MACHINE LEARNING BOOK

$E_{p_\theta}[X(p)] = -\int f(u) \log p(u)$

$W_{k+1} = W_k - \eta \nabla W_k$

HEISENBERG

目录

1-1 Gaussian Distribution-MLE　　Maths Basic　　机器学习 (Machine Learning)　　高斯分布

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

$x_i \in \mathbb{R}^p$. x_i 服从 $N(\mu, \Sigma)$, $\theta = (\mu, \Sigma)$　　MLE (Maximum Likelihood Estimation)

MLE: $\theta_{MLE} = \arg\max P(X|\theta)$, $P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, $P(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1}(x-\mu)\right)$.

令 $p=1$, $\theta = (\mu, \sigma^2)$ 以一维为例　$x \sim N(\mu, \sigma^2)$, $x_i \sim N(\mu, \sigma^2)$

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

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

$$= \sum_{i=1}^{N} \left[\underbrace{\log \frac{1}{\sqrt{2\pi}}}_{\text{无关}} + \underbrace{\log \frac{1}{\sigma}}_{\text{无关}} - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$\mu_{MLE} = \arg\max_{\mu} \log P(X|\theta) = \arg\max_{\mu} \sum_{i=1}^{N} -\frac{(x_i-\mu)^2}{2\sigma^2} = \arg\min_{\mu} \sum_{i=1}^{N} (x_i-\mu)^2$$

求 ∇

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

令 $2\sum_{i=1}^{N} (x_i-\mu)(-1) = 0$

$$\sum_{i=1}^{N} (x_i-\mu) = 0$$

$$\sum_{i=1}^{N} x_i - N\mu = 0$$

$$\boxed{\mu_{MLE} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i}$$ 无偏

$$E[\mu_{MLE}] = \frac{1}{N} \sum_{i=1}^{N} E[x_i] = \frac{1}{N} \sum_{i=1}^{N} \mu = \frac{1}{N} \cdot N \cdot \mu = \mu$$

$$\sigma^2_{MLE} = \arg\max_{\sigma} \log P(X|\theta) = \arg\max_{\sigma} \sum_{i=1}^{N} \left[\log \frac{1}{\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}\right] = \arg\max_{\sigma} \sum_{i=1}^{N} \underbrace{\left[-\log\sigma - \frac{(x_i-\mu)^2}{2\sigma^2}\right]}_{\text{记为 } L(\sigma)}$$

$$\frac{\partial L}{\partial \sigma} = \sum_{i=1}^{N} \left[-\frac{1}{\sigma} - \frac{1}{2}(x_i-\mu)^2 \cdot (-2)\frac{1}{\sigma^3}\right] = \sum_{i=1}^{N} \left[-\frac{1}{\sigma} + \frac{1}{\sigma^3}(x_i-\mu)^2\right]$$

令 $\frac{\partial L}{\partial \sigma} = 0$, 则:

$$\sum_{i=1}^{N} \left[-\frac{1}{\sigma} + \frac{1}{\sigma^3}(x_i-\mu)^2\right] = 0.$$

$$\sum_{i=1}^{N} \left[-\sigma^2 + (x_i-\mu)^2\right] = 0.$$

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

$$\boxed{\sigma^2_{MLE} = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i-\mu_{MLE})^2}$$

有偏估计: $E[\sigma^2_{MLE}] = \frac{N-1}{N}\sigma^2$

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

1.2 $E[\sigma^2_{MLE}] \overset{?}{=} \sigma^2$，相等则无偏 $Var[\mu_{MLE}] = Var[\frac{1}{N}\sum_{i=1}^{N} x_i] = \frac{1}{N^2}\sum_{i=1}^{N} Var(x_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{1}{N}\sigma^2$

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

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

则有：$E[\sigma^2_{MLE}] = E[\frac{1}{N}\cdot\sum_{i=1}^{N}x_i^2 - \mu_{MLE}^2] = E[(\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \mu^2) - (\mu_{MLE}^2 - \mu^2)]$

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

$= E\left(\frac{1}{N}\sum_{i=1}^{N}(x_i^2-\mu^2)\right)$ $\quad E(\mu_{MLE}^2) - E(\mu^2)$

$\frac{1}{N}\sum_{i=1}^{N}E(x_i^2-\mu^2)$ $\quad E(\mu_{MLE}^2) - E^2(\mu_{MLE}) \rightarrow E(\mu) = E(\mu_{MLE})$ 无偏

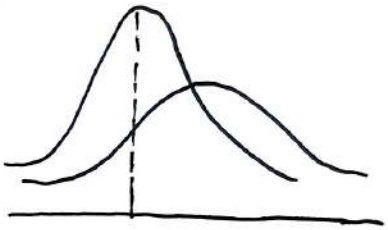
$\frac{1}{N}\sum_{i=1}^{N}(E(x_i^2) - \mu^2)$ $\rightarrow (E(x_i))^2$ $\quad Var(\mu_{MLE})$

$Var(x_i)$ $\quad E(x_i)=\mu$ $\quad \frac{1}{N}\sigma^2$

σ^2

$\frac{1}{N}\sum_{i=1}^{N}\sigma^2$

σ^2



注：$E(x-\bar{x})$ 而非 $E|x-\mu|$

1.3 Gaussian Distribution from σ^2 perspective

1.2 Gaussian Distribution-MLE (unbiased vs biased)

∴极大似然估计中 σ^2 有偏差

高维情况下：高斯分布

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

PDF；二次型；$x \in \mathbb{R}^p$，random vector → r.v.

$x = \begin{pmatrix}x_1\\ \vdots \\ x_p\end{pmatrix}$；$\mu = \begin{pmatrix}\mu_1\\ \vdots \\ \mu_p\end{pmatrix}$；$\Sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p}\\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp}\end{pmatrix}$ $p\times p$

协方差矩阵

Σ：正定的（以下讨论）

：半正定（一般简化）且对称

$(x-\mu)^T\Sigma^{-1}(x-\mu)$：马氏距离（$x$ 与 μ 之间）表示点与一个分布之间的距离。它是一种有效计算两个未知样本集相似度的方法。与欧氏距离不同的是，它考虑到各种特性之间的联系

$1\times p$ $\quad p\times p$ $\quad p\times 1$

$\Sigma = I$，马氏距离 = 欧氏距离

eg. $z_1 = \begin{pmatrix}z_{11}\\ z_{12}\end{pmatrix}$；$z_2 = \begin{pmatrix}z_{21}\\ z_{22}\end{pmatrix}$ 有 $(z_1-z_2)^T\Sigma^{-1}(z_1-z_2)$

$= (z_{11}-z_{21}, z_{12}-z_{22})\cdot\Sigma^{-1}\begin{pmatrix}z_{11}-z_{21}\\ z_{12}-z_{22}\end{pmatrix}$

若 $\Sigma = I$，则 $= (z_{11}-z_{21})^2 + (z_{12}-z_{22})^2$

对于一元：$\sum_{i=1}^{p}\frac{y_i^2}{\lambda_i}$，令 $p=2$，

则 $p=2$，$\Delta = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2}$，假 $\Delta=1$，这是一个椭圆

特征分解 正交

对 $\Sigma = U\Lambda U^T$，$UU^T = U^TU = I$，$\Lambda = diag(\lambda_i)$ $i=1,\cdots,p$，$U = (u_1,\cdots,u_p)_{p\times p}$

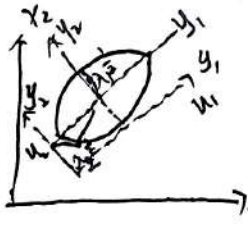
$= (u_1,\cdots,u_p)\begin{pmatrix}\lambda_1 & & \\ & \lambda_2 & \\ & & \ddots & \\ & & & \lambda_p\end{pmatrix}\begin{pmatrix}u_1\\ \vdots \\ u_p\end{pmatrix} = (u_1\lambda_1, u_2\lambda_2, \cdots, u_p\lambda_p)\begin{pmatrix}u_1\\ \vdots \\ u_p\end{pmatrix} = \sum_{i=1}^{p}u_i\lambda_i u_i^T$

$\Sigma^{-1} = (U\Lambda U^T)^{-1} = (U^T)^{-1}\Lambda^{-1}U^{-1} = (U^T)^{-1}\Lambda^{-1}(U^{-1}) = U\Lambda^{-1}U^T$

$\Lambda^{-1} \rightarrow diag(\frac{1}{\lambda_i})$，$i=1,\cdots,p$

$= \sum_{i=1}^{p}u_i\frac{1}{\lambda_i}u_i^T = \sum_{i=1}^{p}u_i\frac{1}{\lambda_i}u_i^T$

$\Delta = \boxed{(x-\mu)^T\Sigma^{-1}(x-\mu)} = (x-\mu)^T\sum_{i=1}^{p}u_i\frac{1}{\lambda_i}u_i^T\cdot(x-\mu)$ P2



$\lambda_1 > \lambda_2$

$\Delta = r \rightarrow p(r)$

取不同值时相当于不同的等高线

$$= \sum_{i=1}^{p} (x-\mu)^T u_i \cdot \frac{1}{\lambda_i} u_i^T (x-\mu)$$

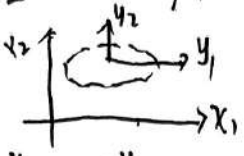
$y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$; $y_i = (x-\mu)^T u_i$. y_i是1维的

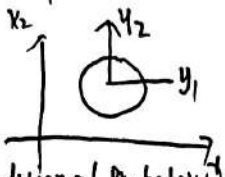
$$= \sum_{i=1}^{p} y_i \frac{1}{\lambda_i} y_i^T = \boxed{\sum_{i=1}^{p} \frac{y_i^2}{\lambda_i}}$$

1.4 Gaussian Distribution - ~~from MLE perspective~~ Limitation

局限性：

① $\Sigma_{p\times p}$ → 参数个数：$C_p^2 + p = \frac{p(p-1)}{2} + p = \frac{p(p+1)}{2} \to O(p^2)$

Σ → 对角矩阵 $\begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \lambda_p \end{pmatrix}$ → 

$\lambda_1 = \lambda_2 = \cdots \lambda_p$ → 

→ 各向同性

Gaussian Distribution - Marginal & Conditional probability

factor analysis (因子分析) → 对角矩阵

P-PCA → 各向同性

1.4.1

$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \begin{matrix} \to m \\ \to n \end{matrix}$ $m+n=p$ $\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$ $\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$ → 对称

求 $P(x_a)$, $P(x_b|x_a)$; $P(x_b)$, $P(x_a|x_b)$ 对称

$$x_a = \underbrace{(I_m \; 0)}_{A} \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{x} + \underbrace{B}_{B=0}$$

Gauss分布定理

定理：已知 $x \sim N(\mu, \Sigma)$, $y = Ax + B$ （$q\times p$ $p\times 1$ 维度↑↑）

结论：$\boxed{y \sim N(A\mu + B, A\Sigma A^T)}$ （$q\times p$ $p\times p$ $p\times q$）

$E[y] = E[Ax+B] = AE(x) + B = A\mu + B$

$Var[y] = Var[Ax+B] = Var[Ax] + Var[B] = A \cdot Var[x] A^T = A\Sigma A^T$

（常数方差为0）

$E[x_a] = A\mu + B = (I_m \; 0)\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} + 0 = \mu_a$

$Var[x_a] = (I_m \; 0)\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} I_m \\ 0 \end{pmatrix} = (\Sigma_{aa} \; \Sigma_{ab})\begin{pmatrix} I_m \\ 0 \end{pmatrix} = \Sigma_{aa}$

$\therefore \boxed{x_a \sim N(\mu_a, \Sigma_{aa})}$

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

令 $\begin{cases} \mu_{b\cdot a} = \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a \\ \Sigma_{bb\cdot a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \end{cases}$ ← 结果定义

$\Sigma_{bb\cdot a}$ → Σ_{aa}的schur complementary

$$x_{b\cdot a} = \underbrace{(-\Sigma_{ba}\Sigma_{aa}^{-1} \; I)}_{A} \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{x}$$

$E[x_{b\cdot a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1} \; I)\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} + \frac{B}{0} = \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a = \mu_{b\cdot a}$

$Var[x_{b\cdot a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1} \; I)\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^T \\ I \end{pmatrix} = \begin{pmatrix} -\Sigma_{ba} + \Sigma_{ba} & -\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} + \Sigma_{bb} \end{pmatrix}\begin{pmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^T \\ I \end{pmatrix}$

~~$-\Sigma_{ba} + \Sigma_{ba}$ $-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma$~~

$= (0 \; -\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} + \Sigma_{bb})\begin{pmatrix} -\Sigma_{aa}^{-1}\Sigma_{ba}^T \\ I \end{pmatrix} = -\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} + \Sigma_{bb} = \Sigma_{bb\cdot a}$

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

$x_b | x_a$.

$\begin{pmatrix} I & 0 \\ -\Sigma_{ba}\Sigma_{aa}^{-1} & I \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix}$

$x_b = x_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1} x_a$ = $I x_{b\cdot a} + \underbrace{\Sigma_{ba}\Sigma_{aa}^{-1} x_a}_{B}$

$E[x_b | x_a] = \begin{pmatrix} I & 0 \\ -\Sigma_{ba}\Sigma_{aa}^{-1} & I \end{pmatrix} \begin{pmatrix} x_a \\ \mu_b \end{pmatrix}$

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

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

x_b与x_a之间的那个式子，体现了x_a与x_b的一个条件关系，对给定的x_a，有固定映射的x_b，所以此时$E(x_b)$为给定条件下的$E(x_b)$，亦即$E(x_b|x_a)$

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

在$x_b | x_a$中，$x_{b\cdot a}$, x_a相互独立.

相互独立

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

$= E(x_{b\cdot a}) + \Sigma_{ba}\Sigma_{aa}^{-1} x_a$.

$\because A = I$

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

$Var[x_b | x_a] = Var[x_b] = Var[x_{b\cdot a} + \underbrace{\Sigma_{ba}\Sigma_{aa}^{-1} x_a}_{B}]$

常数$Var[B] = 0$, 不影响$Var[x_b | x_a] = Var[x_{b\cdot a}] = \Sigma_{bb\cdot a}$

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

$$\boxed{\begin{aligned} &x_a \sim N(\mu_a, \Sigma_{aa}) \\ &x_b \sim N(\mu_b, \Sigma_{bb}) \\ &x_a | x_b \sim N(\mu_{a\cdot b} + \Sigma_{ab}\Sigma_{bb}^{-1} x_b, \Sigma_{aa\cdot b}) \end{aligned}}$$

1.86 Gaussian Distribution - Joint Probability

在上节基础上，依据Gauss分布定理，求：

Gauss分布：$x \sim N(\mu, \Sigma)$, $y = Ax + B$.

$\Rightarrow y \sim N(A\mu + B, A\Sigma A^T)$

已知：$p(x) = N(x | \mu, \Lambda^{-1})$ → precision matrix 精度 = (covariance matrix)$^{-1}$ 协方差

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

求：① $p(y)$, ② $p(x|y)$

解：① $y = Ax + b + \varepsilon$ — x, y, ε：随机变量，A, b：系数；噪声

$\varepsilon \sim N(0, L^{-1})$ — $\varepsilon \perp x$，ε与x独立.

$E[y] = E[Ax + b + \varepsilon] = E[Ax + b] + E(\varepsilon) = A\mu + b$

$Var[y] = Var[Ax + b + \varepsilon] = Var[Ax] + Var[b] + Var[\varepsilon] = A \cdot \Lambda^{-1} A^T + L^{-1}$

$\therefore \boxed{y \sim N(A\mu + b, L^{-1} + A \cdot \Lambda^{-1} A^T)}$

$\Lambda = \Lambda^{-1} \cdot A^T$ ↑ 对称.

② $z = \begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\underbrace{\begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}}_{E[z]}, \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1} A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{bmatrix} \right)$ ← $Var[z]$

$A\Lambda^{-1}$

$p(z) = N\left(\underbrace{\begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}}_{\mu}, \underbrace{\begin{pmatrix} \Lambda^{-1} & \Lambda^{-1} A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}}_{\Sigma} \right)$.

