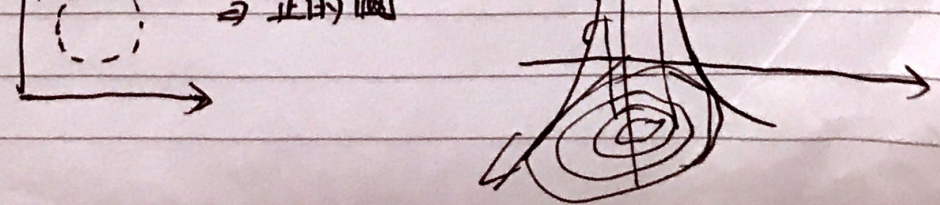


Σ 是 $p \times p$ 维的

$\Sigma_{p \times p}$ 有 → $\frac{p^2+p}{2} = O(p^2)$

① Σ 为对角矩阵 $(\lambda_1 \lambda_2 \dots \lambda_p)$ ⇒ ⇒ 正的椭圆

② 若 $\lambda_1 = \lambda_2 = \lambda_p$ ⇒ 正的圆



2. 高斯分布的局限性：有的数据不能统计为一个

3. 联合概率分布 $X=\begin{pmatrix}X_a\\X_b\end{pmatrix}\begin{matrix}\to n\text{维}\\\to n\text{维}\end{matrix}$ $m+n=p$ $\mu=\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}$ $\Sigma=\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}$

① 边缘概率分布 $P(X_b|X_a)$
条件概率分布 $P(X_a|X_b)$

② 定理：已知 $X\sim N(\mu,\Sigma)$
$Y=AX+B$
$\therefore Y\sim N(A\mu+B,\ A\Sigma A^T)$
$E[Y]=E[AX+B]=A\mu+B$
$Var[Y]=A\Sigma A^T$ （多维的）

③ $E[X_a]=(I_m\ 0)\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix}=\mu_a$
$Var[X_a]=\Sigma_{aa}$
$\therefore X_a\sim N(\mu_a,\Sigma_{aa})$

a和b的关系

$X_b|X_a$
$X_{b\cdot a}=X_b-\Sigma_{ba}\Sigma_{aa}^{-1}X_a$
$\mu_{b\cdot a}=\mu_b-\Sigma_{ba}\cdot\Sigma_{aa}^{-1}\mu_a$
$\Sigma_{bb\cdot a}=\Sigma_{bb}-\Sigma_{ba}\cdot\Sigma_{aa}^{-1}\Sigma_{ab}$

$\Rightarrow X_{b\cdot a}=\underbrace{(-\Sigma_{ba}\Sigma_{aa}^{-1}\cdot I_n)}_{A}\begin{pmatrix}X_a\\X_b\end{pmatrix}$

$E[X_{b\cdot a}]=\mu_b-\Sigma_{ba}\cdot\Sigma_{aa}^{-1}\mu_a=\mu_{b\cdot a}$
$Var[X_{b\cdot a}]=(-\Sigma_{ba}\Sigma_{aa}^{-1}\cdot I_n)\begin{pmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{pmatrix}\begin{pmatrix}-\Sigma_{aa}^{-1}\Sigma_{ba}^{T}\\I\end{pmatrix}$
$=\Sigma_{bb\cdot a}$

$X_b=X_{b\cdot a}+\Sigma_{ba}\Sigma_{aa}^{-1}X_a$
$E[X_b|X_a]=\mu_{b\cdot a}+\Sigma_{ba}\cdot\Sigma_{aa}^{-1}X_a$
$Var[X_b|X_a]=Var[X_{b\cdot a}]=\Sigma_{bb\cdot a}$

$X_b|X_a\sim N(\mu_{b\cdot a}+\Sigma_{ba}\Sigma_{aa}^{-1}X_a,\Sigma_{bb\cdot a})$

已知 $p(x)=N(x|\mu, \Lambda^{-1})$ （Λ^{-1}：精度矩阵）

$p(y|x)=N(y|Ax+b, L^{-1})$

$$p(x|y)=\frac{p(y|x)\cdot p(x)}{p(y)}$$

求：$p(y)$ $p(x|y)$

解：$y=Ax+b+\varepsilon$

$\varepsilon \sim N(0, L^{-1})$

$E[y]=E[Ax+b+\varepsilon]$

$=A\mu+b+0$

方差：$\mathrm{Var}[y]=\mathrm{Var}[Ax+b+\varepsilon]$

$=A\cdot\Lambda^{-1}A^T+L^{-1}$

∴ $y \sim N(A\mu+b;\ L^{-1}+A\Lambda^{-1}A^T)$

前提：（定理）

$x \sim N(\mu, \Sigma)$

$y=Ax+B$

则 $y \sim N(A\mu+B,\ A\Sigma A^T)$

设：$z=\begin{pmatrix}x\\y\end{pmatrix} \sim N\left(\begin{bmatrix}\mu\\A\mu+b\end{bmatrix}, \begin{bmatrix}\Lambda^{-1} & ①\\ ① & L^{-1}+A\Lambda^{-1}A^T\end{bmatrix}\right)$

① $=\Delta=\Lambda^{-1}A^T$ → 求出来的：$\Delta=\mathrm{Cov}(x,y)$

∴ $p(x|y) \sim$ ⟷ 代入求得。

前面求过的：$P(x_b|x_a)=N(\mu_{b\cdot a};\ \Sigma_{bb\cdot a})$

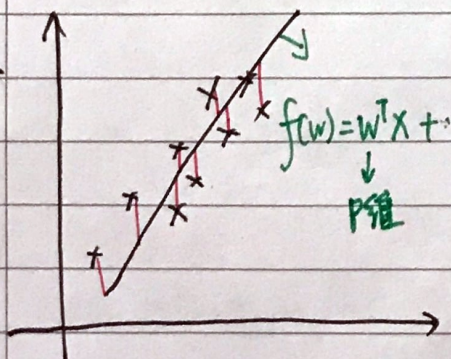
(三) 线性回归

1. 最小二乘法（几何意义：误差的和 // 投影

概率：噪声为高斯的 MLE）

正则化：$\begin{cases}L_1 \to \text{lasso}\\ L_2 \to \text{Ridge，岭回归}\end{cases}$

2.



Data：$D=\{(x_1,y_1)\ (x_2,y_2)\ \cdots\ (x_n,y_n)\}$

$x_i \in \mathbb{R}^p$，$y_i \in \mathbb{R}$，$i=1,2,\cdots N$

$$X=(x_1\ x_2\ \cdots\ x_N)^T=\begin{pmatrix}x_1^T\\x_2^T\\\vdots\\x_N^T\end{pmatrix}=\begin{pmatrix}x_{11}\ x_{12}\ \cdots\ x_{1p}\\\vdots\\x_{N1}\ x_{N2}\ \cdots\ x_{Np}\end{pmatrix}_{N\times p}$$

$$Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_N\end{pmatrix}_{N\times 1}$$

最小二乘法：

$L(w)=\sum_{i=1}^{N}\|w^T x_i - y_i\|^2$ → Loss function

$=\sum_{i=1}^{N}(w^T x_i - y_i)^2$ ⇒ 写成矩阵的形式

向量 × 向量的转置

$=(w^T X^T - Y^T)(Xw - Y)$

$=w^T X^T X w - 2w^T X^T Y + Y^T Y$

$\hat{w} = \arg\min L(w)$

$\frac{\partial L(w)}{\partial w} = 2X^T X w - 2X^T Y = 0$ 对 w 求导

$\therefore X^T X w = X^T Y$

$w = (X^T X)^{-1} X^T Y$

$(X^T X)^{-1} X^T$ → 伪逆

② 概率视角：~~LOS~~ · LSE = MLE

前提 ⇓ (噪声是高斯分布

设 $\varepsilon \sim N(0, \sigma^2)$

$y = f(w) + \varepsilon$

$f(w) = w^T x$ → 拟合

$p(y|x;w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(y-w^T x)^2}{2\sigma^2}\}$

$y = w^T x + \varepsilon$

$\therefore y|x;w \sim N(w^T x, \sigma^2)$ ⇑ 仍然符合高斯分布

MLE =

因为每个点都是独立的

$\mathcal{L}(w) = \log P(Y|X,w) = \log \prod_{i=1}^{N} P(y_i|x_i;w)$

$= \sum_{i=1}^{N} \log \cdots$

$= \sum_{i=1}^{N} \left(\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right)$

$\hat{w} = \arg\max_w \mathcal{L}(w)$

$= \arg\max_w -\frac{1}{2\sigma^2}(y_i - w^T x_i)^2$

$= \arg\min_w (y_i - w^T x_i)^2$ → 就是最小二乘估计.