$E[x|y] = \mu + \Lambda^{-1} A^T (L^{-1} + A\Lambda^{-1} A^T)^{-1} (y - A\mu - b)$

$Var[x|y] = \Lambda^{-1} - \Lambda^{-1} A^T (L^{-1} + A\Lambda^{-1} A^T)^{-1} A \Lambda^{-1}$

$A = cov(x, y) = E[(x - E[x]) \cdot (y - E[y])^T]$

$= E[(x - \mu) \cdot (y - A\mu - b)^T]$

$= E[(x - \mu)(Ax + b + \varepsilon - A\mu - b)^T]$

$= E[(x - \mu)(Ax - A\mu + \varepsilon)^T]$

$= E[(x - \mu)(Ax - A\mu)^T + (x - \mu)\varepsilon^T]$

$= E[(x - \mu)(Ax - A\mu)^T] + E[(x - \mu)\varepsilon^T]$

$= E[(x - \mu)(Ax - A\mu)^T]$ ($x \perp \varepsilon$)

~~$= E[(x - \mu)A(x - \mu)^T]$~~ ($x - \mu \perp \varepsilon$)

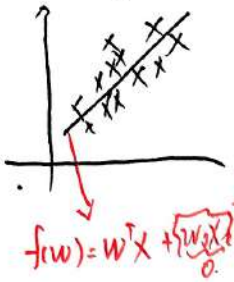
$= E[(x - \mu)(x - \mu)^T A^T]$ $E(x - \mu) \cdot \underbrace{E(\varepsilon)}_{0}$

$= E[(x - \mu)(x - \mu)^T] \cdot A^T$

$= Var[x] \cdot A^T = \Lambda^{-1} \cdot A^T$

2 线性回归 (Linear Regression) ☆

2.1 - Least Square Method



$$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

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

$$X=(x_1,x_2,\cdots,x_N)^T=\begin{pmatrix}x_1^T\\x_2^T\\\vdots\\x_N^T\end{pmatrix}=\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&&\vdots\\x_{N1}&x_{N2}&\cdots&x_{Np}\end{pmatrix}_{N\times p}$$

$N\gg p$

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

最小二乘估计:

$$L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2=\sum_{i=1}^{N}(w^Tx_i-y_i)^2=(w^Tx_1-y_1,\ w^Tx_2-y_2\cdots w^Tx_N-y_N)\begin{pmatrix}w^Tx_1-y_1\\w^Tx_2-y_2\\\vdots\\w^Tx_N-y_N\end{pmatrix}$$

$$=\left[w^T(x_1,x_2\cdots x_N)-(y_1,y_2\cdots y_N)\right]\begin{pmatrix}w^Tx_1-y_1\\w^Tx_2-y_2\\\vdots\\w^Tx_N-y_N\end{pmatrix}=(w^TX^T-Y^T)\cdot(Xw-Y)$$

$$=w^TX^TXw-w^TX^TY-Y^TXw+Y^TY=w^TX^TXw-2w^TX^TY+Y^TY$$

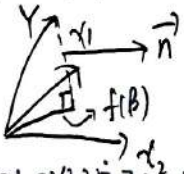
$$\hat{w}=\arg\min L(w)$$

$$\frac{\partial L(w)}{\partial w}=2X^TXw-2X^TY=0.$$

$$\therefore X^TXw=X^TY$$

$$\boxed{w=\underbrace{(X^TX)^{-1}}_{X^+\text{伪逆}}X^TY}\ \text{解析解}$$

几何理解.



$\vec{n}\perp$ 列 P 维空间

$\vec{n}=(Y-XB)$

由 $\vec{a}\perp\vec{b}\Leftrightarrow a^Tb=0$

则有 $(Y-XB)\perp X\Leftrightarrow X^T(Y-XB)=0=\begin{pmatrix}0\\\vdots\\0\end{pmatrix}_{P\times 1}\Leftrightarrow X^TY=X^TXB$

$P\times N,\ N\times 1,\ N\times P,\ P\times 1$

$$\boxed{B=(X^TX)^{-1}X^TY}$$

Y 在 P 维子空间中 (噪声 noise, random)

Y 到 P 维空间最短距离: $\vec{n}$ (垂线段)

$f(B)$ 这一个投影, 就是 $(x_1,\cdots,x_p)$ 的线性组合 $\Leftrightarrow XB$

则有 $\vec{n}=(Y-XB)$, 而由 $\vec{n}\perp$ P 维子空间, 则有 $X^T(Y-XB)=0\Leftrightarrow\boxed{B=(X^TX)^{-1}X^TY}$.

$$f(w)=w^Tx=x^TB$$

$$B=\begin{pmatrix}b_1\\\vdots\\b_p\end{pmatrix}_{p\times 1},\ X^T=(\ \ )_{p\times N}$$

2.2. Least Square Method - from probabilistic perspective

概率视角

$$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

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

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

$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$ x:样本 y值

$P(y|x;w)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-w^Tx)^2}{2\sigma^2}\right]$

$y|x;w$是给定x,w的条件分布，方差来源是误差

这里x,w看作constant

给定特征x和w的条件下，标签y的概率分布

参数

w是模型参数

噪声：$\varepsilon\sim N(0,\sigma^2)$

$y=f(w)+\varepsilon$

$f(w)=w^Tx$

$y=w^Tx+\varepsilon$

$y|x;w\sim N(w^Tx,\sigma^2)$

最小2乘估计：

$L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|_2^2$

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

$\hat{w}=(x^Tx)^{-1}x^TY$

给定为w时，观测到y的概率，取上对log，称为log-Likelihood

MLE：

$L(w)=\log P(Y|x;w)=\log\prod_{i=1}^{N}P(y_i|x_i;w)=\sum_{i=1}^{N}\log P(y_i|x_i;w)=\sum_{i=1}^{N}\log\frac{1}{\sqrt{2\pi}\sigma}+\log\exp\left[-\frac{(y_i-w^Tx_i)^2}{2\sigma^2}\right]$

log-Likelihood

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

定义相似.

MLE

$\hat{w}=\arg\max_w L(w)=\arg\max_w\left[-\frac{1}{2\sigma^2}(y_i-w^Tx_i)^2\right]=\arg\min_w(y_i-w^Tx_i)^2$

则有$\hat{w}=(x^Tx)^{-1}x^TY$，x:样本 y:观测值

最小二乘 $\Leftrightarrow$ MLE

条件（noise is Gaussian Dist）☆

相对似然

2.3. Regularization - Ridge Regression - frequentist

正则化 - 岭回归 - 频率角度

$\sim w$ is constant.

Loss Function：$L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2$ → 目标函数

$\hat{w}=(x^Tx)^{-1}x^TY$

且容易过拟合 ← ①加数据 ②降维选择/特征提取PCA ③正则化

$X_{N\times p}$，N个样本，$x_i\in\mathbb{R}^p$，正常来说$N\gg p$，但也会存在$N<p$甚至$N\ll p$（样本少，维度高），无法求x^{-1}，无法得到解析解的情况，此时引入正则化

正则化框架

$\arg\min_w[L(w)+\lambda P(w)]$ … ①

$L(w)$：loss　$P(w)$：Penalty（惩罚项）

L_1：Lasso，$P(w)=\|w\|$

L_2：Ridge，岭回归，$P(w)=\|w\|_2^2=w^Tw$

权值衰减

对于①→L_2有：$①=J(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2+\lambda w^Tw=(w^Tx^T-Y^T)(xw-Y)+\lambda w^Tw=w^Tx^Txw-2w^Tx^TY+Y^TY+\lambda w^Tw$

PS推导

$=w^Tx^Txw+w^T\lambda w-2w^Tx^TY+Y^TY=w^T(x^Tx+\lambda I)w-2w^Tx^TY+Y^TY$

λ实数可以写到中间

$\hat{w} = \underset{w}{\operatorname{argmin}} J(w)$

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

$w = (X^TX + \lambda I)^{-1}X^TY$ 岭回归下的解析解 与 $w = (X^TX)^{-1}X^TY$ 对比 ← LSE

X^TX → 半正定 ; λI → 对角 ⇒ 正定+对角 ⇒ 可逆

2.4. Regularization - Ridge Regression - Bayesians

P6. $y|x;w \sim N(w^Tx, \sigma^2) \Rightarrow y = w^Tx + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ ← noise

r.v $w \sim N(0, \sigma_0^2)$

Bayesian 中 w is r.v.

Bayesians Perspective:

$w \sim N(0, \sigma_0^2)$, $P(y|w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-w^Tx)^2}{2\sigma^2}\right]$, $P(w) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{\|w\|^2}{2\sigma_0^2}\right]$

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

MAP (最大后验): $\hat{w} = \underset{w}{\operatorname{argmax}} P(w|y) = \underset{w}{\operatorname{argmax}} P(y|w)P(w)$

$P(y|w)P(w) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(y-w^Tx)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2}\right]$

$= \underset{w}{\operatorname{argmax}} \log[P(y|w)P(w)] = \underset{w}{\operatorname{argmax}} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{\sqrt{2\pi}\sigma_0}\right) + \log\exp\left\{-\frac{(y-w^Tx)^2}{2\sigma^2} - \frac{\|w\|^2}{2\sigma_0^2}\right\}$

$= \underset{w}{\operatorname{argmin}} \left\{\frac{(y-w^Tx)^2}{2\sigma^2} + \frac{\|w\|^2}{2\sigma_0^2}\right\} \overset{\text{同乘}\sigma^2}{=} \underset{w}{\operatorname{argmin}} \left\{(y-w^Tx)^2 + \frac{\sigma^2}{\sigma_0^2}\|w\|^2\right\} \cdots ①$

为了简化运算，省略 $= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{N} P(y_i|w)P(w)$

则代入有: $\hat{w}_{MAP} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} [(y_i - w^Tx_i)^2 + \frac{\sigma^2}{\sigma_0^2}\|w\|_2^2] = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{N} [(w^Tx_i - y_i)^2 + \lambda w^Tw] =$

上一节的 $J(w)$. P6. $\lambda = \frac{\sigma^2}{\sigma_0^2}$

$\underset{w}{\operatorname{argmin}} [L(w) + \lambda P(w)] \cdots ①$

Loss ; Penalty.

Regularized LSE ⟺ MAP (noise 为 Gaussian Dist)

(Prior 也是 GD)

最大后验估计

☆

---

3 Linear classification 线性分类

3.1 Linear classification Background

(频率派 → 统计机器学习
贝叶斯派 → 概率图模型

统计机器学习 → Linear Regression $f(w,b) = w^Tx + b$, $x \in \mathbb{R}^p$

- ① 线性 → 属性非线性：特征转换（多项式回归）
  - 全属非线性：线性分类（激活函数是非线性）
  - 参数非线性：神经网络，感知机
- ② 全局性 ⇒ 线性样条回归，决策树
- ③ 数据未知 → PCA、流形.

线性分类 $y = \{0,1\}$

- 硬分类：线性判别分析 (Fisher)
  - 感知机
- 软分类：
  - 生成式：Gaussian Discriminal Analysis
  - 判别式：Logistic Regression

$y \in [0,1]$

线性回归 → 激活函数 / 降维 → 线性分类 $(0,1)$

activation function

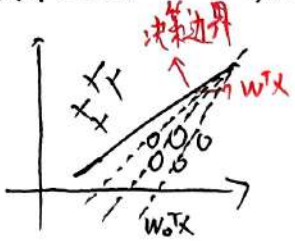
$y = f(w^Tx + b), y \in \{0,1\}$ 或 $[0,1]$

f^{-1}: link function

$f: w^Tx + b \mapsto \{0,1\}$

$f^{-1}: \{0,1\} \mapsto w^Tx + b$

3.2. Linear Classification Perceptron Algorithm (1957年)



D: {被错误分类的样本}

一次一次移动 w_t^Tx

样本集: $\{(x_i,y_i)\}_{i=1}^N$

思想: 错误驱动

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

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

x_i, y_i

$$\left.\begin{matrix}w^Tx_i>0, y_i=+1\\ w^Tx_i<0, y_i=-1\end{matrix}\right\}\Rightarrow\begin{cases}y_iw^Tx_i>0 \text{ 正确分类}\\ y_iw^Tx_i<0 \text{ 错误分类}\end{cases}$$

关于 w 连续函数

策略: loss function:

$$L(w)=\sum_{i=1}^N I\{y_iw^Tx_i<0\}$$

NP hard 不可导, 连续, w 轻易改动, I 有可能为 0 或 1

新 loss function:

$$L(w)=\sum_{x_i\in D}-y_iw^Tx_i$$

$$\nabla_wL=-y_ix_i$$

算法: SGD (随机梯度下降): $w^{(t+1)}\leftarrow w^{(t)}-\lambda\nabla_wL=w^{(t)}+y_ix_i$ (λ 为学习率)

3.3 Fisher - Linear discrimination Analysis

线性判别分析

$$X=(x_1,x_2,\cdots,x_N)^T=\begin{pmatrix}x_1^T\\ x_2^T\\ \vdots\\ x_N^T\end{pmatrix}_{N\times p},\quad Y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_N\end{pmatrix}_{N\times 1}$$

$\{(x_i,y_i)\}_{i=1}^N$, $x_i\in\mathbb{R}^p$, $y_i\in\{+1,-1\}$ ($+1$: C_1, -1: C_2)

$x_{c_1}=\{x_i\mid y_i=+1\}$, $x_{c_2}=\{x_i\mid y_i=-1\}$

$|x_{c_1}|=N_1$, $|x_{c_2}|=N_2$, $N_1+N_2=N$

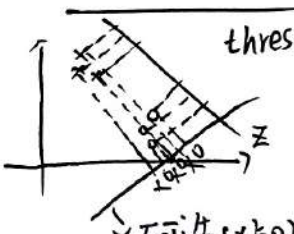
思想: 类内小, 类间大 ⇒ 目标函数: $J(w)=\dfrac{(\bar z_1-\bar z_2)^2}{s_1+s_2}$ → 类间 / → 类内

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

$$J(w)=\frac{(\bar z_1-\bar z_2)^2}{s_1+s_2}$$

分子: $\left(\frac{1}{N_1}\sum_{i=1}^{N_1}w^Tx_i-\frac{1}{N_2}\sum_{i=1}^{N_2}w^Tx_i\right)^2$

$=\left(w^T\left(\frac{1}{N_1}\sum_{i=1}^{N_1}x_i-\frac{1}{N_2}\sum_{i=1}^{N_2}x_i\right)\right)^2=\left(w^T(\bar x_{c_1}-\bar x_{c_2})\right)^2$

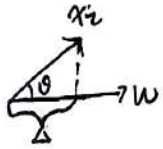


不可选 (x与0) 混合

"松耦合, 高内聚"

类内: 方差足够小

类间: 距离大.



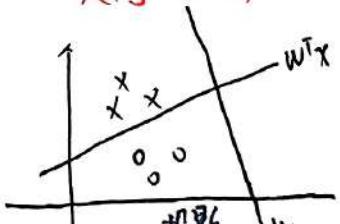
$\Delta=|x_{i2}|\cos\theta$

$x_{i2}w=|x_{i2}||w|\cos\theta$

$\because |w|=1$

$\therefore \Delta=x_{i2}w$

在此处将投影视为 w^Tx_i



C_1, C_2 为均值

S_1, S_2 为方差

$\bar x_{c_1}, \bar x_{c_2}$ 为 x 的均值

S_{c_1}, S_{c_2} 为方差

$\|w\|=1$

投影 $z_i=w^Tx_i$, $\bar z=\frac{1}{N}\sum_{i=1}^N z_i=\frac{1}{N}\sum_{i=1}^N w^Tx_i$

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

类间: $(\bar z_1-\bar z_2)^2$

类内: S_1+S_2

C_1: $\bar z_1=\frac{1}{N_1}\sum_{i=1}^{N_1}w^Tx_i$

$$S_1=\frac{1}{N_1}\sum_{i=1}^{N_1}(w^Tx_i-\bar z_1)(w^Tx_i-\bar z_1)^T$$

C_2: $\bar z_2=\frac{1}{N_2}\sum_{i=1}^{N_2}w^Tx_i$

$$S_2=\frac{1}{N_2}\sum_{i=1}^{N_2}(w^Tx_i-\bar z_2)(w^Tx_i-\bar z_2)^T$$

P8

分母 $= S_1 + S_2$

$$S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}(w^T x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^T x_i)(w^T x_i - \frac{1}{N_1}\sum_{j=1}^{N_1} w^T x_i)^T$$

$$= \frac{1}{N_1}\sum_{i=1}^{N_1} w^T (x_i - \bar{x}_{c_1})(x_i - \bar{x}_{c_1})^T w$$

$$= w^T \cdot \left[\frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - \bar{x}_{c_1})(x_i - \bar{x}_{c_1})^T\right] w$$

S_{c_1} 方差：x_i 与 $\bar{x}_{c_1}$ 的

$$= w^T \cdot S_{c_1} \cdot w$$

分母 $= w^T S_{c_1} w + w^T S_{c_2} w = w^T (S_{c_1} + S_{c_2}) w$

$$\therefore J(w) = \frac{w^T (\bar{x}_{c_1} - \bar{x}_{c_2})(\bar{x}_{c_1} - \bar{x}_{c_2})^T w}{w^T (S_{c_1} + S_{c_2}) w}$$

$$= \frac{w^T S_b w}{w^T S_w w}$$

S_b：between-class 类间方差 $p \times p$　$S_b = (\bar{x}_{c_1} - \bar{x}_{c_2})(\bar{x}_{c_1} - \bar{x}_{c_2})^T$

S_w：with-class 类内方差 $p \times p$　$S_w = S_{c_1} + S_{c_2}$

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

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

$$\Rightarrow S_b w \cdot (w^T S_w w) - w^T S_b w S_w w = 0$$

$$w^T S_b w S_w w = S_b w \cdot (w^T S_w w)$$

$w^T S_b w$：实数 GR　$w^T S_w w$：实数 GR

w^T：$1 \times p$　S_w：$p \times p$　w：$p \times 1$

$$S_w w = \frac{w^T S_w w}{w S_b w} \cdot S_b w$$

$$w = \frac{w^T S_w w}{w^T S_b w} \cdot S_w^{-1} S_b w$$

不关心 w 大小 → 只关心 w 方向。

$$\propto S_w^{-1} S_b w$$

$$\propto S_w^{-1} (\bar{x}_{c_1} - \bar{x}_{c_2}) \underbrace{(\bar{x}_{c_1} - \bar{x}_{c_2})^T}_{1 \times p} \cdot \underbrace{w}_{p \times 1}$$

GIR

$$w \propto \underbrace{S_w^{-1}}_{p \times p} \underbrace{(\bar{x}_{c_1} - \bar{x}_{c_2})}_{p \times 1}$$ 一矩阵 x 向量 → 方向。

若 S_w 是对角阵，且各向同性，$S_w^{-1} \propto 1$，则 $w \propto (\bar{x}_{c_1} - \bar{x}_{c_2})$

3.4. Logistic Regression

Data: $\{(x_i, y_i)\}_{i=1}^N$

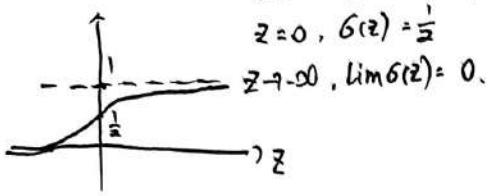
$x_i \in \mathbb{R}^p$, $y_i \in \{0,1\}$

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

$z \to \infty$, $\lim \sigma(z) = 1$

$z = 0$, $\sigma(z) = \frac{1}{2}$

$z \to -\infty$, $\lim \sigma(z) = 0$.



$\sigma: \mathbb{R} \to (0,1)$

$w^T x \mapsto p$

有时我们只得到一个类别的概率，那么我们需要一种能输出$[0,1]$区间的值的函数，考虑二分类模型，我们利用判别模型，希望对$P(C|x)$建模，利用Bayes.

$$P(C_1|x) = \frac{P(x|C_1)\cdot P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

取$a = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$，于是 $P(C_1|x) = \frac{1}{1+\exp(-a)} = \frac{1}{1+e^{-w^T x}}$

上面式子叫 Logistic Sigmoid 函数

$a = w^T x$

$$p_1 = P(y=1|x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}, \quad y=1 \quad (=\varphi(x_i, w))$$

$$p_0 = P(y=0|x) = 1 - P(y=1|x) = \frac{e^{-w^T x}}{1+e^{-w^T x}}, \quad y=0 \quad (=1-\varphi(x_i,w))$$

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

MLE:
$$\hat{w} = \arg\max_w \log P(Y|x) = \arg\max_w \log \prod_{i=1}^N P(y_i|x_i) = \arg\max_w \sum_{i=1}^N \log P(y_i|x_i) = \arg\max_w \underbrace{\sum_{i=1}^N \left(y_i \log p_1 + (1-y_i)\log p_0\right)}_{J(w)}$$

$$= \arg\max_w \sum_{i=1}^N \underbrace{\left[y_i \log \varphi(x_i,w) + (1-y_i)\log(1-\varphi(x_i,w))\right]}_{-\text{cross Entropy}}$$

(max) (min)

MLE $\Rightarrow$ loss function (min cross Entropy)

$$J(w) = \sum_{i=1}^N \left(y_i \log p_1 + (1-y_i)\log p_0\right), \quad p_1' = \left(\frac{1}{1-e^{-a}}\right)' = p_1(1-p_1)$$

损失函数是凹函数，但非二次函数，无法令导数等于0求出闭式解

则 $J'(w) = \sum_{i=1}^N y_i \cdot \frac{p_1(1-p_1)}{p_1}\cdot x_i + y_i p_1 x_i - p_1 x_i = \sum_{i=1}^N (y_i - p_1) x_i$ → 令导数等于0求出闭式解

由于概率值非线性，无法直接求(与Perceptron类似)，可以随机梯度上升(下降)求最大(最小)值

3.5. Gaussian Discriminant Analysis

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^p$, $y_i \in \{0,1\}$

Bayes: $P(y|x) = \frac{P(x|y)P(y)}{P(x)} \propto \underbrace{P(x|y)P(y)}_{P(x,y)\to\text{联合概率}}$ (posterior ∝ Likelihood × prior)

$$\hat{y} = \arg\max_{y\in\{0,1\}} P(y|x) = \arg\max_y P(y)P(x|y)$$ → 生成模型的基本原理

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

y	1	0
p	ϕ	$1-\phi$

$\to \begin{cases} \phi^y, & y=1 \\ (1-\phi)^{1-y}, & y=0 \end{cases} \Rightarrow \phi^y (1-\phi)^{1-y}$ 合并为一项 $y \in \{0,1\}$

GDA:
$x|y=1 \sim N(\mu_1, \Sigma)$
$x|y=0 \sim N(\mu_2, \Sigma)$
$\to N(\mu_1,\Sigma)^y \cdot N(\mu_2,\Sigma)^{1-y}$

log-likelihood $= L(\theta) = \log \prod_{i=1}^N P(x_i, y_i) = \sum_{i=1}^N \log P(x_i|y_i)P(y_i) = \sum_{i=1}^N \left[\log P(x_i|y_i)\right] + \sum_{i=1}^N \left[\log P(y_i)\right] = \sum_{i=1}^N \left[\log P(x_i|y_i) + \log P(y_i)\right]$

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

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

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

$y=1: N_1$, $y=0: N_2$, $N = N_1 + N_2$

$$= \sum_{i=1}^N \left[\underbrace{\log N(\mu_1,\Sigma)^{y_i}}_{①} + \underbrace{\log N(\mu_2,\Sigma)^{1-y_i}}_{②} + \underbrace{\log \phi^{y_i}(1-\phi)^{1-y_i}}_{③\ \text{与}\ p(x)\ \text{无关}}\right] = \sum_{i=1}^N \left[\underbrace{y_i \log N(\mu_1,\Sigma)}_{①} + \underbrace{(1-y_i)\log N(\mu_2,\Sigma)}_{②} + \underbrace{\log \phi^{y_i}(1-\phi)^{1-y_i}}_{③}\right]$$

3.b. 求解GDA参数(写出过程)

求 ϕ: $② = \sum_{i=1}^{N} \log \phi^{y_i}(1-\phi)^{1-y_i} = \sum_{i=1}^{N} y_i \log\phi + (1-y_i)\log(1-\phi)$

$$\frac{\partial ②}{\partial \phi} = \sum_{i=1}^{N}\left[y_i \frac{1}{\phi} + (1-y_i)\cdot\frac{1}{1-\phi}(-1)\right] = \sum_{i=1}^{N}\left[y_i\frac{1}{\phi} - (1-y_i)\frac{1}{1-\phi}\right] \triangleq 0$$

$$\Rightarrow \sum_{i=1}^{N}\left[(1-\phi)y_i - (1-y_i)\phi\right] = 0 \Leftrightarrow \sum_{i=1}^{N}\left[y_i - \phi y_i - \phi + \phi y_i\right] = 0$$

$$\sum_{i=1}^{N}\left[y_i - \phi\right] = 0$$

$y=1, N_1$; $y=0, N_2$ $\quad \sum_{i=1}^{N} y_i = \sum_{i=1}^{N_1} y_1 + \sum_{i=1}^{N_2} y_2 = N_1\cdot 1 + N_2\cdot 0 = N_1$

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

求 μ_1: $② = \sum_{i=1}^{N} \log \mathcal{N}(\mu_1, \Sigma)^{y_i} = \sum_{i=1}^{N} y_i \log \frac{1}{(2\pi)^{\frac{p}{2}}\cdot|\Sigma|} \exp\left\{-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right\}$

$$\mu_1 = \arg\max ① = \arg\max \sum_{i=1}^{N} y_i\cdot\left[-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right] = \arg\max -\frac{1}{2}\sum_{i=1}^{N} y_i\left[(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right]$$

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

($x_i^T\Sigma^{-1}\mu_1$: $1\times p$, $p\times p$, $p\times 1$, $\in\mathbb{R}$; $\mu_1^T\Sigma^{-1}x_i$: $1\times p$, $p\times p$, $p\times 1$, $\in\mathbb{R}$)

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

$\sum_{i=1}^{N} y_i$: Δ; $x_i^T\Sigma^{-1}x_i$: 常数; $2\mu_1^T\Sigma^{-1}x_i - \mu_1^T\Sigma^{-1}\mu_1$: Δ

常数舍弃

$$\frac{\partial \Delta}{\partial \mu_1} = -\frac{1}{2}\sum_{i=1}^{N} y_i(-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu_1) \triangleq 0$$

$$\sum_{i=1}^{N} y_i(\Sigma^{-1}x_i - \Sigma^{-1}\mu_1) \triangleq 0$$

$$\sum_{i=1}^{N} y_i(\mu_1 - x_i) = 0$$

$$\sum_{i=1}^{N} y_i\mu_1 = \sum_{i=1}^{N} y_i x_i$$

$$\boxed{\hat{\mu}_1 = \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} y_i x_i}{N_1}}$$

求 μ_2:

$$\frac{\partial \Delta}{\partial \mu_2} = \sum_{i=1}^{N}(1-y_i)(\mu_2 - x_i) = 0$$

$$\boxed{\hat{\mu}_2 = \frac{\sum_{i=1}^{N} x_i(1-y_i)}{\sum_{i=1}^{N}(1-y_i)} = \frac{\sum_{i=1}^{N} x_i(1-y_i)}{N_1+N_2-N_1} = \frac{\sum_{i=1}^{N} x_i(1-y_i)}{N_2}}$$

求 Σ:

$C_1 = \{x_i \mid y_i = 1, i=1,2,\cdots,n\}$ $\quad |C_1| = N_1$, $|C_2| = N_2$

$C_2 = \{x_i \mid y_i = 0, i=1,2,\cdots,n\}$ $\quad N_1+N_2 = N$

$$\hat{\Sigma} = \arg\max \underset{①+②}{\quad}, \quad ① + ② = \sum_{x_i\in C_1}\log\mathcal{N}(\mu_1,\Sigma) + \sum_{x_i\in C_2}\log\mathcal{N}(\mu_2,\Sigma)$$

通项 $\sum_{i=1}^{N}\log N(\mu,\Sigma) = \sum_{i=1}^{N}\log \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right\}$

$= \sum_{i=1}^{N}\log\frac{1}{(2\pi)^{\frac{p}{2}}} + \log|\Sigma|^{-\frac{1}{2}} + \left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$

常数

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

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

$1\times p \quad p\times p \quad p\times 1$

$\in \mathbb{R}$

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

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

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

$tr(AB) = tr(BA)$

$tr(ABC) = tr(BCA) = tr(CAB)$

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

$S = S^T$

$= 数 \sum_{i=1}^{N} tr\left((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)$

$= \sum_{i=1}^{N} tr\left((x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)$

$= tr\left(\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right)$

NS

$= N\,tr(S\Sigma^{-1})$

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

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

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

$\frac{\partial tr(S_1\Sigma^{-1})}{\partial\Sigma^T} = S_1^T(-1)\Sigma^{-2} = -S_1^T\Sigma^{-2}$

$\frac{\partial ① + ②}{\partial\Sigma} = -\frac{1}{2}\left(N\cdot\frac{|\Sigma|\cdot\Sigma^{-1}}{|\Sigma|} - N_1\cdot S_1^T\Sigma^{-2} - N_2\cdot S_2^T\Sigma^{-2}\right)$

$= -\frac{1}{2}\left(N\Sigma^{-1} - N_1S_1^T\Sigma^{-2} - N_2S_2^T\Sigma^{-2}\right) \triangleq 0.$

则有：$N\Sigma - N_1S_1^T - N_2S_2^T = 0$

$$\hat{\Sigma} = \frac{1}{N}(N_1S_1 + N_2S_2)$$

故有
$$\begin{cases}\hat{\mu}_1 = \frac{\sum_{i=1}^{N} y_i x_i}{N_1}\\ \hat{\mu}_2 = \frac{\sum_{i=1}^{N}(1-y_i)x_i}{N_2}\\ \hat{\Sigma} = \frac{1}{N}(N_1S_1+N_2S_2)\\ \hat{\phi} = \frac{N_1}{N}\end{cases}$$

3.7 Naive Bayes classifier

Data: $\{(x_i,y_i)\}_{i=1}^{N}$

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

给定 x, $y=?$ 0/1

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

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

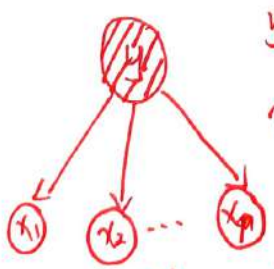
$= \arg\max_y P(y)P(x|y)$

思想：朴素贝叶斯假设

条件独立性假设

最简单的概率图模型

有向图



$y\in\{0,1\}$

$x\in\mathbb{R}^p$

且 $x_i \perp x_j | y \ (i\neq j)$ 相互独立

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

$\begin{cases} x_j 离散 \to x_j \sim Categorical\ Dist \\ x_j 连续 \to x_j \sim N(\mu_j, \sigma_j^2)\end{cases}$

二分类 0/1：$y\sim Bernoulli\ Dist$

多分类：$y\sim Categorical\ Dist$

动机：简化运算

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

做n次 ↑ n=1 ↓ 二项分布

Bernoulli → Binormial etc

& Categorical → Multinormal

↓ 做n次 多项分布

4. Dimensionality Reduction
降维

4.1 Background.

过拟合 {增加data; 正则化 → {Lasso, 岭回归; 降维 ↔ curse of Dimensionality 维度灾难 → 降维方法 {直接降维 → 特征选择; 线性降维 → 特征选择 → PCA, MDS.; 非线性降维 → 流形 {Isomap, LLE → 局部线性嵌入}}}

curse of Dimensionality: 几何角度 (数据稀疏性)

D维:

2维 3维 ... D维

正: 1 1 1

圆: $\pi \cdot 0.5^2$ $\frac{4}{3}\pi \cdot 0.5^3$ ⟸ $k \cdot 0.5^D \to 0$.

① 

$V_{超立方体} = 1$

$V_{超球体} = k \cdot 0.5^D$ $D \to \infty$

说明数据都在球内, 分散的

② 

$r=1$, D维

$V_{外} = k \cdot 1^D = k$

$V_{环形带} = V_{外} - V_{内} = k - k(1-\varepsilon)^D$

则有: $\frac{V_{环}}{V_{外}} = 1 - (1-\varepsilon)^D$, $r=1$, $0<\varepsilon<1$

$= 1 - (1-\varepsilon)^D$

$\lim_{D\to\infty} (1-\varepsilon)^D = 0$, 则有 $\lim_{D\to\infty} \frac{V_{环}}{V_{外}} = 1$

4.2. Sample Mean & Variance matrix

Data: $X = (x_1, x_2 \cdots x_N)^T_{N\times p} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}_{N\times p}$ $1_N = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{N\times 1}$

Sample Mean: $\bar{x}_{p\times 1} = \frac{1}{N}\sum_{i=1}^N x_i = \frac{1}{N} X^T_{p\times N} 1_{N\times 1}$

Sample Covariance: $S_{p\times p} = \frac{1}{N}\sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{N} X^T H X$, $(H = I_N - \frac{1}{N} 1_N 1_N^T)$

$\bar{x}_{p\times 1} = \frac{1}{N}\sum_{i=1}^N x_i = \frac{1}{N}(\underbrace{x_1, x_2 \cdots x_N}_{X^T_{p\times N}}) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{N\times 1} = \frac{1}{N} \cdot X^T \cdot 1_N$

$H = (I_N - \frac{1}{N} 1_N 1_N^T)$

$H^T_{N\times N} = (I_N - \frac{1}{N} 1_N 1_N^T) = H$

$H^2 = H \cdot H = (I_N - \frac{1}{N} 1_N 1_N^T) \cdot (I_N - \frac{1}{N} 1_N 1_N^T)$

$= I_N - \frac{2}{N} 1_N \cdot 1_N^T + \frac{1}{N^2} 1_N 1_N^T 1_N 1_N^T$

$= I_N - \frac{2}{N} 1_N 1_N^T + \frac{1}{N^2} \cdot N\, 1_N 1_N^T$

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

$= H$

则有 $H^n = H$

即 $H = H^T$; $H^2 = H$

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

$= \frac{1}{N}(\underbrace{x_1 - \bar{x}\ \ x_2 - \bar{x} \cdots x_N - \bar{x}}_{(x_1, x_2, \cdots x_N) - (\bar{x}\ \bar{x} \cdots \bar{x})}) \begin{pmatrix} (x_1 - \bar{x})^T \\ (x_2 - \bar{x})^T \\ \vdots \\ (x_N - \bar{x})^T \end{pmatrix}$

$X^T - \bar{x}(1\ 1 \cdots 1)$

$X^T - \bar{x} 1_N^T$

$X^T - \frac{1}{N} X^T 1_N \cdot 1_N^T$ ·

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

H_W → Centering matrix (中心矩阵)

$= \frac{1}{N} X^T H \cdot H^T X = \frac{1}{N} X^T H X$ P13

4.3. PCA - maximum variance perspective

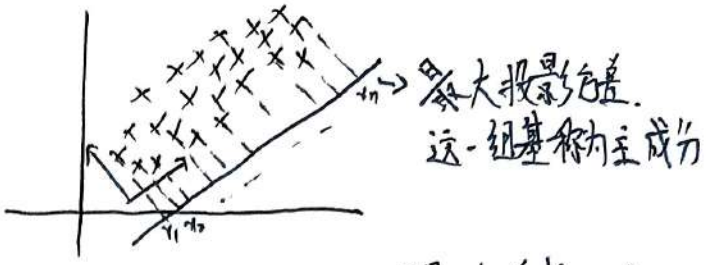
Principal Component Analysis (classical)

一个中心：原始特征空间的重构（相互正交的基）

相关 ⟼ 无关

两个基本点：最大投影方差 ☆-本节重点

最小重构距离 ☆-4.4重点



最大投影方差.