3. 正则化：

Loss Function：$L(w)=\sum_{i=1}^{N}\|(w^Tx_i-y_i\|^2$　　$\hat{w}=(X^TX)^{-1}X^TY$

$X_{N\times p}$　N个样本　p维　理论上 $N\gg P$

若 $N<P$，⇒ 过拟合 ⇒ 可以有很多种拟合方法

① 框架：对 $L(w)$ 加约束

$\arg\min_w [L(w)+\lambda P(w)]$　　L1：Lasso，$P(w)=\|w\|$

L2：Ridge 岭回归　$P(w)=\|w\|^2=w^Tw$

权值衰减

$$\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2+\lambda w^Tw$$

$$=(w^TX^T-Y^T)(Xw-Y)+\lambda w^Tw$$

$$=w^T(X^TX+\lambda I)w-2w^TX^TY+Y^TY$$

$\hat{w}=\arg\min_w J(w)$

$$\frac{\partial J(w)}{\partial w}=2(X^TX+\lambda I)w-2X^TY=0$$

$\hat{w}=(X^TX+\lambda I)^{-1}X^TY$　可逆的

贝叶斯计算后：$P(w|y)=\frac{P(y|w)P(w)}{P(y)}$

$\hat{w}=\arg\max_w P(w|y)$

$=\arg\min\sum_{i=1}^{N}(y_i-w^Tx_i)^2+\frac{\sigma^2}{\sigma_0^2}\|w\|_2^2$　一样

↗ 极大似然估计

LSE ⟺ MLE（噪声为高斯分布）

Regularized LSE ⟺ MAP（噪声为高斯分布）先验也是高斯分布

↘ 最大后验估计

(四) 线性分类

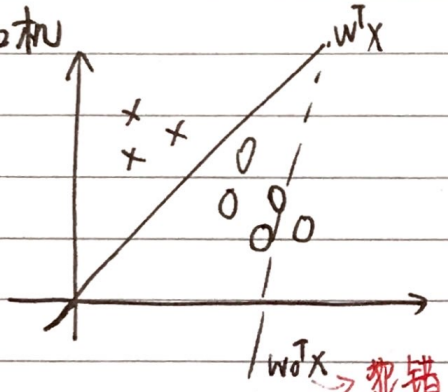
- 硬分类 $y \in \{0,1\}$
  - 线性判别 fisher
  - 感知机
- 软分类 $y \in [0,1]$
  - 生成式：Gaussian Discriminant Analysis
  - 判别式：Logistic Regression

$y = f(w^T x + b)$ f^{-1}：link function

$\{0,1\} \to w^T x + b$

线性回归 ⟹ 激活函数 降维 ⟹ 线性分类 0, 1

① 感知机



错误驱动

模型：$f(x) = \text{sign}(w^T x)$ $x \in R^p$, $w \in R^R$

$$\text{sign}(a) = \begin{cases} +1 & a \geq 0 \\ -1 & a < 0 \end{cases}$$

→ 犯错 (假定是线性可分的)

向正确的移动

策略：Loss function

$$L(w) = \sum_{i=1}^{N} I\{y_i w^T x_i \leq 0\}$$

$w^T x_i > 0$ $y_i = +1$ $\Rightarrow$ 分类正确

$w^T x_i < 0$ $y_i = -1$ $\Rightarrow$ 分类正确 $y_i w^T x_i > 0$

⟹ 小于0则错误分类

$w \Rightarrow w + \delta w$ 整体的值无限趋近于0

$$L(w) = \sum_{x_i \in D} -y_i w^T x_i$$

求导：$\nabla_w L = -y_i x_i$

$w^{(t+1)} \leftarrow w^{(t)} - \lambda \nabla_w L$

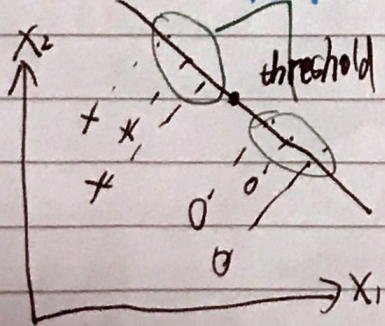
$w^{(t)} + \lambda y_i x_i$

类内小 类间大

threshold 可对多少是x 小于多少是0

且选定合适的轴

② 线性判别分析



投影：$w^T x = z_i$

$$\bar{z} = \frac{1}{N} \sum_{i=1}^{N} w^T x_i$$

协方差 $S_z = \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z})(z_i - \bar{z})^T$

2类数据：C_1：$\bar{z}_1 = \frac{1}{N_1} \sum_{i \in C_1} w^T x_i$

$$S_1 = \frac{1}{N_1} \sum_{i \in C_1} (w^T x_i - \bar{z}_1)(w^T x_i - \bar{z}_1)^T$$

C_2：$\bar{z}_2 = \frac{1}{N_2} \sum_{i \in C_2} w^T x_i$

$$S_2 = \frac{1}{N_2} \sum_{i \in C_2} (w^T x_i - \bar{z}_2)(w^T x_i - \bar{z}_2)^T$$

类间：$(\bar{z}_1 - \bar{z}_2)^2$

类内：$S_1 + S_2$

类内小类间大：

目标函数：$J(w)=\frac{(\bar{z}_1-\bar{z}_2)^2}{s_1+s_2}$ ↑大

$=\frac{[w^T(\bar{x}_{c1}-\bar{x}_{c2})]^2}{w^T(s_{c1}+s_{c2})w}$

$=\frac{w^T(\bar{x}_{c1}-\bar{x}_{c2})(\bar{x}_{c1}-\bar{x}_{c2})^T w}{w^T(s_{c1}+s_{c2})w}$

$=\frac{w^T s_b w}{w^T s_w w}$

$\hat{w}=\arg\max_w J(w)$

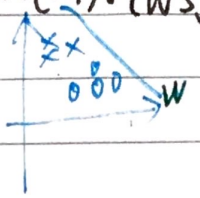
S_b = between-class：类间方差 $p\times p$

S_w = with-class：类内方差 $p\times p$

求最优解：

$\frac{\partial J(w)}{\partial w}=2S_b w(w^T S_w w)^{-1}+w^T S_b w\cdot(-1)\cdot(w^T S_w w)^{-2}\cdot 2S_w\cdot w=0$

$\therefore S_w\cdot w=\frac{w^T S_w w}{w^T S_b w}\cdot S_b\cdot w$ （$\frac{w^T S_w w}{w^T S_b w}$ → 实数）



不关心w的大小，只关心方向

$w=\frac{w^T S_w\cdot w}{w^T S_b\cdot w}S_w^{-1}\cdot S_b\cdot w \propto S_w^{-1}(\bar{x}_{c1}-\bar{x}_{c2})$

只有这部分与方向有关

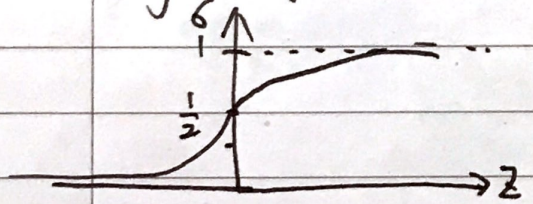
如果S_w^{-1}是对角、各向同性，即$S_w^{-1}\propto I$

则 $w\propto(\bar{x}_{c1}-\bar{x}_{c2})$

② 软输出：概率判别模型

sigmoid function：$\sigma(z)=\frac{1}{1+e^{-z}}$

$\begin{cases} z\to\infty, \lim\sigma(z)\to 1 \\ z\to 0 \quad \text{即 } \sigma(z)=\frac{1}{2} \\ z\to-\infty \quad \lim\sigma(z)=0 \end{cases}$



$\sigma:\mathbb{R}\to(0,1)$ 实数映射到(0,1)区间

若 $w^Tx\mapsto p$

$p_1=P(y=1|x)=\sigma(w^Tx)=\frac{1}{1+e^{-w^Tx}}$ → 记作 $\psi(x;w)$

$p_0=P(y=0|x)=1-P(y=1|x)=\frac{e^{-w^Tx}}{1+e^{-w^Tx}}$

看哪个大就属于哪一类

综合表达

$P(y|x)=p_1^y p_0^{1-y}$

$H(p,q)=-\sum_x p(x)\log q$

交叉熵 = cross Entropy

MLE：$\hat{w}=\arg\max_w \log P(Y|X)=\arg\max_w \log\prod_{i=1}^N P(y_i|x_i)=\arg\max_w \log\prod_{i=1}^N P(y_i|x_i)$

$=\arg\max_w \sum_{i=1}^N \log P(y_i|x_i)=\arg\max_w \sum_{i=1}^N (y_i\log p_i+(1-y_i)\log p_0)$

$=\arg\max_w \sum_{i=1}^N y_i\log\psi(x_i;w)+(1-y)\log(1-\psi(x_i;w))$

← cross Entropy

Gaussian Discriminat Analysis (连续)

④ 软输出：概率生成模型 { Naive Bayes (离散)

判别：$\hat{y} = \underset{y\in\{0,1\}}{\arg\max}\ P(y|x)$　　目的 $P(y=0|x) \overset{比大}{=} P(y=1|x)$ 谁大就属于哪一类

生成：$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$　（贝叶斯模型）

~~目的 $P(y=0|x)$~~ 正比

$P(y|x) \propto P(x|y)P(y)$

posterior　likelihood　prior

$\hat{y} = \underset{y\in\{0,1\}}{\arg\max}\ P(y|x) = \underset{y}{\arg\max}\ P(y)\cdot P(x|y)$

$y \sim \text{Bernoulli}(\phi) \Rightarrow$

y	1	0
ϕ	ϕ	$1-\phi$

$\{ \phi^y \ y=1$；$(1-\phi)^{1-y} \ y=0$

$x|y=1 \sim N(\mu_1, \Sigma)$　$\{ N(\mu,\Sigma)^y$

$x|y=0 \sim N(\mu_2, \Sigma)$　$\{ N(\mu,\Sigma)^{1-y}$

log 似然函数 (log-likelihood)：$L(\theta) = \log \prod_{i=1}^{N} P(x_i, y_i)$

$\theta(\mu_1, \mu_2, \Sigma, \phi)$

$= \sum_{i=1}^{N} \log (P(x_i|y_i)\cdot P(y_i))$

$= \sum_{i=1}^{N} [\log P(x_i|y_i) + \log P(y_i)]$

$\hat{\theta} = \arg\max_{\theta} L(\theta)$

$= \sum_{i=1}^{N} [\log N(\mu_1,\Sigma)^{y_i} \cdot N(\mu_2,\Sigma)^{1-y_i} + \log \phi^{y_i}(1-\phi)^{1-y_i}]$

$= \sum_{i=1}^{N} \log N(\mu_1,\Sigma)^{y_i} + \log(\mu_2,\Sigma)^{1-y_i} + \log \phi^{y_i}(1-\phi)^{1-y_i}]$

①　②　③

$y=1$	N_1	样本
$y=0$	N_2	简化
$N = N_1+N_2$		

求 ϕ：$③ = \sum_{i=1}^{N} \log \phi^{y_i}(1-\phi)^{1-y_i}]$

$\frac{\partial \theta}{\partial \phi} = \sum_{i=1}^{N} y_i \cdot \frac{1}{\phi} - (1-y_i)\cdot\frac{1}{1-\phi} = 0$