这一组基称为主成分

而把 $y_1, y_2, \cdots, y_n$ 这些投影点重构回去，恢复成原样 x_i 的话就叫最小重构距离

这总结了一点

$$\begin{cases} \hat{u}_1 = \arg\max\ u_1^T S u_1 \\ \text{s.t. } u_1^T u_1 = 1 \end{cases}$$

带约束的优化方法

$$\mathcal{L}(u_1, \lambda) = u_1^T S u_1 + \lambda(1 - u_1^T u_1)$$

$$\frac{\partial \mathcal{L}}{\partial u_1} = 2Su_1 - 2\lambda u_1 \triangleq 0$$

$$Su_1 - \lambda u_1 = 0$$

$$Su_1 = \lambda u_1$$

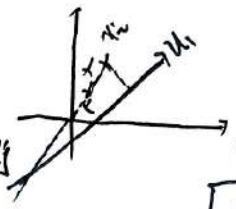
λ → eigen-value（特征值）

u_1 → eigen-vector.

所有点...

$$\begin{cases} \hat{u}_j = \arg\max \sum_{j=1}^{q} u_j^T S u_j \\ \text{s.t. } u_j^T u_j = 1 \end{cases}$$

$$\Rightarrow J = \sum_{i=1}^{p} \lambda_i \quad (\text{大的前}p\text{个})$$



① 中心化（平移）

$x_i - \bar{x}$　　u_1　　$\|u_1\| = 1$

② 投影方差：（已中心化，均值为0就没了）

$$J = \frac{1}{N}\sum_{i=1}^{N}\left((x_i - \bar{x})^T u_1\right)^2$$

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

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

$$= u_1^T S u_1$$

4.4. PCA - minimum error perspective

$$J = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$$ → 前q个都相同

$$= \frac{1}{N} \cdot \sum_{i=1}^{N} \left\| \sum_{\substack{k=q+1 \\ k}}^{p} (x_i^T u_k) u_k \right\|^2$$

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

有：坐标轴 $u_{q+1}, u_{q+2}, \cdots u_p$

坐标 $x_i^T u_{q+1}, \cdots x_i^T u_p$

$a = (a_1, a_2, \cdots a_n)$

$\|a\|^2 = \sum_{i=1}^{n} a_i^2$

中心化：

$$\triangleq \frac{1}{N}\sum_{i=1}^{N}\sum_{k=q+1}^{p} ((x_i - \bar{x})^T u_k)^2$$

$$= \sum_{i=1}^{N} \frac{1}{N} \sum_{k=q+1}^{p} ((x_i - \bar{x})^T u_k)^2 = \sum_{k=q+1}^{p} \sum_{i=1}^{N} \frac{1}{N}((x_i - \bar{x})^T u_k)^2$$

$u_k^T S u_k$

$$= \sum_{k=q+1}^{p} \cdot u_k^T S u_k$$

s.t. $u_k^T u_k = 1$

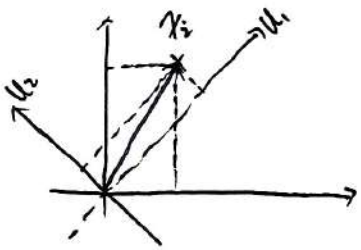
这讨论了所有点

$$\begin{cases} u_k = \arg\min \cdot \sum_{k=q+1}^{p} u_k^T S u_k \\ \text{s.t. } u_k^T u_k = 1 \end{cases}$$

（$u_{q+1}, \cdots, u_p$ 线性相关以下个解，参考4.3. 求最小较小的 λ_i）

重构余量中的 λ

$$J_{\min} = \sum_{k=q+1}^{p} \lambda_k \quad (\text{余量中最小的这些}\lambda)$$



$$x_i = (x_i^T u_1)u_1 + (x_i^T u_2)u_2$$

投影

对于 $x_1, x_2, \cdots, x_p$ 经过去中心化、重构，有 $u_1, u_2, \cdots, u_p$

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

$$\hat{x}_i = \sum_{k=1}^{q} (x_i^T u_{(k)}) \cdot u_k$$ → 重构后前q维

$$J = \frac{1}{N}\sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2$$

N个样本点的最小重构代价

方差矩阵 S 进行特征值分解 $S = GKG^T;\ G^T G = I$

$$K = \begin{bmatrix} k_1 & & \\ & k_2 & \\ & & \ddots \\ & & & k_p \end{bmatrix} \quad k_1 \geq k_2 \geq \cdots \geq k_p$$

4.5. PCA - SVD perspective (SVD: 奇异值分解)

参考 4.2 中数据格式

中心化: $H X = U \Sigma V^T$

$S = \frac{1}{n} X^T H X$

为方便推导, 先省略 $\frac{1}{n}$

则 $S = X^T H X = X^T H^T H X = V \Sigma U^T \cdot U \Sigma V^T = V \cdot \Sigma^2 V^T$ → 特征分解

$V^T V = I$

$V^T V = V V^T = I$

Σ 对角

$H^n = H$

$H^T = H$

V 是特征向量 代表方向

方向. 坐标为投影: $HX \cdot V$

由上节 4.4. $S = G k G^T$ 得: $G = V$, $k = \Sigma^2$ ☆

① 先进行中心化; ② 再进行奇异值分解; 效果相同

$T \triangleq H X X^T H^T = U \Sigma V^T \cdot V \Sigma U^T = U \Sigma^2 U^T$

T 和 S 有相同的 eigenvalue

eigen-vector ↑

S: 特征分解: 得到方向(主成分). 然后 $HX \cdot V$ → 坐标: $U \Sigma V^T V = U \Sigma$

T: 特征分解: 直接得到坐标. 称为主坐标分析 (principle coordinate analysis) PCoA

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

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

$= U \Sigma^3$

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

- Σ^2: 特征值
- $U\Sigma$: 特征向量

, 而有: 坐标: $HX \cdot V = U \Sigma$ 知

此方式下的特征向量就是坐标

对 T 作得到 $U\Sigma$, 而 $U\Sigma$ 刚好就为坐标矩阵

坐标矩阵 ↑

特征向量 组成矩阵

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

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

$= U \Sigma^3 = U \Sigma \Sigma^2$

特征值

根据 n 与 p 的大小适当选取 PCA 或 PCoA. 维度高用 PCoA.

4.6 PPCA - probabilistic PCA

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

↓ observed data (观测得到的数据)

↳ latent variable (指没观测的变量)

$\begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \ddots & \\ & & & \sigma^2 \end{bmatrix}$ → 各向同性

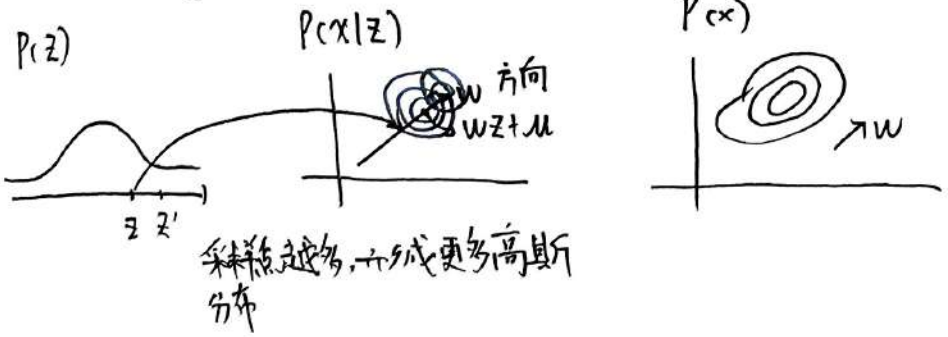
$z \sim N(0_{q \times 1}, I_{q \times q})$, $x = W z + u + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_p)$

(W: $p \times q$, z: $q \times 1$)

Linear Gaussian Model

z, $x|z$, x, $z|x$

P-PCA
- Inference: $P(z|x)$ → ① 求 $P(z|x)$
- Learning: W, u, σ^2 → EM (期望最大) → ② 计算参数



采样点越多, 合成更为高斯分布

$z \sim N(0, I)$

$x = Wz + \mu + \varepsilon$

$\varepsilon \sim N(0, \sigma^2 I)$

$z \perp \varepsilon$

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

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

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

$E(x) = E[Wz+\mu+\varepsilon] = E[Wz+\mu] + E[\varepsilon] = \mu$, $Var[x] = Var[Wz+\mu+\varepsilon] = Var[Wz+\mu] + Var[\varepsilon]$

$= Var[Wz] + Var[\varepsilon] = WIW^T + \sigma^2 I = WW^T + \sigma^2 I$

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

$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$, $x \sim N\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right)$ ① 下面为中公式回忆

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

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

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

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

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

$\begin{cases} E[x_b|x_a] = E[x_{b\cdot a}] + \Sigma_{ba}\Sigma_{aa}^{-1} x_a = \mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1} x_a = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a) \\ Var[x_b|x_a] = Var[x_{b\cdot a}] = \Sigma_{bb\cdot a} \end{cases}$

$x_b|x_a \sim N(\mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a), \Sigma_{bb\cdot a})$ ②

$\begin{pmatrix} x \\ z \end{pmatrix} \sim N\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix},\right.$

P14 参考

$Cov = E[(x-\mu)(z-0)^T] = E[(x-\mu)\cdot z^T] = E[(Wz+\varepsilon)\cdot z^T] = E[Wz\cdot z^T] + \underbrace{E[\varepsilon z^T]}_{E[\varepsilon]\cdot E[z]^T = 0} = E[Wz\cdot z^T]$

$= WE[z\cdot z^T] = W\cdot I = W$

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

利用①② 即可求出 $x_b|x_a$ 即 $x|z$

5. Support Vector Machine

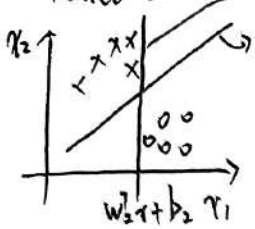
5.1 hard-margin SVM - model definition

SVM有三宝，间隔，对偶，核技巧

hard-margin SVM → 最大间隔分类器

soft-margin SVM

kernel SVM → 非线性，泛化误差不好，对噪声很敏感，运算慢



$f(w) = sign(w^T x + b)$ → 判别模型

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$

$\max margin(w,b)$ s.t. $\begin{cases} w^T x_i + b > 0, & y_i = +1 \\ w^T x_i + b < 0, & y_i = -1 \end{cases}$ ① ⇔ 等价 ⇒ $y_i(w^T x_i + b) > 0$ ②, $i = 1, 2, \cdots, n$

$margin(w,b) = \min_{\substack{w,b \\ x_i \\ i=1,2,\cdots,n}} distance(w, b, x_i)$ ← 点到直线距离中最小的那个

$= \min_{\substack{w,b,x_i \\ i=1,\cdots,n}} \frac{1}{\|w\|} |w^T x_i + b|$

$distance = \frac{1}{\|w\|} |w^T x_i + b|$

$\Rightarrow \max_{w,b} \min_{\substack{x_i \\ i=1,\cdots,n}} \frac{1}{\|w\|} |w^T x_i + b| \Rightarrow \max_{w,b} \min_{\substack{x_i \\ i=1,\cdots,n}} \frac{1}{\|w\|} y_i(w^T x_i + b)$

约束 → $y_i(w^T x_i + b) > 0$ 总可以调整w,b对应r=1，r=1 不影响

$\Rightarrow \max_{w,b} \frac{1}{\|w\|} \min_{\substack{x_i \\ i=1,\cdots,n}} y_i(w^T x_i + b)$

$y_i(w^T x_i + b) > 0 \Rightarrow \exists r > 0, s.t. \min y_i(w^T x_i + b) = r$

$\Rightarrow \max_{w,b} \frac{1}{\|w\|}$ s.t. $\min y_i(w^T x_i + b) = 1$

↳ $y_i(w^T x_i + b) \geq 1, i = 1, \cdots, n$

$\Rightarrow$ $\min_{w,b} \|w\| \Rightarrow \min_{w,b} \frac{1}{2} w^T w$ 二次

s.t. $y_i(w^T x_i + b) \geq 1$, for $\forall i = 1, \cdots, n$ N个约束

convex optimization

5.2 hard-margin SVM - model solution

Data $= \{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$, $\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{pmatrix}$

① $\begin{cases} \min_{w,b} \frac{1}{2} w^T w \\ s.t. \ y_i(w^T x_i + b) \geq 1 \Leftrightarrow 1 - y_i(w^T x_i + b) \leq 0 \\ \text{for } i = 1, 2, \cdots, N \end{cases}$ 带约束

$\mathcal{L}(w, b, \lambda) = \frac{1}{2} w^T w + \sum_{i=1}^N \lambda_i \underbrace{[1 - y_i(w^T x_i + b)]}_{\leq 0}$ ($\lambda_i \geq 0$) 无约束

① ② → Primal Problem

② $\begin{cases} \min_{w,b} \max_{\lambda} \mathcal{L}(w, b, \lambda) \\ s.t. \ \lambda_i \geq 0 \end{cases}$ 这是一种化简优化问题的思想

①② 是等价的

解释表述：

在 $\lambda_i \geq 0$ 时，

如果 $1 - y_i(w^T x_i + b) > 0$, $\max_{\lambda} \mathcal{L}(w, b, \lambda) = \frac{1}{2} w^T w + \infty = \infty$

如果 $1 - y_i(w^T x_i + b) \leq 0$, $\max_{\lambda} \mathcal{L}(w, b, \lambda) = \frac{1}{2} w^T w$, $\Leftrightarrow \min_{w,b} \max_{\lambda} \mathcal{L}(w, b, \lambda) = \min_{w,b} \frac{1}{2} w^T w$

在约化问题②中，$\Delta > 0$ 中的 (w,b) 解会被舍，且 w^*, b^* 一定出自于 $\Delta \leq 0$.

利用强对偶 $\begin{cases}\max\limits_{\lambda}\min\limits_{w,b}\mathcal{L}(w,b,\lambda)\\ s.t.\ \lambda_i\geq 0.\end{cases}$ → dual problem

$\min\max\mathcal{L}\geq\max\min\mathcal{L}$ 弱对偶关系

$\min\max\mathcal{L}=\max\min\mathcal{L}$ 强对偶关系

梯中，凸优化，$\frac{1}{2}w^Tw$ 二次，且二次优化问题，无互满足强对偶关系约束线性，需证明，参看优化

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

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

$\mathcal{L}(w,b,\lambda)=\frac{1}{2}w^Tw+\sum_{i=1}^{N}\lambda_i-\sum_{i=1}^{N}\lambda_i y_i w^Tx_i-\underbrace{\sum_{i=1}^{N}\lambda_i b y_i}_{\sim①\ \therefore=0}$

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

$\frac{\partial\mathcal{L}}{\partial w}=\frac{1}{2}\cdot 2\cdot w-\sum_{i=1}^{N}\lambda_i y_i x_i\triangleq 0\Rightarrow\boxed{w=\sum_{i=1}^{N}\lambda_i y_i x_i}$，将其代入 $\mathcal{L}(w,b,\lambda)$

$\mathcal{L}(w,b,\lambda)=\frac{1}{2}(\sum_{i=1}^{N}\lambda_i y_i x_i)^T(\sum_{j=1}^{N}\lambda_j y_j x_j)-\sum_{i=1}^{N}\lambda_i y_i(\sum_{j=1}^{N}\lambda_j y_j x_j)^T\cdot x_i+\sum_{i=1}^{N}\lambda_i$

$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^Tx_j-\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_j^Tx_j+\sum_{i=1}^{N}\lambda_i$

$=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^Tx_j+\sum_{i=1}^{N}\lambda_i$

$\begin{cases}\max\limits_{\lambda}\mathcal{L}(②)(w,b,\lambda)\\ s.t.\ \lambda_i\geq 0\end{cases}\Rightarrow\begin{cases}\min\limits_{\lambda}\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^Tx_j-\sum_{i=1}^{N}\lambda_i\\ s.t.\ \lambda_i\geq 0\\ \sum_{i=1}^{N}\lambda_i y_i=0\end{cases}$

解出 w^*,b^*

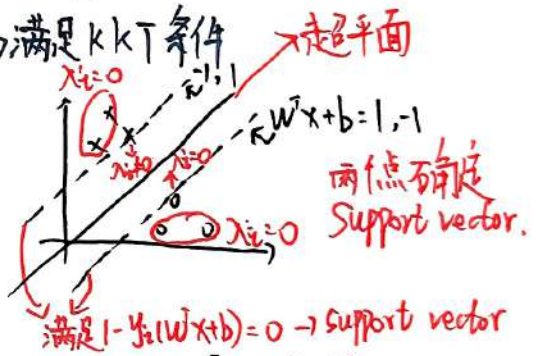
KKT条件：(Kamush-Kuhn-Tucker) → 原对偶问题具有强对偶关系 $\Leftrightarrow$ 满足KKT条件

$\begin{cases}\frac{\partial\mathcal{L}}{\partial w}=0,\ \frac{\partial\mathcal{L}}{\partial b}=0\\ \lambda_i[1-y_i(w^Tx_i+b)]=0\\ \lambda_i\geq 0\\ 1-y_i(w^Tx_i+b)\leq 0\end{cases}$ slackness complementary 松弛互补

→ 当 $1-y_i(w^Tx_i+b)=0$ 时，λ_i 不等于0.

Support vector

$\boxed{w^*=\sum_{i=1}^{N}\lambda_i x_i y_i,\ b^*=y_k-\sum_{i=1}^{N}\lambda_i x_i^Ty_i x_k}$ 对偶最优解，最优拉格朗日算子 $y_k=\pm 1$



$\exists(x_k,y_k)$, s.t. $1-y_k(w^Tx_k+b)=0\Leftrightarrow y_k(w^Tx_k+b)=1\Leftrightarrow y_k^2(w^Tx_k+b)=y_k\Leftrightarrow w^Tx_k+b=y_k$

$\therefore b^*=y_k-w^Tx_k=y_k-\sum_{i=1}^{N}\lambda_i x_i^Ty_i x_k$

$\therefore f(x)=\text{sign}(w^*x+b^*)$，决策平面 $w^{*T}x+b^*$

w^*：是data的线性组合，只有在Support vector上的点才会对 w^* 有用，不在的都会让 $\lambda_i=0$，这是KKT条件.

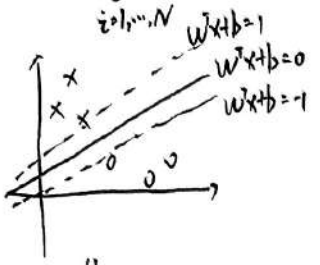
5.3 Soft-Margin SVM

soft: 允许一点点错误: $\min \frac{1}{2}w^Tw + loss$

Hard-margin SVM:

Data $= \{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$

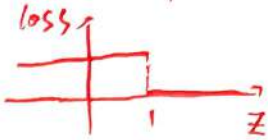
$$\begin{cases} \min \frac{1}{2}w^Tw \\ \text{s.t. } y_i(w^Tx_i+b) \geq 1 \quad i=1,\cdots,N \end{cases}$$



Loss difination: 指示函数

① $Loss = \sum_{i=1}^N I\{y_i(w^Tx_i+b) < 1\}$ ⇒ loss 0/1: $\begin{cases} 1, z<1 \\ 0, otherwise \end{cases}$

关于 w 不连续



跳跃

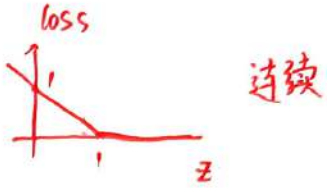
令 $z = y_i(w^Tx_i+b)$ 点数

② loss: 距离 (hinge loss) 较连续

$$\begin{cases} \text{if } y_i(w^Tx+b) \geq 1, \ loss = 0 \\ \text{if } y_i(w^Tx+b) < 1, \ loss = 1 - y_i(w^Tx+b) \end{cases}$$

$$\Rightarrow Loss = \max\{0, 1 - y_i(w^Tx+b)\}$$

$$Loss_{max} = \max\{0, 1-z\} \rightarrow$$



连续

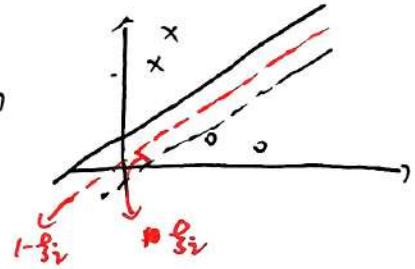
⇓

Soft-margin SVM (Hinge Loss):

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}w^Tw + C\underbrace{\sum_{i=1}^N \max\{0, 1-y_i(w^Tx_i+b)\}}_{Loss} \\ \text{s.t. } y_i(w^Tx_i+b) \geq 1 \quad i=1,\cdots,N \end{cases}$$

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

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}w^Tw + C\sum_{i=1}^N \xi_i \\ \text{s.t. } y_i(w^Tx_i+b) \geq 1-\xi_i \\ \xi_i \geq 0 \end{cases}$$



求解过程与 H-SVM 雷同

① Lagrange function

② Dual function

③ KKT condition

④ Lagrange Multiplier

5.4 Constraint Optimization - Weak Duality

Primal Problem:

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} f(x) \\ \text{s.t. } m_i(x) \leq 0, \ i=1,\cdots,M \\ \text{等价 } n_j(x) = 0, \ j=1,\cdots,N \end{cases}$$

Lagrangian function:

$$\mathcal{L}(x, \lambda, \eta) = f(x) + \sum_{i=1}^M \lambda_i m_i + \sum_{j=1}^N \eta_j n_j$$

Dual problem (原问题的无约束形式) (关于 x 的函数)

$$\begin{cases} \min\limits_x \max\limits_{\lambda,\eta} \mathcal{L}(x, \lambda, \eta) \Rightarrow x \in \{\text{好的 } x \text{ 集合}\} \\ \text{s.t. } \lambda_i \geq 0 \end{cases}$$

$\rightarrow p$

如果 x 违反约束 $m_i(x)$, 即 $m_i(x) > 0$, $\max\limits_{\lambda} \mathcal{L} \rightarrow \infty$

如果 x 符合 $m_i(x) \leq 0$, $\max\limits_{\lambda} \mathcal{L} \rightarrow \pm +\infty$, $\max\limits_{\lambda} \mathcal{L} = f(x)$

$$\min_x \max_\lambda \mathcal{L} = \min_x \{\underbrace{\max_\lambda \mathcal{L}}_{\text{好的 } x}, \underbrace{+\infty}_{\text{坏的 } x}\} = \min_x \max_\lambda \mathcal{L}$$

Dual Problem. (About $\mathcal{L}(x, \lambda, \eta)$)

$d \leq$

$$\begin{cases} \max\limits_{\lambda,\eta} \min\limits_x \mathcal{L}(x, \lambda, \eta) \quad \text{(Dual problem is about } \lambda, \eta \text{'s function)} \\ \text{s.t. } \lambda_i \geq 0 \end{cases}$$

Weak Duality:

Dual problem $\leq$ Primal problem

$d \leq p$

Proof: $\max \min \mathcal{L} \leq \min \max \mathcal{L}$

即 $\max\limits_{\lambda,\eta} \min\limits_x \mathcal{L}(x, \lambda, \eta) \leq \min\limits_x \max\limits_{\lambda,\eta} \mathcal{L}(x, \lambda, \eta)$

$\min\limits_x \mathcal{L}(x, \lambda, \eta) \leq \mathcal{L}(x, \lambda, \eta) \leq \max\limits_{\lambda,\eta} \mathcal{L}(x, \lambda, \eta)$

则有 $A(\lambda, \eta) \leq B(x)$

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

$$\max_{\lambda,\eta} \min_x \mathcal{L}(x, \lambda, \eta) \leq \min_x \max_{\lambda,\eta} \mathcal{L}(x, \lambda, \eta)$$

This complete the proof

5.5. 对偶性的几何解释：

$$\begin{cases} \min\limits_{x\in\mathbb{R}^n} f(x) \\ \text{s.t. } m_i(x)\le 0 \end{cases}$$ D·定义域，$D=\mathrm{dom} f \cap \mathrm{dom} m_i$

便于讨论，只取不等式约束，且只取一个条件

$L(x,\lambda)=f(x)+\lambda m_1(x)$，$\lambda\ge 0$.

$p^*=\min f(x)$ （原问题的最优解）

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

$G=\{(m_1(x),f(x))\,|\,x\in D\}=\{(u,t)\,|\,x\in D\}$

$p^*=\inf\{t\,|\,(u,t)\in G, u\le 0\}$（纵轴上的t），$d^*=\max\limits_{\lambda} g(\lambda)$

p^*就是投影的inf

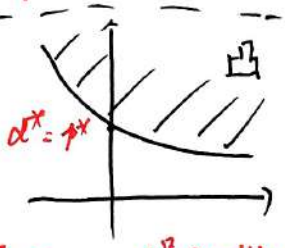
$d^*=\max\limits_{\lambda}\min\limits_{x} L(x,\lambda)=\max\limits_{\lambda}\underbrace{\min\limits_{x}(t+\lambda u)}_{=g(\lambda)}=\max\limits_{\lambda} g(\lambda)$，

$g(\lambda)=\inf\{t+\lambda u\,|\,(u,t)\in G\}$ ⇒ 一次函数与G的交集的inf.

一次函数，含$t+\lambda u=\Delta$

可以任取t.λ去平移、相交

此情况 $d^*\neq p^*$

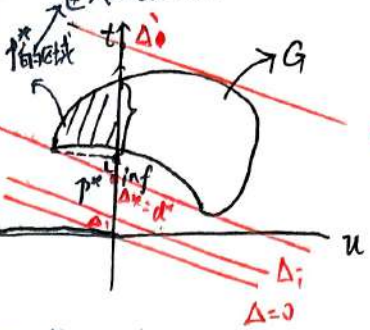


↑↑↑ 强Duality

区域在t上的投影，$\therefore p^*=\inf\{t\,|\,(u,t)\in G, u\le 0\}$

取t.

$p^*=\inf$ 取不确界的t



Δ'不在$g(\lambda)$中，也不是该λ.t组合下的inf Δ.

由于$t+\lambda u=\Delta$，可知$g(\lambda)$就是

$\inf\{\Delta\,|\,(u,t)\in G\}$

取且Δ是有限的，由图上知，不同的λ.u.直线会平移产生不同的Δ，

即同一个斜率下得到的第一个切线，就是$g(\lambda)$的集合一元素，而不同斜率所得到的可数的$\{\Delta_1,\cdots,\Delta_n\}$就是$g(\lambda)$当t

而$d^*=\max g(\lambda)$，可知，由图$\Delta^*=d^*$

两点确定一条直线时，

这就是Weak Duality.

5.6. Slater Condition - Duality Problem

slater condition：

$\exists\,\hat{x}\in \mathrm{relint}\, D$, s.t. $\forall i=1,\cdots,M$, $m_i(\hat{x})<0$

凸 + slater ⇒ 强 Duality

还可以是：else conditions.

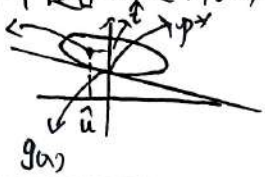
relint：relative interior：explain.

有边界就删去边界，没有边界更好，之后取内部区域

凸二次规化：f是凸，m_i仿射 n_j仿射 天然满足②

① 对于大多数凸优化，slater成立

② 放松的slater：if M中有k个仿射函数，则只需找忽略掉的M-k个是否满足$m_i(x)\le 0$



区域或左半边本包含边界的区域中存在一个点

5.7 karush - kuhn - Tucker (KKT) Condition

$$\begin{cases} \min f(x), & f(x)\text{可微} \\ \text{s.t. } m_i(x)\le 0, & i=1,\cdots,m \\ n_j(x)=0, & j=1,\cdots,N \end{cases}$$

$\lambda=\begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_M\end{pmatrix}$ $y=\begin{pmatrix}y_1\\ \vdots\\ y_N\end{pmatrix}$

$p^*\to x^*$ 时

$L(x,\lambda,y)=f(x)+\sum_{i=1}^{M}\lambda_i m_i+\sum_{j=1}^{N} y_j n_j$

$g(\lambda,y)=\min\limits_{x} L(x,\lambda,y)$

Dual problem：

$$\begin{cases} \max\limits_{\lambda,y} g(\lambda,y) \\ \text{s.t. } \lambda\ge 0 \end{cases}$$ $d^*\to\lambda^*,y^*$ 时

convex + slater ⇒ strong Duality ⇔ KKT (To solve the λ^*,y^*,x^*)

($d^*=p^*$)

KKT：
- 可行条件：$\begin{cases} m_i(x^*) \le 0 \ ① \\ n_j(x^*) = 0 \ ② \\ \lambda^* \ge 0 \ ③ \end{cases}$
- 互补松弛性：$\lambda_i^* m_i = 0 \ (\forall i=1,\cdots,M) \ ④$
- 梯度为零：$\left.\frac{\partial \mathcal{L}(x,\lambda^*,\gamma^*)}{\partial x}\right|_{x=x^*} = 0 \ ⑤$

互补松弛性与梯度为零都是从不等号变等号得出

$d^* = \max_{\lambda,\gamma} g(\lambda,\gamma) = g(\lambda^*,\gamma^*)$

$= \min_x \mathcal{L}(x,\lambda^*,\gamma^*)$

$\le \mathcal{L}(x,\lambda^*,\gamma^*) \quad \forall x \in \text{dom} f$，$\Rightarrow x^*$也成立 → 取等号

$\min \mathcal{L}(x,\lambda^*,\gamma^*) = \mathcal{L}(x^*,\lambda^*,\gamma^*)$

故 $\left.\frac{\partial \mathcal{L}}{\partial x}\right|_{x=x^*} = 0$，$\nabla f|_{x=x^*} = 0$

strong $= \mathcal{L}(x^*,\lambda^*,\gamma^*)$

duality $= f(x^*) + \sum_{i=1}^{M} \lambda_i^* m_i + \sum_{j=1}^{N} \gamma_j^* n_j$

$\le f(x^*)$（$\lambda_i \ge 0$，$m_i \le 0$；$n_j = 0$）这里只能取等号 → $\sum_{i=1}^{M} \lambda_i^* m_i = 0 \Rightarrow \lambda_i m_i = 0 \ (\forall i=1,\cdots,M)$

$= p^*$

---

6. Kernel Method

6.1 Kernel Method Introduction

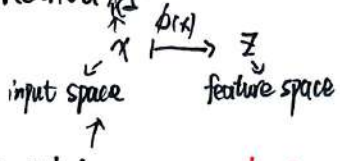
kernel Method → 从思想角度

kernel Trick → 从计算角度

kernel Function

① 非线性带来高维转换（从模型角度） $x \to \phi(x)$

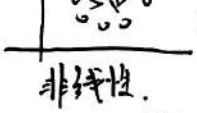
② 对偶表示带来内积（从优化角度） $x_i^T x_j$



线性可分	一点点错误	严格非线性
PLA（与初始值相关）	Pocket Algorithm	$\phi(x)$ + PLA ① ②
Hard-Margin SVM	soft-Margin SVM	$\phi(x)$ + Hard-Margin SVM ① ②（只①不②✓；①②同时到达这里⑥）

看①：

关于这类：

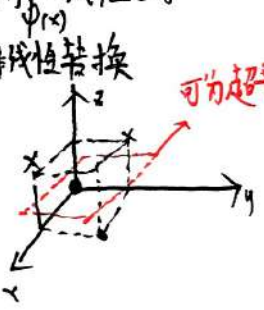


非线性.