$\Rightarrow \hat{\phi} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{N_1}{N}$

求 μ_1：$① = \sum_{i=1}^{N} \log N(\mu_1,\Sigma)^{y_i}$　($y_i=0$ 时约掉，因为连加)

$= \sum_{i=1}^{N} y_i \log \frac{1}{(2\pi)^{\frac{p}{2}}\cdot|\Sigma|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1))$

(μ_2 同方法)　$\mu_1 = \arg\max_{\mu_1} ① = \arg\max \sum_{i=1}^{N} y_i(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1))$

$\Delta = -\frac{1}{2}\sum_{i=1}^{N} y_i(x_i^T\Sigma^{-1}x_i - 2\mu_i^T\Sigma^{-1}x_i + \mu_1^T\Sigma^{-1}\mu_1)$

$\frac{\partial \Delta}{\partial \mu_1} = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} y_i x_i}{N_1}$

求 Σ：$\hat{\Sigma} = \arg\max_{\Sigma} ①+②$

$\sum_{i=1}^{N} y_i$

$C_1 = \{x_i \mid y_i = 1, i = 1 \cdots N\}$

$C_2 = \{x_i \mid y_i = 0, i = 1 \cdots N\}$

$|C_1| = N_1 \quad |C_2| = N_2 \quad N_1 + N_2 = N$

$① + ② = \sum_{x_i \in C_1} \log N(\mu_1, \Sigma) + \sum_{x_i \in C_2} \log N(\mu_2, \Sigma)$

$\log NC(\mu, \Sigma) = \log \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\}$

$= \log \frac{1}{(2\pi)^{\frac{p}{2}}} + \log \frac{1}{|\Sigma|^{\frac{1}{2}}} + (-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$

$= \sum_{i=1}^{N} [C - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)]$

$p \times 1 \quad 1 \times p \quad p \times p \quad p \times 1 \Rightarrow$ 实数

$= C - \frac{1}{2}N\log|\Sigma| - \frac{1}{2}\mathrm{tr}[(x_i - \mu)^T \cdot (x_i - \mu) \cdot \Sigma^{-1}]$ 其中：样本 N 的话 $S = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T$

$= C - \frac{1}{2}N\log|\Sigma| - \frac{1}{2}\mathrm{tr}(N S \cdot \Sigma^{-1})$

引入概念：

$\mathrm{tr}(AB) = \mathrm{tr}(BA)$

$\mathrm{tr}(ABC) = \mathrm{tr}(CAB) = \mathrm{tr}(BCA)$

可以任意交换次序

$\frac{\partial \mathrm{tr}(AB)}{\partial A} = B^T$

$\frac{\partial |A|}{\partial A} = |A| A^{-1}$

$\therefore ① + ② = -\frac{1}{2}N_1 \log|\Sigma| - \frac{1}{2}N_1 \mathrm{tr}(S_1 \cdot \Sigma^{-1}) - \frac{1}{2}N_2 \log|\Sigma| - \frac{1}{2}N_2 \mathrm{tr}(S_2 \Sigma^{-1}) + C$

$= -\frac{1}{2}N\log|\Sigma| - \frac{1}{2}N_1 \mathrm{tr}(S_1 \cdot \Sigma^{-1}) - \frac{1}{2}N_2 \mathrm{tr}(S_2 \Sigma^{-1}) + C$

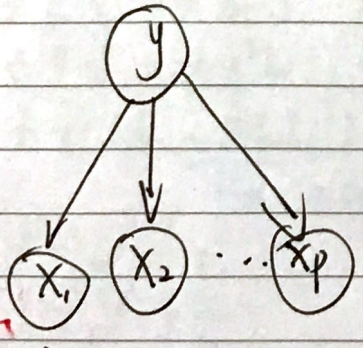
$\therefore \frac{\partial ① + ②}{\partial \Sigma} = -\frac{1}{2}(N\Sigma^{-1} - N_1 S_1 \Sigma^{-2} - N_2 S_2 \Sigma^{-2})$

$= 0$

$\therefore \hat{\Sigma} = \frac{1}{N}(N_1 S_1 + N_2 S_2)$

⑤ Naive Bayes 朴素贝叶斯：最简单的概率图模型

$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(y)P(x|y)}{P(x)} \propto P(y)P(x|y)$



x_i 与 x_j 之间被 y 阻断

$x_i \perp x_j \mid y \quad (i \neq j)$ 独立的表达

$P(x|y) = \prod_{j=1}^{p} P(x_j|y)$

x 离散 $\Rightarrow x_j \sim$ Categorial 分布

x 连续 $\Rightarrow x_j \sim N(\mu_j, \sigma_j^2)$

2分类 (0/1分类)：y 属于 Bernoulli Dist

多分类 0, 1, ... k：y 属于 Categorial Dist

做 n 次 → 二项式：Binormial

做 n 次 → 多项式分布 Maltinormial

目的：简化运算

① $x_i \in \mathbb{R}^p$，$y \in \{0, 1\}$

给定 x，y ? 0/1

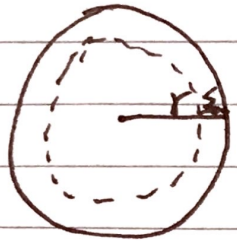
$\hat{y} = \arg\max_y P(y|x)$

$= \arg\max_{y \in \{0,1\}} \frac{P(x,y)}{P(x)} = \arg\max_y P(y) \cdot P(x|y)$

(五) 降维

① 过拟合 { 增加 Data
正则化 (加约束 L(w) to P(w))
降维 { 直接降维：特征选择
线性降维：PCA，多维空间偏移
非线性降维：流形 }

维度灾难：



r=1　$V_{超球体} = K \cdot r^D$　D维

$V_{外} = K \cdot 1^D = K$

~~$V_{内} = K - K \cdot (1-\varepsilon)^D$~~　$V_{环} = K - K \cdot (1-\varepsilon)^D$

$\frac{V_{环}}{V_{外}} = \frac{K - K(1-\varepsilon)^D}{K} = 1 - (1-\varepsilon)^D \quad 0 < \varepsilon < 1$

$\lim_{D \to +\infty} (1-\varepsilon)^D \to 0$　∴ 维度越大，比例接近于1

② Data: $X = (X_1\ X_2\ \cdots\ X_N)^T_{N \times p} = \begin{pmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{N \times p}$

样本均值：$\bar{X}_{p \times 1} = \frac{1}{N} \sum_{i=1}^{N} X_i$

样本方差：$S_{p \times p} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})^T$

$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i = \frac{1}{N} (X_1\ X_2\ \cdots\ X_N) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{N \times 1} = \frac{1}{N} X^T 1_N$

$S = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})^T$

$= \frac{1}{N} [X^T - \bar{X} 1_N^T] \cdot (X_i - \bar{X})^T$

$= \frac{1}{N} \{ X^T - \frac{1}{N} X^T 1_N 1_N^T \} \cdot (1_N - \frac{1}{N} 1_N \cdot 1_N^T)^T \cdot X$

$= \frac{1}{N} X^T (1_N - \frac{1}{N} 1_N 1_N^T) \cdot (1_N - \frac{1}{N} 1_N \cdot 1_N^T)^T \cdot X$

H_N → centering matrix

$= \frac{1}{N} X^T \cdot H \cdot H^T \cdot X$

$= \frac{1}{N} X^T H \cdot X$

$H = 1_N - \frac{1}{N} 1_N \cdot 1_N^T$

$H^T = H$（对称矩阵）

$H^2 = H \cdot H = H$

$H^n = H$

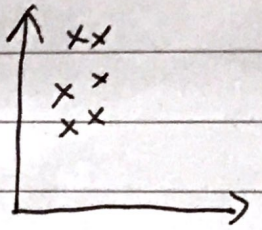
③ 一个中心：原始特征空间的重构

相关 ⟷ 无关

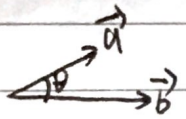
两个基本点：{ 最大投影方差 → 使投影到线上的点尽量散开
最小重构距离 }

对数据的处理：① 中心化 $x_i - \bar{x}$

先-移动



前提：



$\vec{a}\cdot\vec{b} = |\vec{a}|\cdot|\vec{b}|\cdot\cos\theta$ 令 $|\vec{b}|=1$

则 $\vec{a}$ 的投影 $|\vec{a}|\cdot\cos\theta = \vec{a}\cdot\vec{b}$

对应到矩阵为 $a^T b$ （ ）（ ）

由于中心化 → 均值 = 0

∴ 方差 $= \frac{1}{n}\sum_{i=1}^{n}((x_i-\bar{x})^T u_1)^2$ 实数可以任意交换位置 设 $u_1^T\cdot u_1 = 1$

$= \sum_{i=1}^{n}\frac{1}{n} u_1^T (x_i-\bar{x})\cdot(x_i-\bar{x})^T\cdot u_1$

$= u_1^T\left(\sum_{i=1}^{n}\frac{1}{n}(x_i-\bar{x})\cdot(x_i-\bar{x})^T\right)u_1$

$= u_1^T S\cdot u_1$

∴ 最大投影方差 $\hat{u}_1 = \arg\max u_1^T S\cdot u_1$ 前提 $u_1^T\cdot u_1 = 1$

求解时，设 $f(u_1,\lambda) = u_1^T S u_1 + \lambda(1-u_1^T u_1)$ 拉格朗日求解法

$\frac{\partial f}{\partial u_1} = S\cdot 2u_1 - \lambda 2u_1 = 0$

协方差矩阵 特征值

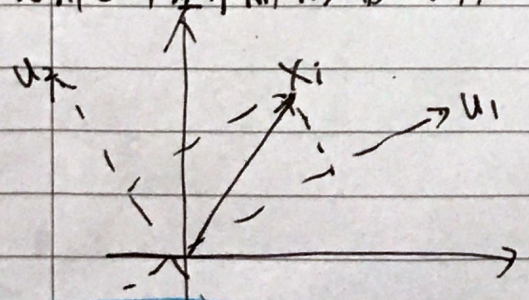
∴ $Su_1 = \lambda u_1$

特征向量

②.