① PLA → 多层感知机（神经网路）→ DP

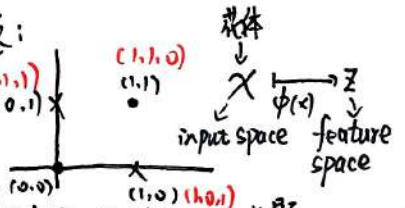
② 非线性可分 → 线性可分

非线性转换



转换举例

异或问题：



① PLA ✗ DP ✓ ② 转换

此题：$x = (x_1, x_2) \xrightarrow{\phi(x)} Z(x_1, x_2, |x_1 - x_2|)^T$

dim=2 → dim=3

Cover Theorem：高维比低维更易线性可分

困难：① $\phi(x)$ 难求；② $\phi(x_i)^T \phi(x_j)$ 计算大.

直接求内积，而不是求 $\phi(x)$，再求 $\langle \phi(x), \phi(x) \rangle$

Kernel Function：

$K(x,x') = \phi^T(x)\phi(x) = \langle \phi(x), \phi(x') \rangle$ 利用 kernel Function 直接求内积，而不是求 $\phi(x)$，再求 $\langle \phi(x), \phi(x) \rangle$

$\forall x, x' \in \mathcal{X}$，$\exists \phi: \mathcal{X} \to Z$ s.t. $K(x,x') = \phi(x)^T\phi(x)$

对称 $K(x,x')$ 是一个核函数

eg. $K(x,x') = \exp\left\{-\frac{\|x-x'\|^2}{2\sigma^2}\right\}$ x, x' 为样本 "Gauss kernel Function"

这个skill：kernel Trick，省去找 $\phi(x)$，求 $\phi^T(x)\phi(x)$ 或径向基函数（RBF）

kernel Function 适用场景：线性不可分，需要通过低维到高维变换，但 $\phi(x)$ 难求，直接使用 kernel Function，eg：上例的 Gauss Function

看②：Hard-Margin SVM.

Primal problem：

$\begin{cases} \min\limits_{w,b} \frac{1}{2} w^T w \ \text{凸优化} \\ \text{s.t. } y_i(w^T x_i + b) \ge 1 \ (N\text{个约束}) \end{cases}$

$\Rightarrow$

Dual problem：

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

此处 $x_i^T x_j$ 即 $\phi(x_i)^T \phi(x_j)$

由此引出 Kernel Function，不去计算 $\phi(x)$，直接得到 $\phi(x_i)^T \phi(x_j)$

线性核：$K(x_i, x_j) = x_i^T x_j$

多项式核：$K(x_i, x_j) = (x_i^T x_j)^d$ — $d \ge 1$ 为多项式的次数

高斯核：$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$ — $\sigma > 0$ 为高斯核的带宽（width）

拉普拉斯核：$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{\sigma}\right)$ — $\sigma > 0$

Sigmoid 核：$K(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$ — tanh 为双曲正切函数，$\beta > 0$，$\theta < 0$

6.2 Positive Definite kernel

$\mathcal{X}$ → input space

kernel function: $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $\forall x, z \in \mathcal{X}$, 则称 $k(x,z)$ 为 kernel function

Positive Definite kernel: $K: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, $\forall x, z \in \mathcal{X}$,

①定义 有 $k(x,z)$, if $\exists \phi: \mathcal{X} \mapsto \mathbb{R}$, $\phi \in \mathcal{H}$ → (Hilbert space)

s.t. $k(x,z) = \langle \phi(x), \phi(z) \rangle$, 那么则称 $k(x,z)$ 为正核函数

Positive Definite kernel: $K: \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, $\forall x, z \in \mathcal{X}$

②定义 有 $k(x,z)$. if $k(x,z)$ 满足如下两条性质:

① 对称性

② 正定性

那么则称 $k(x,z)$ 为正定核函数

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

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 matrix 是半正定的

$K = [k(x_i, x_j)]$

$\lim_{n \to \infty} k_n = k \in \mathcal{H}$ 对极限操作是封闭的

元素为函数

introduce Hilbert space: 完备的, 可能是无限维的, 被赋予内积运算的线性空间

$f, g \in \mathcal{H}$

向量空间 (保加法, 保数乘)

{ 对称性 / 正定性 / 线性

$\langle f, g \rangle = \langle g, f \rangle$

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

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

证明①②等价: 由于对称性明显, 这里讨论正定性:

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

$\Rightarrow$

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: 内积拥有对称性质, 即 $\langle \phi(x), \phi(z) \rangle = \langle \phi(z), \phi(x) \rangle$

$\therefore k(x,z) = k(z,x)$

$\therefore k(x,z)$ 满足对称性质

欲证 Gram matrix 半正定:

Gram matrix: $K = [k(x_i, x_j)]_{N \times N}$ $\quad k_{ij} = k(x_i, x_j)$

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

$$\alpha^T \cdot K \cdot \alpha = (\alpha_1 \ \alpha_2 \cdots \alpha_N) \begin{pmatrix} k_{11} & k_{12} & \cdots & k_{1N} \\ k_{21} & k_{22} & \cdots & k_{2N} \\ \vdots & \vdots & & \vdots \\ k_{N1} & k_{N2} & \cdots & k_{NN} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j k_{ij} = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle = \sum_{i=1}^N \alpha_i \phi(x_i)^T \cdot \sum_{j=1}^N \alpha_j \cdot \phi(x_j)$$

$$= \left[\sum_{i=1}^N \alpha_i \phi(x_i)\right]^T \sum_{j=1}^N \alpha_j \phi(x_j) = \left\langle \sum_{i=1}^N \alpha_i \phi(x_i), \sum_{j=1}^N \alpha_j \phi(x_j) \right\rangle = \left\| \sum_{i=1}^N \alpha_i \phi(x_i) \right\|^2 \geq 0$$

$\therefore K$ 是半正定的

$\Leftarrow$

对 K 进行分解, 对于对称矩阵 $K = V \Lambda V^T$, 那么令 $\phi(x_i) = \sqrt{\lambda_i} V_i$, 其中 V_i 是特征向量, 于是就构造了 $k(x,z) = \sqrt{\lambda_i} V_i^T \sqrt{\lambda_j} V_j = \langle \phi(x_i), \phi(x_j) \rangle$

此时 Λ 为对角矩阵 $\Lambda = E$

小结: 对于严格可分的 Data, Hard-Margin SVM 选定一个超平面, 保证所有数据到这个超平面距离最大, 对这个平面施加约束, 即 $y_i(W^T x_i + b) = 1$, 得到一个 convex optimization prob, 并且所有约束 conditions 都是仿射 function, 且满足 slate 条件, 将 Primal prob $\Rightarrow$ Dual prob, 得到等价解, 并求出约束数: $\max_{\lambda} -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^N \lambda_i$, s.t. $\lambda_i \geq 0$.

对超平面参数的求解采用 strong Duality's KKT Condition.

$$\begin{cases} \frac{\partial L}{\partial w} = 0, \ \frac{\partial L}{\partial b} = 0 \\ \lambda_k (1 - y_k (w^T x_k + b)) = 0 \quad \text{slackness complementary} \\ \lambda_i \geq 0 \\ 1 - y_i (w^T x_i + b) \leq 0 \end{cases} \Rightarrow \text{解:} \begin{cases} \hat{w} = \sum_{i=1}^N \lambda_i y_i x_i \\ \hat{b} = y_k - w^T x_k = y_k - \sum_{i=1}^N \lambda_i y_i x_i^T x_k, \ \exists k, \ 1 - y_k (w^T x + b) = 0 \end{cases}$$

当允许一点错误, Soft-Margin SVM, 加入 Hinge Function 容错大计, $\Rightarrow \operatorname{argmin}_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i$ s.t. $y_i (W^T x_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$, $i = 1, \cdots N$

对完全不可分, 采用特征转换, 在 SVM 中, 引入 Positive Definite kernel 对内积进行变换, 只要满足对称、正定性, 就可以做核技巧

7. Exponential Family Distribution

充分统计量 sufficient statistics → online learning

7.1. Background

$p(x|\eta) = h(x)\exp(\eta^T\phi(x) - A(\eta))$ 指数族分布

η: 参数向量, $x \in \mathbb{R}^n$ $\eta = \eta(\theta)$ $\theta \to \eta$ $p(x|\eta)$是概率密度函数

$A(\eta)$: log partition function 配分函数 $A(\eta) = A(\eta(\theta))$

指数族分布 → 充分统计量 / 共轭 / 最大熵 / 广义线性模型 / 概率图模型 / 变分推断

$p(x|\theta) = \frac{1}{z}\hat{p}(x|\theta)$ 归一化因子: $z = \int_x \hat{p}(x|\theta)dx$ 配分函数

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

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

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

$p(x|\eta) = h(x)\cdot\exp(\eta^T\phi(x))\cdot\exp(-A(\eta)) = \frac{1}{\exp(A(\eta))}h(x)\cdot\exp(\eta^T\phi(x)) = \frac{1}{z}\hat{p}(x|\eta)$

$\exp(A(\eta)) = z \Rightarrow A(\eta) = \log z$ ↘ partition function $\Rightarrow$ log partition function

广义线性模型: 线性组合 w^Tx; link function: $\eta(\cdot)$ 指数族分布; $y(x,w)$ 指数族分布

线性回归: $y|x,w \sim \mathcal{N}(w^Tx, \beta^{-1})$

分类: $y|x$ Bernoulli

共轭: $p(z|x) \propto p(x|z)\,p(z)$ Beta / 二项分布 / Beta

prior: ①共轭 → 计算方便 ②最大熵 → 无信息先验 ③Jeffrey

7.2. Gaussian Distribution

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

$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2)\right\} = \exp\log(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(x^2 - 2\mu x) - \frac{\mu^2}{2\sigma^2}\right\}$

$= \exp\log(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}\begin{pmatrix}-2\mu & 1\end{pmatrix}\begin{pmatrix}x\\x^2\end{pmatrix} - \frac{\mu^2}{2\sigma^2}\right\} = \exp\left\{\underbrace{\begin{pmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{pmatrix}}_{\eta^T}\underbrace{\begin{pmatrix}x\\x^2\end{pmatrix}}_{\phi(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log 2\pi\sigma^2\right)}_{A(\eta)}\right\}$

设 $\eta = \begin{pmatrix}\eta_1\\\eta_2\end{pmatrix} = \begin{pmatrix}\frac{\mu}{\sigma^2}\\-\frac{1}{2\sigma^2}\end{pmatrix} \Rightarrow \begin{cases}\mu = -\frac{\eta_1}{2\eta_2}\\ \sigma^2 = -\frac{1}{2\eta_2}\end{cases} \Rightarrow \frac{\mu^2}{2\sigma^2} = \frac{-\eta_1^2}{4\eta_2}$

$A(\eta) = \frac{-\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right)$; $h(x) = 1$; $\eta^T = \begin{pmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{pmatrix}$; $\phi(x) = \begin{pmatrix}x\\x^2\end{pmatrix}$

故 $p(x|\theta) = \exp\left\{\eta^T\phi(x) - A(\eta)\right\}$, Gauss Distribution is Exponential Family Distribution

7.3 log partition function

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

$\int_{-\infty}^{+\infty} p(x|\eta)dx = 1$, 则两边同时积分 $\Rightarrow$ $\exp(A(\eta)) = \int h(x)\exp(\eta^T\phi(x))dx$ $\Rightarrow$ 两边对η求偏导

$\Rightarrow \exp(A(\eta))\cdot A'(\eta) = \frac{\partial}{\partial\eta}\left(\int h(x)\exp(\eta^T\phi(x))dx\right)$

$= \int h(x)\exp(\eta^T\phi(x))\phi(x)dx$

$A'(\eta) = \frac{\int h(x)\exp(\eta^T\phi(x))\phi(x)dx}{\exp(A(\eta))} = \int h(x)\exp(\eta^T\phi(x) - A(\eta))\cdot\phi(x)dx = \int p(x|\eta)\cdot\phi(x)dx$

$= E_{p(x|\eta)}[\phi(x)]$ ← $\phi(x)$的期望

$\therefore A'(\eta) = E_{p(x|\eta)}[\phi(x)]$; $A''(\eta) = \mathrm{Var}[\phi(x)] \geq 0$ $A(\eta)$ is convex function

验证: $E[\phi(x)] = \begin{pmatrix}E(x)\\E(x^2)\end{pmatrix}$

$E(x)$ 若 $x \sim \mathcal{N}(\mu, \sigma^2)$, 则 $E(x) = \mu$.

检验 $A'(\eta) = ?\ \mu$.

$A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right)$

$\frac{\partial A(\eta)}{\partial\eta_1} = -\frac{\eta_1}{2\eta_2}$, 代入 $\begin{cases}\eta_1 = \frac{\mu}{\sigma^2}\\ \eta_2 = \frac{1}{-2\sigma^2}\end{cases} \Rightarrow A'(\eta) = \mu$ ✓✓

7.4 MLE & Sufficient Statistics

$D = \{x_1, x_2, \cdots, x_N\}$

$$\eta_{MLE} = \arg\max_{\eta} \log P(D|\eta) = \arg\max_{\eta} \log \prod_{i=1}^{N} P(x_i|\eta) = \arg\max_{\eta} \sum_{i=1}^{N} \log p(x_i|\eta)$$

$$= \arg\max_{\eta} \sum_{i=1}^{N} \log [h(x_i) \cdot \exp(\eta^T \phi(x_i) - A(\eta))] = \arg\max_{\eta} \sum_{i=1}^{N} [\log h(x_i) + \eta^T \phi(x_i) - A(\eta)]$$

$$= \arg\max_{\eta} \sum_{i=1}^{N} (\eta^T \phi(x_i) - A(\eta))$$

$$\frac{\partial}{\partial \eta}\left(\sum_{i=1}^{N} (\eta^T \phi(x_i) - A(\eta))\right) = \sum_{i=1}^{N} \frac{\partial}{\partial \eta} (\eta^T \phi(x_i) - A(\eta)) = \sum_{i=1}^{N} [\phi(x_i) - A'(\eta)] = \sum_{i=1}^{N} \phi(x_i) - N A'(\eta) \triangleq 0$$

$$A'(\eta)_{MLE} = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i)$$ ⇒ $A'(\eta)$ 关于 η 的函数，则 $\eta = (A')^{-1}(\eta)$ 反函数，解出 $\eta_{MLE} = A'^{(-1)}(\eta)$

$\phi(x_i)$: sufficient statistics

7.5 Maximum Entropy Perspective

信息量：$-\log p$

熵：$E_{p(x)}[-\log p] = \int -p(x) \cdot \log p(x)\, dx$

若离散 $= -\sum_{x} p(x) \cdot \log p(x)$

$H[p] = -\sum_{x} p(x) \log p(x)$

最大熵 ⟺ 等可能

假设 x 是离散的

x	1	2	$\cdots$	k
p	p_1	p_2	$\cdots$	p_k

$\sum_{i=1}^{k} = 1$.

$$\max H[p] = \max -\sum_{i=1}^{k} p_i \log p_i \Leftrightarrow \min \sum_{i=1}^{k} p_i \log p_i$$

$$\text{s.t.} \sum_{i=1}^{k} p_i = 1, \quad p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_k \end{pmatrix}$$

→ 均匀分布 $p(x)$

对某个 p_i 求导（因为 p_i 是离散型）

$$\frac{\partial \mathcal{L}}{\partial p_i} = \log p_i + 1 - \lambda \triangleq 0$$

$$\hat{p}_i = \exp(\lambda - 1)$$

∵ $\hat{p}_i$ = constant

∴ $\hat{p}_1 = \hat{p}_2 = \cdots = \hat{p}_k = \frac{1}{k}$

∴ $p(x)$ 是均匀分布

在没有任何已知情况下，熵达到最大的分布是均匀分布

$\hat{p}_i = \arg\max H[p]$，Lagrangian Function

$\hat{p}_i = \arg\min \sum_{i=1}^{k} p_i \log p_i$

$$\therefore \mathcal{L}(\eta, \lambda) = \sum_{i=1}^{k} p_i \log p_i + \lambda \left(1 - \sum_{i=1}^{k} p_i\right)$$

7.6 Maximum Entropy Perspective

$$P(x|\eta) = h(x) \exp\{\eta^T \phi(x) - A(\eta)\} = \frac{1}{z(\eta)} h(x) \exp\{\eta^T \phi(x)\}$$

最大熵原理（满足已知事实的约束）

最大熵 ⟺ 等可能

Data: $\{x_1, x_2, \cdots, x_N\}$ 既定事实

Empirical Distribution 经验分布 $\hat{p}(X = x) = \hat{p}(x) = \frac{count(x)}{N}$，eg: $\hat{p}(x = x_i) = \frac{count(x_i)}{N}$

$E_{\hat{p}}[x]$, $Var_{\hat{p}}[x]$

$f(x)$ 是任意关于 x 的函数向量 $f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_Q(x) \end{pmatrix}$, $\Delta = \begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_Q \end{pmatrix}$

$$E_{\hat{p}}[f(x)] = \Delta \rightarrow (\text{已知})$$

$p(x) \to p(x)$

$H[p] = -\sum_{x} p(x) \log p(x)$

则最大熵原理：

$$\begin{cases} \min \sum_{x} p(x)\log p(x) \\ \text{s.t. } \sum_{x} p(x) = 1,\ E_p[f(x)] = E_{\hat{p}}[f(x)] = A \text{ (已constant)} \end{cases} \to p(x) \text{ 指数族分布}$$

比7.5多出的一个条件：服从事实

$$L(p, \lambda_0, \lambda) = \sum_x p(x)\log p(x) + \lambda_0\left(1 - \sum_x p(x)\right) + \lambda^T\left(A - E_p[f(x)]\right), \quad E_p[f(x)] = \sum_x p(x) f(x)$$

$$\frac{\partial L}{\partial p(x)} = \sum_x [\log p(x) + 1] - \sum_x \lambda_0 - \sum_x \lambda^T f(x) \triangleq 0$$

$$\sum_x [\log p(x) + 1 - \lambda_0 - \lambda^T f(x)] = 0$$

（$\log p(x)$ 项为 0）

$$\therefore \log p(x) + 1 - \lambda_0 - \lambda^T f(x) = 0$$

$$\log p(x) = \lambda^T f(x) + \lambda_0 - 1$$

$$p(x) = \exp\{\lambda^T f(x) + \lambda_0 - 1\} = \exp\left\{\underbrace{\lambda^T}_{\eta^T}\underbrace{f(x)}_{\phi(x)} - \underbrace{(-\lambda_0 + 1)}_{A(\eta)}\right\} = \exp\{\eta^T \phi(x) - A(\eta)\}$$

$$= \exp\{(\lambda_0, \lambda)(1, f(x))^T - 1\}$$

这就是指数族分布

$\begin{cases} \eta = \begin{pmatrix} \lambda_0 \\ \lambda \end{pmatrix} \to \text{自然参数} \\ \phi(x) = \begin{pmatrix} 1 \\ f(x) \end{pmatrix} \to \text{sufficient statistic} \\ A(\eta) = [illegible] \to \text{partition function} \end{cases}$

$\therefore p(x)$ 是满足指数族分布

作用：不知数据集不关心是什么分布时，用指数族分布 (Gauss Distribution)

与7.5的区别：约束条件不同

	最大熵约束条件
均匀分布	仅限制取值范围(如 $x \in [a,b]$)，无其它统计量约束
指数族分布	存在矩约束(如已知 $E[x] = \mu$，或更一般 $E[f_i(x)] = \mu_i$)

"在信息约束下，做最无偏的概率推断"

"有约束但信息不全，且需'保持中立，不添假设'"时用最大熵

eg. 已知随机变量均值/方差，但不知分布类型，正态分布是最大熵时。

---

8. Probabilistic Graphical Models

概率图
- Representation 表示
  - 有向图 Bayesian Network
  - 高斯图(连续)：Gaussian BN / Gaussian MN
  - 无向图 Markov Network
- Inference 推断
  - 精确推断
  - 近似推断
    - 确定性近似（变分推断）
    - 随机近似（MCMC）
- Learning 学习
  - 参数学习
    - 完备数据：有向 / 无向
    - 隐变量 EM
  - 结构学习

8.1 PGM Background

高维随机变量 $P(x_1, x_2, \cdots, x_p)$

- 边缘概率 $P(x_i)$
- 条件概率 $P(x_j | x_i)$

Sum Rule: $P(x_1) = \int P(x_1, x_2)\, dx_2$

Product Rule: $P(x_1, x_2) = P(x_1)\cdot P(x_2|x_1) = P(x_2)\cdot P(x_1|x_2)$

eg. $P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)\cdot P(x_3|x_1, x_2)$

Chain Rule: $P(x_1, x_2, \cdots, x_p) = \prod_{i=1}^{p} P(x_i | x_1, x_2, \cdots, x_{i-1})$

Bayesian Rule: $P(x_2|x_1) = \frac{P(x_1, x_2)}{P(x_1)} = \frac{P(x_1, x_2)}{\int P(x_1, x_2)dx_2} = \frac{P(x_2)\cdot P(x_1|x_2)}{\int P(x_2)\cdot P(x_1|x_2)dx_2}$

困境：维度高，计算复杂，$P(x_1, x_2, \cdots, x_p)$ 计算太大 $x_1, x_2, \cdots, x_n$

↓ (Naive Bayes) $P(x|y) = \prod_{i=1}^{p} P(x_i|y)$

简化 → 相互独立 $P(x_1, x_2, \cdots, x_p) = \prod_{i=1}^{p} P(x_i)$

Markov Property → $x_j \perp x_{i>j+1} \mid x_j, j<i$ P25 (HMM 有次 Markov 假设)

给定现在，将来过去相互独立

条件独立性 → $x_A \perp x_B \mid x_C$

x_A, x_B, x_C 是集合，且不相交

8.2. Bayesian Network - condition independence

$P(x_1, x_2, \cdots, x_p) = P(x_1) \cdot \prod_{i=2}^{p} P(x_i | x_{1:i-1})$ → chain Rule

条件独立性：$x_A \perp x_C | x_B$

因子分解 factorization：$P(x_1, x_2, \cdots, x_p) = \prod_{i=1}^{p} P(x_i | x_{pa(i)})$ （有向图）

$x_{pa(i)}$ 是 x_i 的父集合

关于给定证明中的关系时，使用 chain Rule 以及因子分解

拓扑排序

①

tail → head

在a处，连接2个tail 称为(tail to tail)

因子分解

$P(a,b,c) = P(a) P(b|a) P(c|a)$

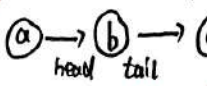
$P(a,b,c) = P(a) \cdot P(b|a) \cdot P(c|a,b)$ chain Rule

$\Rightarrow P(c|a) = P(c|a,b)$

$\Rightarrow c \perp b | a$

必须给定a，若a被观测，则路径被阻塞，即 b, c 独立

②

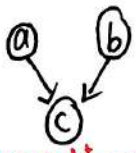


=> head to tail

$a \perp c | b$ 若b被观测，则路径被阻塞

独立

③



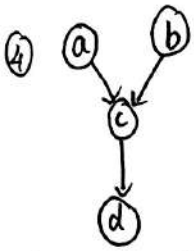
=> head to head

默认情况下，$a \perp b$，若c被观测路径是阻塞的，若c被观测，则路径是通的

$P(a,b,c) = P(a)P(b) \cdot P(c|a,b)$ 因子分解 $\Rightarrow P(c|b) = P(c|a)$

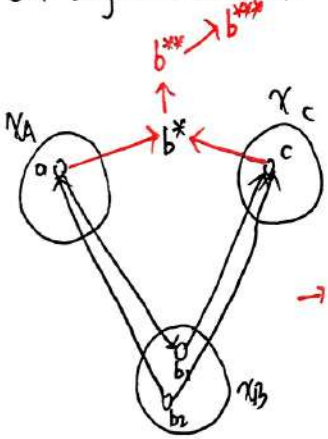
$P(a,b,c) = P(a) P(c|a) P(c|a,b)$ chain Rule $\Rightarrow a \perp b$

④



若d被观测，则路径连通

8.3. Bayesian Network - D-Seperation



→ 全局 Markov Property

根据 8.2 ①②③④ 作一个更加 generious 的 result

$P(x_i | x_{-i}) = \frac{P(x_i, x_{-i})}{P(x_{-i})} = \frac{P(x)}{\int_{x_i} P(x) dx_i} = \frac{\prod_{j=1}^{p} p(x_j | x_{pa(j)})}{\int_{x_i} \prod_{j=1}^{p} p(x_j | x_{pa(j)}) dx_i}$

x_i 与其他所有元素间的关系

Δ / Δ ← factorization 8.2中

与 x_i 相关 Δ

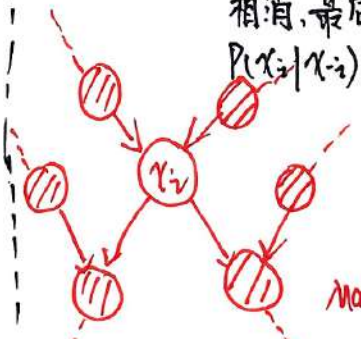
与 x_i 无关 Δ

→ 在分母中，与 x_i 无关的 Δ 可以被提出积分号，而由于分母也可以提出一个 Δ，二者相消，最后就变为与 Δ 相关的函数，即

$P(x_i | x_{-i}) = f(\Delta)$

$\propto p(x_i | x_{pa(i)}) \cdot p(x_{child(i)} | x_i, x_{parent(child(i))})$

双亲



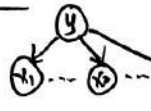
Markov Blanket

即在 Markov Network 中，一个元素 x_i 与所有网络元素之间的关系，等于所有与它相关的，即 Markov Blanket.

8.4. Bayesian Network - example

Bayesian Network

- 单一：Naive Bayes → $P(x|y) = \prod_{i=1}^{n} P(x_i | y=1)$



Naive Bayes

$x_1 \perp x_2 | y$

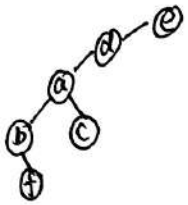
- 混合：GMM
- 时间：Markov chain / Gaussian Process（无限维）
  - → 动态模型
    - HMM（离散）
    - LDS（Kalman Filter）（连续，线性）Gaussian
    - Particle Filter（非高斯，非线性）
- 连续：Gaussian Bayesian Network

P26.

8.5. Markov Random Field - Representation - 条件独立性

① $x_A \perp x_C \mid x_B$
Global Markov

② Local Markov



$a \perp \{\text{全集} - a - \text{邻居}\} \mid \{\text{邻居}\}$
$a \perp \{e, f\} \mid \{b, c, d\}$

pairwise
③ 成对 Markov
$x_i \perp x_j \mid x_{-i-j} \quad (i \neq j)$

条件独立性体现在三个方面：①全局 ②局部 ③成对
且 ①<=>②<=>③ 相互等价

factorization:

图：
团：一个关于节点的集合，集合中的节点之间相互都是连通的
最大团：无法加入新节点，加入后就不是团

概率分布：归一化因子
$P(x) = \frac{1}{Z} \prod_{i=1}^{K} \phi(x_{ci})$
$Z = \sum_{x} \prod_{i=1}^{K} \phi(x_{ci}) = \sum_{x_1 x_2 \cdots x_p} \prod_{i=1}^{K} \phi(x_{ci})$

$C_1, C_2, \cdots C_K$ 表示一个团
x_C 表示所对应团的一个集合
定义在团上
$\phi(x)$ 是一个函数，加上归一化因子后才是概率

需证明二者等价
才能说明这是一个概率图的 factorization

8.6. Markov Random Field - Representation - Factorization

D-separation
通过：条件独立性
① Global Markov Property: $x_A \perp x_C \mid x_B$ 如果 A, B, C, sep(A, C | B), 那么 $x_A \perp x_C \mid x_B$
② Local Markov Property: $x_i \perp x_{-i-nb(i)} \mid x_{nb(i)}$, $nb(i)$: ~~neighborhood~~ Neighbor of node i
③ Pairwise Markov property: $x_i \perp x_j \mid x_{-i-j}$
①<=>②<=>③

factorization: 无向图，有向图以后。
$P(x) = \frac{1}{Z} \prod_{i=1}^{K} \psi(x_{ci})$,
C_i: 最大团
x_{ci}: 最大团随机变量集合
$Z = \sum_{x} \prod_{i=1}^{K} \psi(x_{ci}) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_p} \prod_{i=1}^{K} \psi(x_{ci})$
$\psi(x_{ci})$: 势函数，必须为正

如果一个概率分布，可以写成基于团上的因子分解，根据 Hammesley-Clifford 可以证明它是马尔可夫随机场：①<=>②<=>③ <=> 因子分解（基于最大团）

$\psi(x_{ci}) = \exp\{-E(x_{ci})\} > 0$
Energy Function

$P(x)$ 称为 Gibbs Distribution / Boltzman Distribution

$P(x) = \frac{1}{Z} \prod_{i=1}^{K} \psi(x_{ci}) = \frac{1}{Z} \prod_{i=1}^{K} \exp\{-E(x_{ci})\}$
$= \frac{1}{Z} \exp\{-\sum_{i=1}^{K} E(x_{ci})\}$

形式上与指数族分布相似 <=> 指数族分布 $\rightarrow P(x) = h(x) \cdot \exp\{\eta^T \phi(x) - A(\eta)\}$
$= \frac{1}{Z(\eta)} h(x) \exp\{\eta^T \phi(x)\}$

最大熵原理 => 指数族分布（Gibbs 分布）
Markov Random Field <=> Gibbs Distribution

8.7 Inference - Introduction

Inference：求概率：$P(x) = P(x_1, x_2, \cdots, x_p)$
边缘概率：$P(x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \sum_{x_p} P(x)$ 除 x_i 对其余变量的所有取值求和
条件概率：$P(x_A \mid x_B)$ $x = x_A \cup x_B$
MAP Inference：$P(z \mid x)$ $\hat{z} = \arg\max_z P(z \mid x) \propto \arg\max P(z, x)$

Inference
- 精确推断
  - variable elimination (VE)
  - Belief Propagation (BP) 别称 Sum-Product Algorithm (树状结构)
  - Junction Tree Algorithm (普通图结构)
- 近似推断
  - Loop Belief Propagation (有环图)
  - Monte Carlo Inference: Importance Sampling, MCMC
  - variational Inference:

8.8. Inference - Variable Elimination (乘法分配律)

Tasks: Inference (给定 $P(\cdot) = P(Y_1, Y_2, \cdots, Y_p)$)

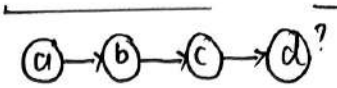
边缘概率: $P(Y_i) = \sum_{Y_1, Y_2, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_p} P(Y_1, Y_2, \cdots, Y_p)$

条件概率: $P(Y_A | Y_B)$

MAP $\hat{Y}_A = \arg\max_{Y_A} P(Y_A | Y_B) = \arg\max_{Y_A} P(Y_A, Y_B)$

$P(x) = \prod_{x_c} \phi_c(Y_c)$

缺点: ① 重复计算 ② ordering → NP-hard



(假 a,b,c,d 均是离散二值变量 r.v., $a,b,c,d \in \{0,1\}$)

$$P(d) = \sum_{a,b,c} P(a,b,c,d) = \sum_{a,b,c} P(a) \cdot P(b|a) \cdot P(c|b) \cdot P(d|c)$$

$= P(a=0) \cdot P(b=0|a=0) P(c=0|b=0) P(d=0|c=0)$
$+ P(a=1) P(b=0|a=1) P(c=0|b=0) P(d=0|c=0)$
$+ \cdots$
$+ P(a=1) P(b=1|a=1) P(c=1|b=1) P(d=0|c=1)$
$= 8$ 项乘积

$$= \sum_{b,c} P(c|b) \cdot P(d|c) \cdot \underbrace{\sum_a P(a) \cdot P(b|a)}_{\phi_a(b)} \quad \left(\sum_a P(a,b) = P(b)\right)$$

$$= \sum_c P(d|c) \cdot \underbrace{\sum_b P(c|b) \cdot \phi_a(b)}_{\phi_b(c)}$$

$$= \phi_c(d)$$

(乘法对加法分配律)

8.9. Inference - Belief Propagation

Variable Elimination: (乘法分配律)

$P(a,b,c,d,e) = P(a) \cdot P(b|a) \cdot P(c|b) \cdot P(d|c) \cdot P(e|d)$

Forword Algorithm

$$P(e) = \sum_{a,b,c,d} P(a,b,c,d,e) = \sum_d P(e|d) \cdot \sum_c P(d|c) \cdot \sum_b P(c|b) \cdot \sum_a P(b|a) P(a)$$

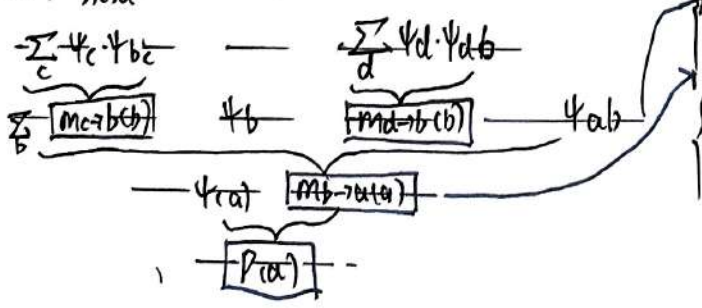
$m_{d\to e}(e) \to m_{c\to d}(d) \to m_{b\to c}(c) \to m_{a\to b}(b)$

$$P(c) = \sum_{a,b,d,e} P(a,b,d,e) = \left(\sum_b P(c|b) \cdot \sum_a P(b|a) P(a)\right) \cdot \left(\sum_d P(d|c) \sum_e P(e|d)\right)$$

Forword-Backword Algorithm

为解决重复计算, 引入 Belief Propagation:

$$P(a) = \sum_{b,c,d} P(a,b,c,d)$$ 结合①和图①, 从 c,d 往回



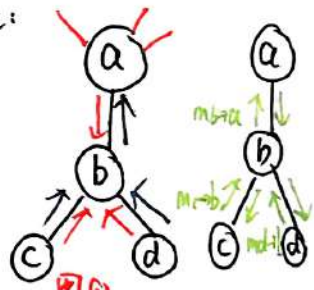
$$\begin{cases} m_{b\to a}(Y_a) = \sum_{Y_b} \psi_{ab} \psi_b \cdot m_{c\to b}(Y_b) \cdot m_{d\to b}(Y_b) \\ P(a) = \psi_a m_{b\to a}(Y_a) \\ m_{j\to i}(Y_i) = \sum_{Y_j} \psi_{ij} \psi_j \prod_{k \in NB(j)\setminus i} m_{k\to j}(Y_j) \\ P(Y_i) = \psi_i \prod_{k \in NB(i)} m_{k\to i}(Y_i) \end{cases}$$

b 的除 a 外的 Neighbor

NB: neighbor

如图中 a 可能有多个 neighbor

BP: Chain-Tree: 有向-无向



图①

求 $P(a)$

(4个点, 3边)

$$P(a,b,c,d) = \frac{1}{Z} \psi_a(a) \cdot \psi_b(b) \cdot \psi_c(c) \cdot \psi_d(d) \cdot \psi_{a,b}(a,b) \cdot \psi_{bc}(b,c) \cdot \psi_{bd}(b,d) \cdots ①$$

$$P(a) = \sum_{b,c,d} P(a,b,c,d), \quad P(b) = \sum_{a,c,d} P(a,b,c,d)$$

蓝线 红线

...