对坐标空间的重构：相当于重新定义坐标轴位置

最小重构代价（2个基本点的第2点）



在 u_1 轴上的坐标值 单位向量

$x_i = (x_i^T u_1)\cdot u_1 + (x_i^T\cdot u_2)u_2$ （2维）

点积就是个横向量

$x_i = \sum_{k=1}^{p}(x_i^T u_k)\cdot u_k$ （p维）

$\Rightarrow \begin{cases} u_k = \arg\min \sum_{k=q+1}^{p} u_k^T\cdot S\cdot u_k \text{ 最小重构代价} \\ \text{s.t. } u_k^T u_k = 1 \text{ (前提)} \end{cases}$

降维后：$\hat{x}_i = \sum_{k=1}^{q}(x_i^T u_k)\cdot u_k$

$J = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \hat{x}_i\|^2$ → 相当于对坐标求平方

$= \frac{1}{n}\sum_{i=1}^{n}\|\sum_{k=q+1}^{p}(x_i^T u_k)u_k\|^2$

$= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=q+1}^{p}(x_i^T u_k)^2$ ↳ 实数

为啥投影中心也为0 $= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=q+1}^{p}((x_i-\bar{x})^T u_k)^2$

重构代价 $= \sum_{k=q+1}^{p} u_k^T S u_k$

主成分分析 ①SVD角度

中心矩阵 ← 数据矩阵 $n\times p$ → 奇异值分解

$HX = U\Sigma V^T$ → SVD $\begin{cases} U^TU=I \\ V^TV=VV^T=I \to \text{正交的} \\ \Sigma \text{ 对角} \end{cases}$ $\Sigma^T=\Sigma$ (对角矩阵)

$S=\frac{1}{n}X^THX$

$=X^TH\cdot HX = V\Sigma U^T\cdot U\Sigma V^T = V\Sigma^2V^T$

设定的 G, K

之前有: $S=GKG^T$ $\longrightarrow G=V$ $K=\Sigma^2$

设 $T=HXX^TH = U\Sigma V^T\cdot V\Sigma U^T = U\Sigma^2U^T$

$HX\cdot V = U\Sigma V^TV = U\Sigma$

$T=U\Sigma^2U^T =$

$T\cdot(U\Sigma) = U\Sigma^2U^TU\Sigma = U\Sigma^3 = (U\Sigma)\Sigma^2$

$n\times n$ T 和 S 有相同的特征(eigen value)

$p\times p$ S = 特征分解: 得到方向(主成分) $H\cdot X\cdot V$ → 坐标

T = 特征分解: 直接得到坐标

$U\Sigma$: 特征向量组成的矩阵; Σ^2: 特征值矩阵

↓ 主坐标分析: (principle coordinate analysis)

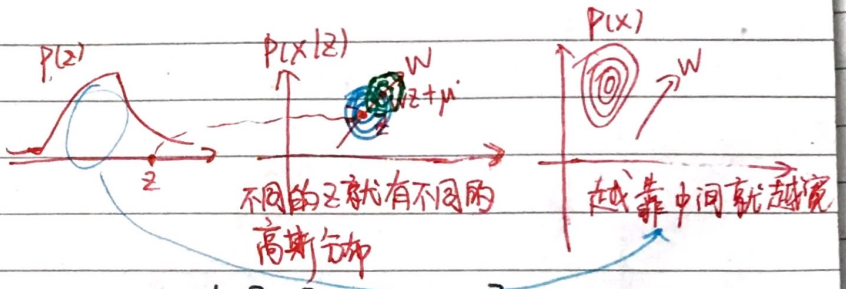
PCoA

② 概率角度. P-PCA

$x\in\mathbb{R}^p$ $z\in\mathbb{R}^q$ $q<p$

↳ latent variable

↳ observable data



假设:
- $z\sim N(0_q, I_q)$
- $x=Wz+\mu+\varepsilon$
- $\varepsilon\sim N(0, \sigma^2 I_p)$ 噪声, 对角矩阵.
- $\varepsilon \perp z$ 独立的

P-PCA:
- Inference: $P(z|x)$
- Learning: W, μ, σ^2 → EM算法 (求参数)

$E[x|z]=E[Wz+\mu+\varepsilon] = Wz+\mu$

$Var[x|z] = Var[Wz+\mu+\varepsilon) = \sigma^2 I$

$x|z \sim N(Wz+\mu, \sigma^2 I)$

$E[x] = E[Wz+\mu+\varepsilon] = \mu$

$Var[x] = Var[Wz+\mu+\varepsilon] = Var[Wz]+Var[\varepsilon]$

$= W\cdot I\cdot W^T+\sigma^2 I = WW^T+\sigma^2 I$

$x\sim N(\mu, WW^T+\sigma^2)$

z, 且 $x|z$, $x\ z|x$

那么 $\begin{pmatrix}x\\z\end{pmatrix} \sim N\left(\begin{bmatrix}\mu\\0\end{bmatrix}, \begin{bmatrix}WW^T+\sigma^2 & W\\ W & I\end{bmatrix}\right)$

求 $z|x$ (高斯第6节内容)

$Cov(x,z) = E[(x-\mu)(z-0)^T] = E[(Wz+\varepsilon)z^T]$

$= E[Wzz^T] = WE[zz^T] = W$

~~$z|x$~~ $x=\begin{pmatrix}x_a\\x_b\end{pmatrix}$ $x\sim N\left(\begin{bmatrix}\mu_a\\\mu_b\end{bmatrix}, \begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}\right)$

$x_{b\cdot a} = x_b - \Sigma_{ba}\Sigma_{aa}^{-1}x_a$

$\mu_{b\cdot a} = \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a$

$\Sigma_{bb\cdot a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$

$x_{b\cdot a}\sim N(\mu_{b\cdot a}, \Sigma_{bb\cdot a})$

又 $x_b = x_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a$

$x_b|x_a \sim N(\mu_{b\cdot a}+\cdots, \Sigma_{bb\cdot a})$

$E[x_b|x_a] = E[x_{b\cdot a}] + \Sigma_{ba}\cdot\Sigma_{aa}^{-1}x_a = \mu_{b\cdot a}+\Sigma_{ba}\cdot\Sigma_{aa}^{-1}x$

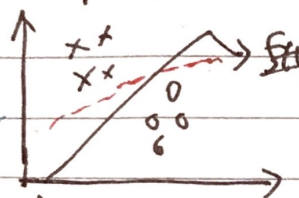
$= \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a-\mu_a)$

$Var[x_b|x_a] = Var[x_{b\cdot a}] = \Sigma_{bb\cdot a}$

第六讲 支持向量机 SVM 有三宝：间隔，对偶，核技巧

SVM
- hard-margin SVM 硬间隔
- soft-margin SVM 软间隔
- kernel SVM 约束优化

SVM 最初是为了解决2分类问题



能正确合开的有无数条

⇓

SVM：选出最好的一条

$f(w)=\text{sing}(w^Tx+b)$ 符号函数

① 最大间隔分类器 $\max \text{margin}(w,b)$

$$\text{st.}\begin{cases} w^Tx_i+b>0 & y_i=+1 \\ w^Tx_i+b<0 & y_i=-1 \end{cases} \Rightarrow y_i(w^Tx_i+b)>0 \quad i=1,2\cdots N$$

~~$y_i(w^Tx_i+b)<0$~~ 同号

定 $\text{margin}(w,b)=\min_{w,b,x_i} \text{distance}(w,b,x_i)=\min_{\substack{w,b,x_i \\ i=1\cdots N}} \frac{1}{\|w\|}|w^Tx_i+b|$

点到直线的距离

(x_i,y_i)

distance → margin 找到所 distance 中最小的 distance

$$\text{distance}=\frac{1}{\|w\|}\cdot|w^Tx_i+b|$$

$$\therefore \begin{cases} \max \text{margin}(w,b)=\max_{w,b}\min_{x_i}\frac{1}{\|w\|}|w^Tx_i+b| \\ \text{st. } y_i(w^Tx_i+b)>0 \end{cases}$$

去掉绝对值：$y_i(w^Tx_i+b)$

$$\max_{w,b}\min_{x_i}\frac{1}{\|w\|}y_i(w^Tx_i+b)$$

$$=\max_{w,b}\frac{1}{\|w\|}\left(\min_{\substack{x_i \\ i=1,2\cdots N}} y_i(w^Tx_i+b)\right)$$

$r=1$（令 $r=1$，简化运算，因为 w^Tx_i+b 直线可以 w 任意比例缩放）

$$=\max_{w,b}\frac{1}{\|w\|}=\min_{w,b}\|w\|$$

$$\begin{cases} \min_{w,b}\|w\| \\ \text{st. } \min y_i(w^Tx_i+b)=1 \end{cases}$$

可写成 $y_i(w^Tx_i+b)\geq 1$

如何→变成→规化问题

$$\Rightarrow \begin{cases} \min_{w,b}\frac{1}{2}w^Tw & \text{→二次} \\ \text{st. } y_i(w^Tx_i+b)\geq 1 \quad \text{for } \forall i=1\cdots N & \text{N个约束} \end{cases}$$

求导简化

接上面：

写成拉格朗日：

$$\mathcal{L}(w,b,\lambda)=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b))$$

($\lambda_i\geq 0$，$1-y_i(w^Tx_i+b)\leq 0$)

问题转化为：$\begin{cases}\min\limits_{w,b}\max\limits_{\lambda}\mathcal{L}(w,b,\lambda)\\ s.t.\ \lambda_i\geq 0\end{cases}$

(从带约束 ↓ 无约束)

若：$1-y_i(w^Tx_i+b)>0$ $\max\to\infty$

若 $1-y_i(w^Tx_i+b)<0$ $\max$ 则一定存在 $\max=\frac{1}{2}w^Tw+0$

$$\min_{w,b}\left(\frac{1}{2}w^Tw,\infty\right)=\min_{w,b}\frac{1}{2}w^Tw$$

对偶问题：$\begin{cases}\max\limits_{\lambda}\min\limits_{w,b}\mathcal{L}(w,b,\lambda)\\ s.t.\ \lambda_i\geq 0\end{cases}$

$\min\max\mathcal{L}\geq\max\min\mathcal{L}$ (宁为凤尾，不为鸡头)

弱对偶关系

$\min\max\mathcal{L}=\max\min\mathcal{L}$

强对偶关系

对于 $\min\mathcal{L}(w,b,\lambda)$

$$\frac{\partial\mathcal{L}}{\partial b}\triangleq 0\Rightarrow\sum_{i=1}^{N}\lambda_iy_i=0$$ 将其代入 $\mathcal{L}(w,b,\lambda)$

$$\mathcal{L}(w,b,\lambda)=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i-\sum_{i=1}^{N}\lambda_iy_iw^Tx_i$$

$$\frac{\partial\mathcal{L}}{\partial w}=\frac{1}{2}\cdot 2\cdot w-\sum_{i=1}^{N}\lambda_iy_ix_i\triangleq 0\Rightarrow w=\sum_{i=1}^{N}\lambda_iy_ix_i$$

将其代入 $\mathcal{L}(w,b,\lambda)$

有：$$\mathcal{L}(w,b,\lambda)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_jy_iy_jx_i^Tx_j+\sum_{i=1}^{N}\lambda_i$$

↓ 相当于是 min 了 (因为是最小值的最优解的值)

再求：$\max\limits_{\lambda}-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_jy_iy_jx_i^Tx_j+\sum_{i=1}^{N}\lambda_i$

$s.t.\ \lambda_i\geq 0$

$\sum_{i=1}^{N}\lambda_iy_i=0$ 也是一个向量

KKT 条件：$\frac{\partial\mathcal{L}}{\partial w}=0$ $\frac{\partial\mathcal{L}}{\partial b}=0$

$\lambda_i(1-y_i(w^Tx_i+b))=0$

$\lambda_i\geq 0$

$1-y_i(w^Tx_i+b)\leq 0$

是数据 data 相关的线性组合

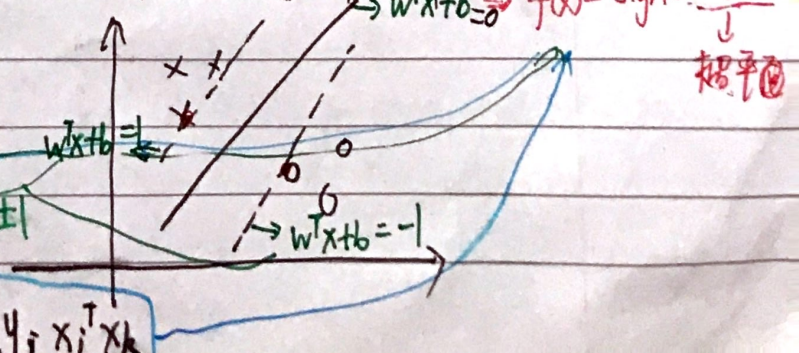
$w^*=\sum_{i=1}^{N}\lambda_iy_ix_i$

不在2条线上的点对 w 的求解都无意义

计算 b：$\exists(x_k,y_k)$，$s.t.\ 1-y_k(w^Tx_k+b)=0$

$$b^*=y_k-w^Tx_k=y_k-\sum_{i=1}^{N}\lambda_iy_ix_i^Tx_k$$

原、对偶问题具有强对偶关系 $\Leftrightarrow$ 满足 KKT 条件



② soft-margin SVM 软间隔分类

$$\begin{cases} \min_{w,b} \frac{1}{2}w^Tw \\ s.t.\ y_i(w^Tx_i+b) \geq 1 \\ i=1\cdots N \end{cases}$$

hard 是基于样本本身就是可分的，soft：数据不可分/或存在噪声

soft 思想：允许一点错误

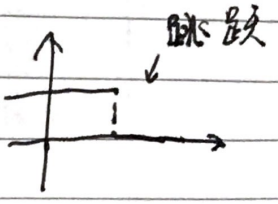
$\min \frac{1}{2}w^Tw + loss$

① $loss = \sum_{i=1}^{N} I\{y_i(w^Tx_i+b)<1\}$

但由此时 w 不连续

令 $z = y(w^Tx+b)$

$loss\ 0/1 = \begin{cases} 1 & z<1 \\ 0 & otherwise \end{cases}$

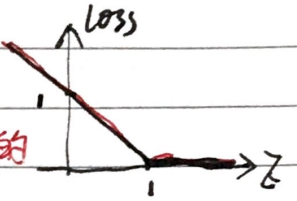


② loss：距离 → hinge loss

若 $y_i(w^Tx_i+b) \geq 1$，$loss = 0$

若 ------ <1，$loss = 1 - y_i(w^Tx_i+b)$

$loss = \max\{0, 1-y_i(w^Tx_i+b)\}$ → 函数本身是连续的

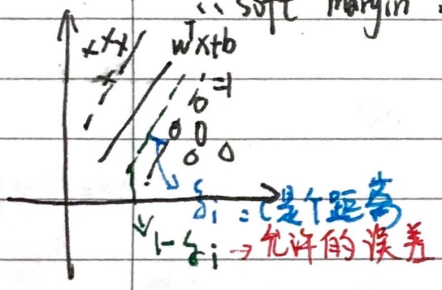


∴ soft margin：$\min_{w,b} \frac{1}{2}w^Tw + C\sum_{i=1}^{N} \max\{0, 1-y_i(w^Tx_i+b)\}$

一般不用 max：

引入 $\xi_i = 1-y_i(w^Tx_i+b)$，$\xi_i \geq 0$

$$\begin{cases} \min_{w,b} \frac{1}{2}w^Tw + C\sum_{i=1}^{N}\xi_i \\ s.t.\ y_i(w^Tx_i+b) \geq 1-\xi_i \end{cases}$$



③ 约束优化问题

等价的

拉格朗日函数：$\zeta(x,\lambda,\eta) = f(x) + \sum_{i=1}^{M}\lambda_i m_i + \sum_{j=1}^{N}\eta_j n_j$

(原问题的无约束形式) SVM：$\begin{cases} \min_x \max_{\lambda,\eta} \zeta(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0 \end{cases}$ ⇒ 自动排除了不好的 x 的集合 (好：$m_i(x)<0$)

(原是关于 x 的函数)

对偶问题：$\begin{cases} \max_{\lambda,\eta} \min_x \zeta(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0 \end{cases}$

(对…是关于 λ,η 的函数)

证明弱对偶性：$\max\min \zeta \leq \min\max \zeta$

$$\underbrace{\min_x \zeta}_{A(\lambda,\eta)} \leq \zeta(x,\lambda,\eta) \leq \underbrace{\max_{\lambda,\eta} \zeta(x,\lambda,\eta)}_{B(x)}$$

$A(\lambda,\eta) \leq B(x)$

$\max A(\lambda,\eta) \leq \min B(x)$

即 $\max_{\lambda,\eta}\min_x \zeta \leq \min_x \max_{\lambda,\eta} \zeta$

④ 对偶性的几何解释

凸优化问题 $\begin{cases} \min\limits_{x\in D} f(x) \\ s.t.\ m_i(x)\le 0 \end{cases}$ D：定义域 $x\in D$

拉格朗日问题 $\phi(x,\lambda)=f(x)+\lambda m_i(x)$ ，$\lambda\ge 0$

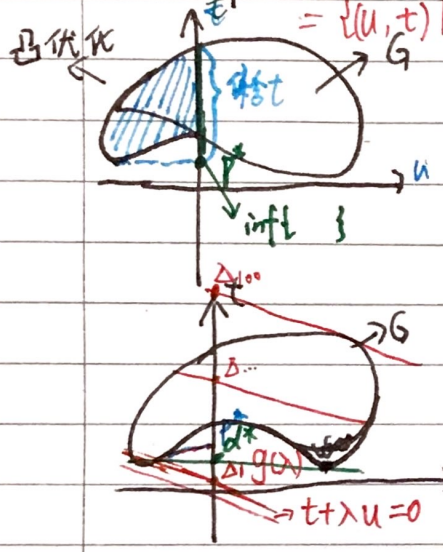
$p^*=\min f(x)$ 原问题最优解

$d^*=\max\limits_{\lambda}\min\limits_{x}\phi(x,\lambda)$ 对偶最优解

引入集合的定义：

$G=\{(m_i(x),f(x))\mid x\in D\}$ 二维坐标系的一个区域

$=\{(u,t)\mid x\in D\}$



此时：$p^*=\min f(x)=\min t$

$p^*=\inf\{t\mid (u,t)\in G, u\le 0\}$ → 下确界（集合中）= min的意思

$d^*=\max\limits_{\lambda}\min(t+\lambda u)$

$g(\lambda)$

$=\max\limits_{\lambda} g(\lambda)$

其中 $g(\lambda)=\inf\{t+\lambda u\mid (u,t)\in G\}$

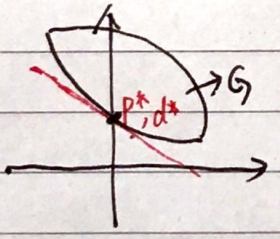
往上平移：最先切的点就是 $g(\lambda)$

$d^*=\max\limits_{\lambda} g(\lambda)$：变动斜率，使同时切到两处

此时的 $g(\lambda)$ 是 $\max\limits_{\lambda} g(\lambda)$

有：$d^*\le p^*$

若：G的集合为：



此时的 $p^*=d^*$

凸优化 + slater条件 $\Rightarrow d^*=p^*$ （强对偶）

并不是充要。（充分非必要，还是有其他情况推出强对偶）

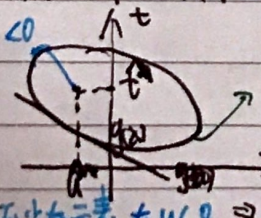
⑤ slater条件：

$\exists \hat{x}\in \text{relint } D$ relint：相对内部 例：(D) 去掉边界的D

$s.t.\ \forall i=1\sim m,\ m_i(\hat{x})<0$

I. 对于大多数凸优化，slater成立

II. 放松slater：若M中有k个仿射函数，m-k



凸二次规化问题天然满足I或II

②至少有元素，在u<0 ⇒ 便得有gp这样的切线。

⑥ KKT条件

$p^* \to x^*$

$d^* \to \lambda^*; \eta^*$

KKT条件给出了 x^*, λ^*, η^* 的一种关系

$$\begin{cases} m_i(x^*) \le 0 \\ n_j(x^*) = 0 \\ \alpha^* \ge 0 \end{cases}$$ 必须满足

KKT：
- 可行条件
- 互补松弛：$\lambda_i^* m_i = 0 \quad \forall i = 1, 2, \cdots m$
- 梯度为0：$\left.\frac{\partial f(x, \lambda^*, \eta^*)}{\partial x}\right|_{x=x^*} = 0$

Ⅱ. 互补松弛：

$$d^* = \max_{\lambda,\eta} g(\lambda, \eta) = g(\lambda^*, \eta^*)$$
$$= \min_x \mathcal{L}(x, \lambda^*, \eta^*) \le \mathcal{L}(x, \lambda^*, \eta^*) \quad \forall x \text{ 成立，则 } x^* \text{ 也能}$$
$$\le \mathcal{L}(x^*, \lambda^*, \eta^*)$$
$$= f(x^*) + \sum_{i=1}^{m} \lambda_i^* m_i + \sum_{j=1}^{n} \eta_j^* n_j \quad (n_j = 0) \to \text{只能取等}$$
$$\le f(x^*) \quad (\sum \lambda_i^* m_i \le 0)$$
$$= p^* \quad (\sum \lambda_i^* m_i = 0, \ \lambda_i \ge 0, \ m_i \le 0) \quad \forall i = 1, 2 \cdots, m$$

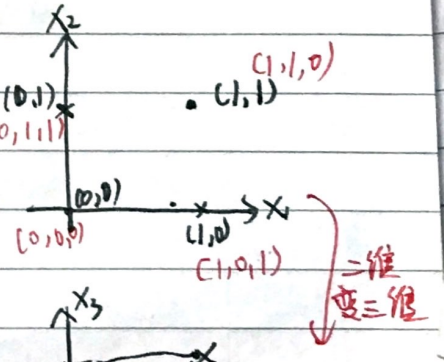
不等号变等号 ⟺ $\sum \lambda_i = 0$ 或 $m_i = 0$

Ⅲ. 梯度为0 ：（$x = x^*$ 时取最小值）

$$\min \mathcal{L}(x, \lambda^*, \eta^*) = \mathcal{L}(x^*, \lambda^*, \eta^*)$$
$$\left.\frac{\partial \mathcal{L}}{\partial x}\right|_{x=x^*} = 0$$

⑦ 核方法 Kernel Function

线性可分	一点错误	严格非线性
PLA	Pocket Algorithm	$\phi(x)$ + PLA
Hard-margin SVM	Soft-margin SVM	$\phi(x)$ + Hard-margin



Ⅰ. 若数据不能线性可分（例如圆等等）

Ⅰ. PLA → 多层感知机（神经网络）→ deep learning

Ⅱ. 非线性可分问题 → 转化为线性可分问题

$\phi(x)$ = 非线性转换

例：$X \xrightarrow{\phi(x)} Z$

input space　　feature space

$X = (x_1, x_2)$ = 维

$\phi(x)$

$Z = (x_1, x_2, (x_1 - x_2)^2)$ 三维

Cover Theorem：高维比低维更易线性可分

Ⅱ. 对偶问题：对偶表示带来内积

Primal problem：

$$\begin{cases} \min_{w,b} \frac{1}{2} w^T w \\ s.t.\ y_i(w^T x_i + b) \geq 1 \quad (N\text{个约束}) \end{cases}$$

Dual Problem

$$\begin{cases} \min \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \underbrace{x_i^T x_j}_{\text{内积}} - \sum_{i=1}^{N} \lambda_i \\ s.t.\ \lambda_i \geq 0 \quad \text{for all } i = 1, 2 \cdots N \\ \sum_{i=1}^{N} \lambda_i y_i = 0 \end{cases}$$

变成 $\phi(x_i)^T \phi(x_j)$，但 $\phi(x_i)$ 在现实情况下维度可能会非常高.

Kernel Function：

$$K(x, x') = \underbrace{\phi(x)^T \phi(x')}_{\text{input space}} = \underbrace{\langle \phi(x), \phi(x') \rangle}_{\phi(x)\text{函数 feature space}}$$

$\forall x, x' \in \mathcal{X}$，$\exists \phi: \mathcal{X} \to \mathcal{Z}$

s.t. $K(x, x') = \phi(x)^T \phi(x')$

则称 $K(x, x')$ 是一个核函数

例：假设已经找到一个核函数 $K(x, x') = \exp\left(-\frac{(x-x')^2}{2\sigma^2}\right)$

则不用求 $\phi(x)$，则可直接得到内积.

② 正定核的定义

通常情况下的核函数为正定核函数

核：$K: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$\forall x, z \in \mathcal{X}$，则称 $K(x, z)$ 为核函数

正定核： $K: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$\forall x, z \in \mathcal{X}$，有 $K(x, z)$

若 $\exists$ 一个 $\phi(\cdot): \mathcal{X} \mapsto \mathbb{R}$，$\phi \in \mathcal{H}$（希尔伯特空间）

s.t. $K(x, z) = \langle \phi(x), \phi(z) \rangle$

那么则称 $K(x, z)$ 为正定核函数

正定核： $K: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$\forall x, z \in \mathcal{X}$，有 $K(x, z)$

若 $K(x, z)$ 满足如下两条性质：

Ⅰ. 对称性

Ⅱ. 正定性

则称 $K(x, z)$ 为正定核

①对称性 $\Leftrightarrow K(x, z) = K(z, x)$

②正定性 $\Leftrightarrow$ 任取 N 个元素 $x_1, x_2, \cdots x_N \in \mathcal{X}$

对应的 Gram 矩阵 matrix 是半正定的 $K = [K(x_i, x_j)]$

要证： $K(x, z) = \langle \phi(x), \phi(z) \rangle$ 等价 $\Leftrightarrow$ Gram matrix 半正定

Hilbert Space：完备的，可能是无限维的，被赋予内积的线性空间

向量空间（加法和数乘）

对称性，正定性，线性

$\langle f, g \rangle = \langle g, f \rangle$，$\langle f, f \rangle \geq 0$，"=" $\Leftrightarrow f = 0$

$\langle r_1 f_1 + r_2 f_2, g \rangle = r_1 \langle f_1, g \rangle + r_2 \langle f_2, g \rangle$

$\{K_n\}$ 对极限是封闭的 $\lim_{n \to \infty} K_n = K \in \mathcal{H}$

正定核一必要性证明 (⇒)

已知 $k(x,z)=\langle\phi(x),\phi(z)\rangle$，证：Gram matrix 半正定，且 $k(x,z)$ 对称

证：$k(x,z)=\langle\phi(x),\phi(z)\rangle$

$k(z,x)=\langle\phi(z),\phi(x)\rangle$

又∵ 内积运算本身具有对称性，即 $\langle\phi(x),\phi(z)\rangle=\langle\phi(z),\phi(x)\rangle$

∴ $k(x,z)=k(z,x)$

∴ $k(x,z)$ 满足对称性

欲证 Gram matrix 半正定 ------> 方法：① 特征值 ≥ 0

Gram matrix：$K=[k(x_i,x_j)]_{N\times N}$ ② $\forall\alpha\in\mathbb{R}^n$，$\alpha^T A\cdot\alpha\geq 0$

即证，对 $\forall\alpha\in\mathbb{R}^N$，$\alpha^T K\alpha\geq 0$

$$\alpha^T\cdot K\alpha=(\alpha_1\ \alpha_2\ \cdots\ \alpha_N)\begin{pmatrix}k_{11}&k_{12}&\cdots&k_{1N}\\k_{21}&k_{22}&\cdots&k_{2N}\\\vdots&\vdots&&\vdots\\k_{N1}&k_{N2}&\cdots&k_{NN}\end{pmatrix}\begin{pmatrix}\alpha_1\\\alpha_2\\\vdots\\\alpha_N\end{pmatrix}$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j k_{ij}$$

$k(x_i,x_j)$ 就是内积

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\langle\phi(x_i),\phi(x_j)\rangle$$

$$=\alpha_i\sum_{i=1}^{N}\alpha_i\phi(x_i)^T\cdot\sum_{j=1}^{N}\alpha_j\phi(x_j)$$

$$=\left[\sum_{i=1}^{N}\alpha_i\phi(x_i)\right]^T\cdot\sum_{j=1}^{N}\alpha_j\phi(x_j)$$

$$=\left\langle\sum_{i=1}^{N}\alpha_i\phi(x_i),\sum_{j=1}^{N}\alpha_j\phi(x_j)\right\rangle$$

$$=\left\|\sum_{i=1}^{N}\alpha_i\phi(x_i)\right\|^2\geq 0$$

∴ K 是半正定的

∴ 必要性得证

(八) 指数族分布

① $p(x|\eta) = h(x)\exp(\eta^T \phi(x) - A(\eta))$

线性 ; 充分统计量 ⇒ 样本的一个函数

例如 样本均值

η: 参数(向量), $x \in \mathbb{R}^p$

$A(\eta)$: log partition function

log 配分函数

例: $p(x|\theta) = \frac{1}{Z}\hat{p}(x|\theta)$

归一化因子

$\int p(x|\theta)dx = \int \frac{1}{Z}\hat{p}(x|\theta)dx$

$= \frac{1}{Z}\int \hat{p}(x|\theta)dx$

$\therefore Z = \int \hat{p}(x|\theta)dx$

$p(x|\eta) = h(x)\cdot \exp(\eta^T\phi(x))\exp(-A(\eta))$

$= \frac{1}{\exp(A\eta)}\, h(x)\cdot\exp(\eta^T\phi(x))$

$= \frac{1}{Z}\cdot \hat{p}(x|\eta)$

$\therefore Z = \exp(A\eta)$

$A(\eta) = \log Z$

充分统计量 共扼 最大熵(无信息先验)

指数族分布

变分推断 概率图模型 → 无向图 = RBM

广义线性模型 ← link function → (激活函数)^T

广义线性回归 W^Tx

Ⅱ. 指数族分布: $y|x \sim$ 指数族分布

线性回归: $y|x \sim N(\mu, \sigma)$

共扼: $y|x \sim$ Bernoulli, $y|x \sim$ possion 等

因为它是指数族, 所以有这个性质

$P(z|x) \propto P(x|z)\cdot P(z)$

后验 先验

共扼的效果: 先验和后验有个相同的分布形式

例: Beta 伯二项式 Beta

prior

① 共扼 → 计算方便

② 最大熵 → 无信息先验

③ Jerrif

②

EM算法：EM期望最大

MLE：$P(x|\theta)$

$\theta_{MLE} = \arg\max_\theta P(x|\theta)$

→ 简化 $\arg\max_\theta \log P(x|\theta)$ log likelihood

① $\theta^{(t+1)} = \arg\max_\theta \int_z \log P(x,z|\theta) \cdot P(z|x,\theta^{(t)})\, dz$ 证明是 θ^{t+1} 与 θ^t 之间的关系

完整数据　后验

$E_{z|x,\theta^{(t)}}[\log P(x,z|\theta)]$

② 会有 $\theta^{(t)} \to \theta^{(t+1)}$

$\log P(x|\theta^{(t)}) \le \log P(x|\theta^{(t+1)})$ → 证明这个等式：

收敛性

$\log P(x|\theta) = \log P(x,z|\theta) - \log P(z|x,\theta)$

同时求期望：左边 $= \int_z P(z|x,\theta^{(t)}) \log P(x|\theta)\, dz$

$= \log P(x|\theta) \int_z P(z|x,\theta^{(t)})\, dz$ → 只跟 z 相关，所以积分=1

$= \log P(x|\theta)$

右边 $= \int_z P(z|x,\theta^{(t)}) \cdot \log P(x,z|\theta)\, dz - \int_z P(z|x,\theta^{(t)}) \cdot \log P(z|x,\theta)\, dz$

$Q(\theta,\theta^{(t)})$　$H(\theta,\theta^{(t)})$

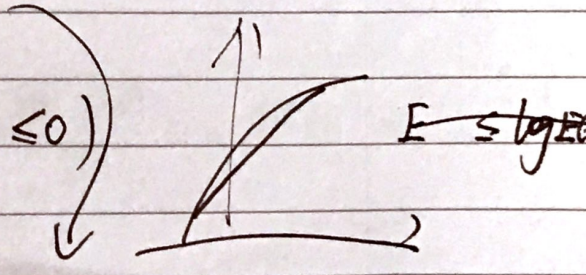
要证：$\begin{cases} Q(\theta^{(t+1)},\theta^{(t)}) \ge Q(\theta^{(t)},\theta^{(t)}) \\ H(\theta^{(t+1)},\theta^{(t)}) \le H(\theta^{(t)},\theta^{(t)}) \end{cases}$

$H(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)})$

$= \int_z P(z|x,\theta^{(t)}) \log P(z|x,\theta^{(t+1)})\, dz - \int_z P(z|x,\theta^{(t)}) \log P(z|x,\theta^{(t)})\, dz$

$= \int_z P(z|x,\theta^{(t)}) \log \frac{P(z|x,\theta^{(t+1)})}{P(z|x,\theta^{(t)})}\, dz$

$(= -KL(P(z|x,\theta^{(t)}) \| P(z|x,\theta^{(t+1)})) \le 0)$



$E[\log X] \le \log E(X)$

$\therefore \log \int_z P(z|x,\theta^{(t+1)})\, dz = \log$

$= 0$

证明收敛性

③ 证明公式从何

重要点

② 证明公式从何而来：

$$\log P(X|\theta) = \log P(X,Z|\theta) - \log P(Z|X,\theta)$$
$$= \log \frac{P(X,Z|\theta)}{q(Z)} - \log \frac{P(Z|X,\theta)}{q(Z)}$$ （log 的一减一加后恒等）

两边求积分：

左边 $= \int_Z q(Z)\cdot \log P(X|\theta)\, dZ = \log P(X|\theta)\int_Z q(Z)\cdot dZ = \log P(X|\theta)$

右边 $= \int_Z q(Z) \log \frac{P(X,Z|\theta)}{q(Z)}\, dZ - \int_Z q(Z)\cdot \log \frac{P(Z|X,\theta)}{q(Z)}\, dZ$

ELBO：evidence lower bound　　相对熵：$KL(q(Z)\,||\,P(Z|X,\theta))$

$$\log P(X|\theta) = ELBO + KL(q\,||\,p)$$

$KL(q\,||\,p) \geq 0$ → 后验

$\therefore \log P(X|\theta) \geq ELBO$ [当 q 与 p 相等时，取等号]

$\therefore \hat{\theta} = \arg\max_\theta ELBO$

$= \arg\max_\theta \int q(Z) \log \frac{P(X,Z|\theta)}{q(Z)}\, dZ$ （取等号时，$q(Z) = P(Z|X,\theta^{(t)})$）

$= \arg\max_\theta \int q(Z|X,\theta^{(t)}) \log \frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t)})}\, dZ$

$\frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t)})}$ → 与 θ 无关（$\theta^{(t)}$ 是常数）

$= \arg\max_\theta \int P(Z|X,\theta^{(t)}) \log P(X,Z|\theta)\, dZ$ → EM 公式

③ 从另一个角度推公式　　X：observed data　$X = \{x_1, x_2 \cdots x_N\}$

$\log P(X|\theta) = \log \int_Z P(X,Z|\theta)\, dZ$ 直接引入 Z，所以求积分

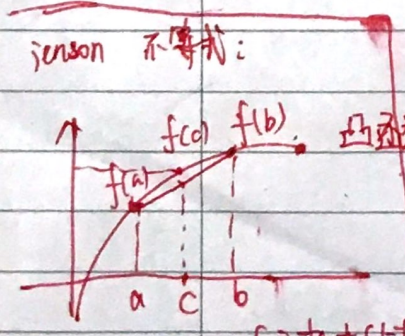
$= \log \int_Z \frac{P(X,Z|\theta)}{q(Z)} \cdot q(Z)\cdot dZ$ （第一讲：积分可理解为求其期望）

$= \log E_{q(Z)}\left[\frac{P(X,Z|\theta)}{q(Z)}\right]$

$\geq E_{q(Z)}\left[\log \frac{P(X,Z|\theta)}{q(Z)}\right]$ 在 $\frac{P(X,Z|\theta)}{q(Z)} = C$ 时取 "="

后验 → ELBO

Jenson 不等式：



$c = ta + (1-t)b$

$f(c) = f(ta + (1-t)b) \geq t f(a) + (1-t) f(b)$ (特点)

$q(Z) = \frac{1}{C} P(X,Z|\theta)$

$1 = \int q(Z)\, dZ = \int_Z \frac{1}{C} P(X,Z|\theta)\, dZ$

因为是概率密度函数

$= \frac{1}{C}\int_Z P(X,Z|\theta)\, dZ$

$= \frac{1}{C} P(X|\theta)$

$P(X|\theta) = C$

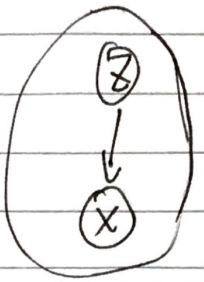
$\therefore q(Z) = \frac{1}{P(X|\theta)} \cdot P(X,Z|\theta) = P(Z|X,\theta)$

后验

HMM / GMM → 模型　GD → 算法，解决优化问题的一个迭代算法
EM → 算法，解决估计问题的一个迭代算法

① 从狭义 EM → 广义 EM
② 为什么狭义 EM 是 EM 的一特
③ 真正的 EM，EM 的变种，问题等

没：模型：概率生成模型
Data : X
latent variable → Z
θ : model parameter



$P(X|\theta)$
EM是用于估计参数 θ 的

$\hat{\theta} = \arg\max P(X|\theta)$
$= \arg\max \log P(X|\theta)$ → 为了方便起见
P(X) 未知，所以才有归纳偏置

引入 P(X, Z)

$P(X) = \int_Z P(X, Z)\, dZ$ 全概率

$P(X) = \frac{P(X,Z)}{P(Z,X)}$

其中 $\log P(X|\theta) = ELBO + KL(q\|p) \geq \mathcal{L}(q,\theta)$

$\log P(X) = \log \frac{P(X,Z)}{P(Z,X)}$

$= \log P(X,Z) - \log P(Z,X)$

$= \log \frac{P(X,Z)}{q(Z)} - \log \frac{P(Z|X)}{q(Z)}$

$$\begin{cases} ELBO = E_{q(z)}\left[\log \frac{P(X,z|\theta)}{q(z)}\right] = \mathcal{L}(q,\theta) \\ KL(q\|p) = \int q(z) \cdot \log \frac{q(z)}{P(z|X,\theta)}\, dz \end{cases}$$

KL ≥ 0，q = p 时取等号

但有时 P(z|X, θ) 有时无法计算出

广义 EM：

E-step：θ 固定时，q → p，KL 越小 ⇔ ELBO 越大

$\hat{q} = \arg\min_q KL(q\|p) = \arg\max_q \mathcal{L}(q,\theta)$

在右面两边同时求期望：

左边：$E_{q(z)}[\log P(X)] = \int q(z) \log P(X)\, dz = \log P(X)$

右边：$\int q(z) \cdot \log \frac{P(X,z)}{q(z)}\, dz - \int q(z) \log \frac{P(z,X)}{q(z)}\, dz$

ELBO　　KL(q‖p)

M-step：固定 $\hat{q}$ 时，$\theta = \arg\max_\theta \mathcal{L}(\hat{q}, \theta)$

更一般规范：

广义 EM：

E-step：$q^{(t+1)} = \arg\max_q \mathcal{L}(q, \theta^{(t)})$

M-step：$\theta^{(t+1)} = \arg\max_\theta \mathcal{L}(q^{(t+1)}, \theta)$

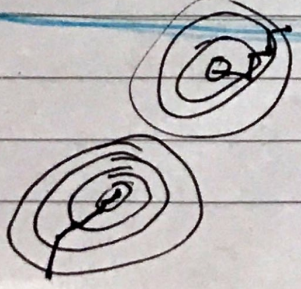
$\mathcal{L}(q,\theta) = E_q[\log P(X,z) - \log q]$

$= E_q[\log P(X,z)] - E_q[\log q]$

$= E_{q(z)}[\log P(X,z|\theta)] + H(q(z))$

熵：$\int q(z) \cdot \frac{1}{\log(q(z))}\, dz$

坐标上升法：



梯度上升法

M-E 的可以广泛所谓

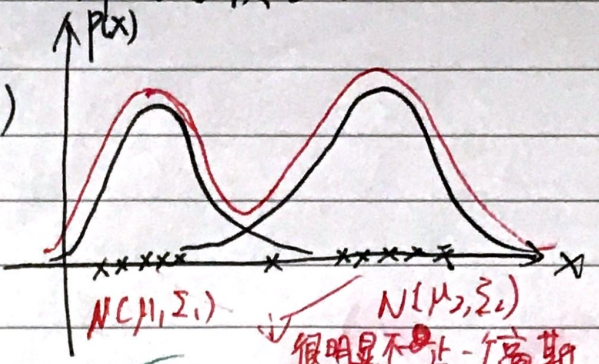
(十一) GMM : Gaussian mixture Model 高斯混合模型.

① 从几何角度来看：

关于多个高斯的加权平衡（叠加）

$p(x)=\sum_{k=1}^{K}\alpha_k N(\mu_k,\Sigma_k)$，$\sum_{k=1}^{K}\alpha_k=1$

α_k：权重



很明显不止一个高斯

② 从混合模型角度来看：（生成模型）

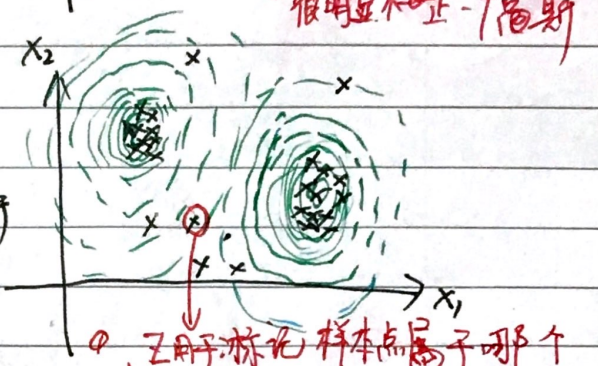
X：observed variable

引进 Z：latent variable：对应的样本X是属于哪一个高斯分布

↳ 离散的随机变量

Z	1	2	⋯	K
p(z)	p_1	p_2	⋯	p_K

$\sum_{k=1}^{K} p_k=1$



① 假设有一个骰子，有六个面（例子）

每个面都有概率，概率更大的更重

Z的子标记样本点属于哪个高斯分布

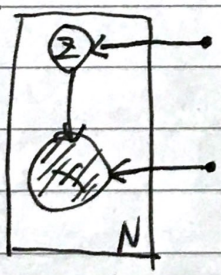
②有的点，2个都属于，但得看离哪个近

X ∩ C_1（概率更大）

C_2（概率更小）

X ∩ Z.

二. 概率图：



$p=\{p_1,p_2,\cdots p_K\}$ （p是Z的参数）

(μ_k,Σ)

$$p(x)=\sum_{z} p(x,z)$$
$$=\sum_{k=1}^{K} p(x,z=C_k)$$
$$=\sum_{k=1}^{K} p(z=C_k)\cdot p(x|z=C_k)$$
$$=\sum_{k=1}^{K} p_k\cdot N(x|\mu_k,\Sigma_k)$$

p_k：概率

θ = parameter → $\theta=\{p_1\ p_2\cdots p_N\ \mu_1\mu_2\cdots\mu_N\ ,\ \Sigma_1\Sigma_2\cdots\Sigma_k\}$

$\theta_{MLE}=\arg\max_{\theta}\log p(x)$ → 极大释然估计无法求得

⑧ 用EM求解GMM

EM: $\theta^{(t+1)} = \arg\max E_{Z|X,\theta^{(t)}}[\log P(X,Z|\theta)]$

Q函数 $Q(\theta,\theta^{(t)})$

$Q(\theta,\theta^{(t)}) = \int_Z \log P(X,Z|\theta)\cdot P(Z|X,\theta^{(t)})\,dZ$

Z是离散的

E-step

$= \sum_Z \log \prod_{i=1}^N P(X_i,Z_i|\theta)\cdot \prod_{i=1}^N P(Z_i|X_i,\theta^{(t)})$

$= \sum_{Z_1Z_2\cdots Z_N} \sum_{i=1}^N \log P(X_i,Z_i|\theta) \prod_{i=1}^N P(Z_i|X_i,\theta^{(t)})$

$= \sum_{Z_1Z_2\cdots Z_N} [\log P(X_1,Z_1|\theta)+\log(X_2,Z_2|\theta)+\cdots+\log P(X_N,Z_N|\theta)] \prod_{i=1}^N P(Z_i|X_i,\theta^{(t)})$

⑧ $\because \sum_{Z_1Z_2\cdots Z_N} \log P(X_1,Z_1|\theta) \prod_{i=1}^N P(Z_i|X_i,\theta^{(t)})$

$= \sum_{Z_1} \log P(X_1,Z_1|\theta)\, P(Z_1|X_1,\theta^{(t)})$

$\therefore$ 原式 $= \sum_{Z_1} \log P(X_1,Z_1|\theta)\cdot P(Z_1|X_1,\theta^{(t)}) + \cdots + \sum_{Z_N} \log P(X_N,Z_N|\theta) P(Z_N|X_N,\theta^{(t)})$

$= \sum_{i=1}^N \sum_{Z_i} \log P(X_i,Z_i|\theta)\, P(Z_i|X_i,\theta^{(t)})$

$= \sum_{i=1}^N \sum_{Z_i} \log p_{Z_i} N(X_i|\mu_{Z_i},\Sigma_{Z_i}) \cdot \frac{p_{Z_i}\cdot N(X|\mu_{Z_i}^{(t)},\Sigma_{Z_i}^{(t)})}{\sum_{k=1}^K p_k^{(t)}\cdot N(X_i|\mu_k^{(t)}\Sigma_k^{(t)})}$

$= \sum_{k=1}^K \sum_{i=1}^N [\log p_k + \log N(X_i|\mu_k\Sigma_k)]\, P(Z_i=C_k|X_i,\theta^{(t)})$

- $P(X,Z) = P(Z)P(X|Z)$
  $= p_Z \cdot N(X|\mu_Z,\Sigma_Z)$
- $P(Z|X) = \frac{P(X,Z)}{P(X)} = \frac{p_Z N(X|\mu_Z,\Sigma_Z)}{\sum_{k=1}^K p_k\cdot N(X|\mu_k,\Sigma_k)}$

⑧ M-step $\theta^{(t+1)} = \arg\max Q(\theta,\theta^{(t)})$

求 $p_k^{(t+1)} = (p_1^{(t+1)}, p_2^{(t+1)}, \cdots p_K^{(t+1)})$

$p_k^{(t+1)} = \arg\max_{p_k} \sum_{k=1}^K \sum_{i=1}^N \log p_k \cdot P(Z_i=C_k|X_i,\theta^{(t)})$ s.t. $\sum_{k=1}^K p_k = 1$

带约束的最大值求解

$L(p,\lambda) = \sum_{k=1}^K \sum_{i=1}^N \log p_k \cdot P(Z_i=C_k|X_i,\theta^{(t)}) + \lambda(\sum_{k=1}^K p_k - 1)$

↑ 与 p_k 无关

$\frac{\partial L}{\partial p_k} = \sum_{i=1}^N \frac{1}{p_k}\cdot P(Z_i=C_k|X_i,\theta^{(t)}) + \lambda \triangleq 0$

$\Rightarrow \sum_{i=1}^N P(Z_i=C_k|X_i,\theta^{(t)}) + p_k\cdot\lambda = 0$

$k=1\cdots K \Rightarrow \sum_{i=1}^N \sum_{k=1}^K P(Z_i=C_k|X_i,\theta^{(t)}) + \sum_{k=1}^K p_k\cdot\lambda = 0$

$\sum_{k=1}^K P(Z_i=C_k|X_i,\theta^{(t)}) = 1$, $\sum_{k=1}^K p_k = 1$

$\Rightarrow N+\lambda = 0 \quad \therefore \lambda = -N$

$\therefore p_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^N P(Z_i=C_k|X_i,\theta^{(t)})$

$p^{(t+1)} = (p_1^{(t+1)}, p_2^{(t+1)} \cdots p_N^{(t+1)})$