若把绿线全求出, $P(a)P(b)P(c)P(d)$ 自然也就出来

☆不需直接求边缘概率 ($P(a), P(b), P(c), P(d)$)

只需求 $m_{i\to j}$

Belief propagation

8.09 Inference - Belief propagation

$m_{j\to i} = \sum_j \psi_{ij}\,\psi_j \prod_{\text{邻居}} m_{k\to j}$ (边 点 路径)

$m_{b\to a} = \sum_b \psi_{ab}\,\psi_b\, m_{c\to b}\, m_{d\to b}$ (self, children)

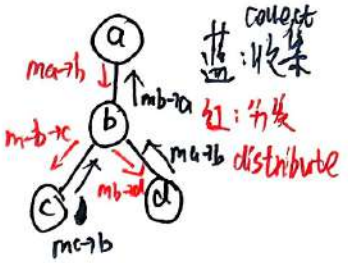
belief(b) = $\psi_b \cdot$ children

$m_{b\to a} = \sum_b \psi_{ab} \cdot \text{belief}(b)$

法II BP = VE + Caching

↳ (直接求 $m_{ij} \Rightarrow P(x_i)$)

图的遍历



BP (Sequential Implementation) 序

① Get root, assume a is root

② collect Message → for x_i in NB(root): collect Msg(x_i) 递归

③ distribute Message → for x_j in NB(root): distribute(x_j)

可得 m_{ij} for all $i,j \in V$

从而 $P(x_k)$, $k \in V$

法III Parellel Implementation 分布式

随机找一点 → 每一个点 collection Message

可以并行运行 → 迅速得到收敛

信息传到某个节点后，向所有邻居扩散，以此类推地传播

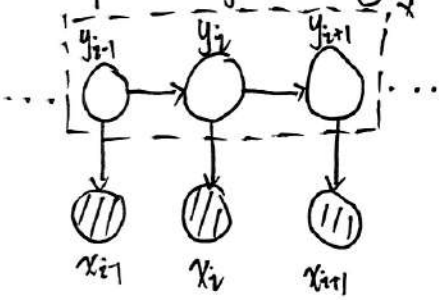
8.10 Inference - Max Product Algorithm

Graph = $\{V, E\}$

(I) 边缘概率：$E = \{e_1, e_2, \cdots, e_k\}$, $P(E)$ (likelihood)

(II) 条件概率：$P(Y|E)$ (posterior) $\quad x = (Y, Z)$ $P(Y|E) = \sum_Z P(x|E)$

(III) MAP (Decoding)：$\hat{x} = \arg\max_x P(x|E)$ $\quad \hat{Y} = \arg\max_Y P(Y|E) = \arg\max_Y \sum_Z P(x|E)$



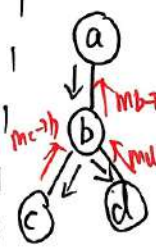
Decoding：$\hat{Y} = \arg\max_Y P(Y|x)$

↳ ① viterbi Algorithm：动态规划问题

Belief propagation (Tree) | Sum-product

$m_{j\to i}(x_i) = \sum_{x_j} \psi_j(x_j)\cdot\psi_{ij}(x_i,x_j) \prod_{k\in NB(j)-i} m_{k\to j}(x_j)$

$P(x_i) = \psi_i(x_i) \cdot \prod_{k\in NB(i)} m_{k\to i}(x_i)$



此处：$m_{b\to a}$, $m_{c\to b}$, $m_{d\to b}$ 是经过了前一个节点的最大概率

Max-product：

① BP 改进；② Viterbi 的推广

$(\hat{a},\hat{b},\hat{c},\hat{d}) = \arg\max_{a,b,c,d} P(x_a, x_b, x_c, x_d | E)$ | MAX-product

如 $m_{b\to a}$，即经过 c, d 后，从 b 到 a 的最大概率，即 c, b, d 联合概率最大

① 改进：

$m_{j\to i} = \max_{x_j} \psi_j \cdot \psi_{ij} \cdot \prod_{k\in NB(i)-i} m_{k\to i}$

改为：

$(x_a^*, x_b^*, x_c^*, x_d^*) = \arg\max_{x_a,x_b,x_c,x_d} P(x_a,x_b,x_c,x_d|E)$

$x_d^* = \arg\max_{x_d} \psi_d \cdot \psi_{bd}$

c 最优路径 $x_c^* = \arg\max_{x_c} \psi_c \cdot \psi_{bc}$

回溯：$(x_a^*, x_b^*, x_c^*, x_d^*)$

$x_b^* = \arg\max_{x_b} \psi_b \cdot \psi_{ab} \cdot m_{c\to b} \cdot m_{d\to b}$

$x_a^* = \arg\max_{x_a} \psi_a \cdot m_{b\to a}$

$m_{c\to b} = \max_{x_c} \psi_c \cdot \psi_{bc}$ 与 c 相关的所有概率 关于 c 的函数

$m_{d\to b} = \max_{x_d} \psi_d \cdot \psi_{bd}$ 关于 d 的函数

不向图无法用概率修容

$m_{b\to a} = \max_{x_b} \psi_b \cdot \psi_{ab} \cdot m_{c\to b} \cdot m_{d\to b}$ b 的函数

从底向上可求出 $\max P(x_a, x_b, x_c, x_d)$

从上向底回溯可求出 $x_a^*, x_b^*, x_c^*, x_d^*$

$\max P(\text{root}) = \max_{x_a} \psi_a \cdot m_{b\to a}$ a 的函数

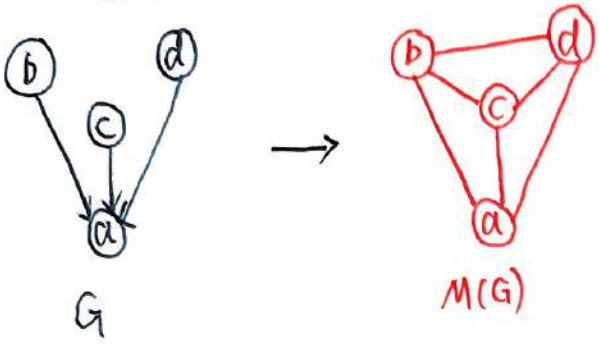
8.1. Supplement - ~~Factor~~ Moral Graph

原因：想把有向图转为无向图来研究。

有向图：head-to-head（V结构）

作法：① $\forall x_i \in G$，将 Parent(x_i) 两两连接

② 将G中的有向弧替换成无向边

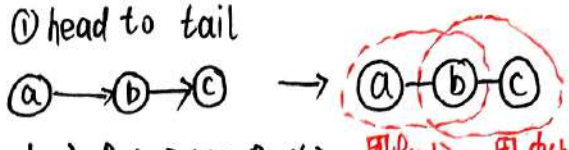


G　　M(G)

Sep(A,B|C) <=> D-sep(A,B|C)

无向图中　　有向图

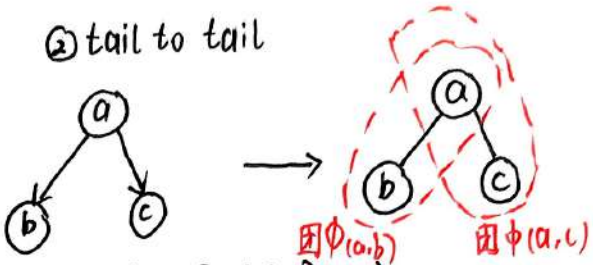
① head to tail



$P(a,b,c) = P(a)\cdot P(b|a)\cdot P(c|b)$

$\psi(a,b)$　$\psi(b,c)$

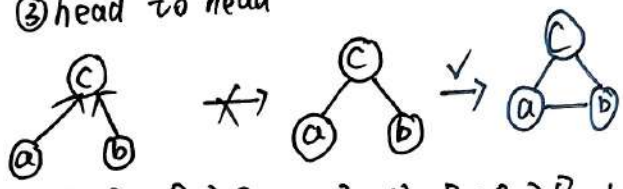
② tail to tail



$P(a,b,c) = P(a)\cdot P(b|a)\cdot P(c|a)$

$\psi(a,b)$　$\psi(a,c)$

③ head to head



$P(a,b,c) = P(a)P(b)\cdot P(c|a)\cdot P(c|b) = \underbrace{P(a)P(b)P(c|a,b)}_{\psi(a,b,c)}$

按①②做法，(a,b,c)显然不为一个团

8.2 Supplement - Factor Graph

有向图：$P(x) = \prod_i^n P(x_i|x_{pa(i)})$

无向图：$P(x) = \frac{1}{Z}\prod_{i=1}^{K}\phi_{ci}(x_{ci})$，最大团集合

道德图：有向图（树）→ 无向图（树）（引入环）（Tree-Like Graph）

① BP只能对树操作，因子图可以去环

② 简便

因式分解本身对应一个特殊的因子图

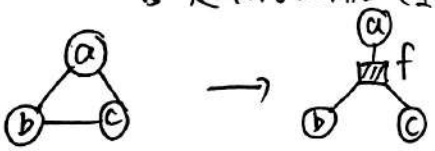
因子图：看作是对因式分解的进一步分解

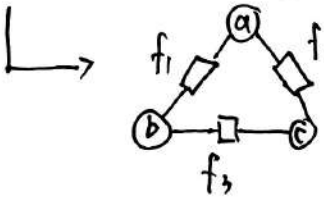
因子图：$x = x_1, \cdots, x_p$

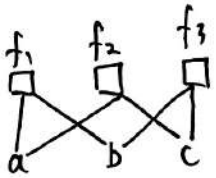
$P(x) = \prod_S f_S(x_S)$

S：图的节点子集

x_S：是x的随机变量子集







---

9. EM Algorithm

9.1 EM Introduction

可以得到解析解的例子：MLE：$P(x|\theta)$

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

EM：$\theta^{(t+1)} = \arg\max_\theta \int_z \log P(x,z|\theta)\cdot P(z|x,\theta^{(t)})dz$

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

M-步就是使其最大

收敛性

$\log P(x|\theta^{(t)}) \le \log P(x|\theta^{(t+1)})$ …①

EM：最大化对数似然，保证下一步都比上一步大，能达到极值

证①式：$P(x,z|\theta) = P(z|x,\theta)\cdot P(x|\theta)$

$P(x|\theta) \overset{\checkmark}{=} \frac{P(x,z|\theta)}{P(z|x,\theta)}$

$\log P(x|\theta) = \log\left(\frac{P(x,z|\theta)}{P(z|x,\theta)}\right)$

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

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

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

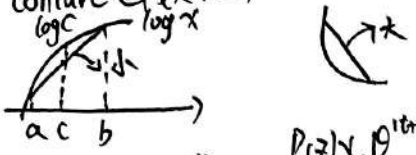
$= \log P(x|\theta)\cdot \int_z P(z|x,\theta^{(t)})dz$

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

对任意随机变量 X（满足 $X>0$ 且 $E[X]$ 存在），若 f 是凹函数，则：$f(E[X]) \geq E[f(X)]$

或由 Jensen 不等式：

concave 函数（凹）　convex（凸）



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

$$\boxed{E[\log x] \leq \log E(x)}, \quad f(E[x]) \geq E[f(x)] \quad f(x)=\log x$$

$$\leq \log \int_z P(z|x,\theta^{(t+1)}) dz = \log 1 = 0$$

$\therefore H(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \leq 0.$

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

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

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

由 P30 $\theta^{(t+1)} = \arg\max$ 得出

$\Leftarrow Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta,\theta^{(t)})$

$\Leftarrow \boxed{Q(\theta^{(t+1)},\theta^{(t)}) \geq Q(\theta^{(t)},\theta^{(t)})}$ → 自然成立 ①

θ是变量对符号没有影响，自然成立 也有替换

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

只证：$H(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \leq 0$

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

$= \int_z P(z|x,\theta^{(t)}) \cdot \log \frac{P(z|x,\theta^{(t+1)})}{P(z|x,\theta^{(t)})} dz$ ↓ kullback-Leiber divergence

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

≤ 0

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

$$\boxed{-H(\theta^{(t+1)},\theta^{(t)}) \geq -H(\theta^{(t)},\theta^{(t)})} \quad ②$$

注意变量替换：$\log P(x|\theta) \to \log P(x|\theta^{(t)}) = Q(\theta^{(t)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \leq \log P(x|\theta^{(t+1)})$

由①② → $Q(\theta^{(t)},\theta^{(t)}) - H(\theta^{(t)},\theta^{(t)}) \leq Q(\theta^{(t+1)},\theta^{(t)}) - H(\theta^{(t+1)},\theta^{(t)})$

故 $\boxed{\log P(x|\theta^{(t)}) \leq \log P(x|\theta^{(t+1)})}$ 收敛

上述推导并不严密：P30中 $\log P(x|\theta)$ 中是取 θ，因为我们证①式时写的是 $P(x,z|\theta) = P(z|x,\theta) \cdot P(x|\theta)$，若把 θ 改为 $\theta^{(t)}$ 则有 $\log P(x|\theta^{(t)}) = \log P(x,z|\theta^{(t)}) - \log P(z|x,\theta^{(t)})$ 之后证明关于变量替换部分就明显了，我用绿笔画出

9.2 ELBO + KL Divergence

MLE：$\theta_{MLE} = \log P(x|\theta)$

x：observed data　$x = \{x_1,\dots,x_N\}$　x_i 独立同分布，非独立时候好拆开大义表示有样本　$\log P(x) = \log \prod_{i=1}^N P(x_i) = \sum_{i=1}^N \log P(x_i)$

z：unobserved data (latent variable)

(x,z)：complete data

θ：parameter

E-step：$P(z|x,\theta^{(t)}) \to E_{z|x,\theta^{(t)}}[\log P(x,z|\theta)]$

M-step：$\theta^{(t+1)} = \arg\max_\theta E_{z|x,\theta^{(t)}}[\log P(x,z|\theta)]$

收敛性：$\log P(x|\theta^{(t)}) \leq \log P(x|\theta^{(t+1)})$

下面证明上M：

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

$= \log \frac{P(x,z|\theta)}{q(z)} - \log \frac{P(z|x,\theta)}{q(z)}, \quad q(z) \neq 0$

同时对左右关于 $q(z)$ 作期望

左边 $= \int_z q(z) \log P(x|\theta) dz = \log P(x|\theta) \cdot \int_z q(z) dz = \log P(x|\theta)$

右边 $= \underbrace{\int_z q(z) \cdot \log \frac{P(x,z|\theta)}{q(z)} dz}_{ELBO = \text{evidence lower bound}} - \underbrace{\int_z q(z) \log \frac{P(z|x,\theta)}{q(z)} dz}_{KL(q(z)\|P(z|x,\theta))}$

对数似然的下界

P31

故有：$\log P(x|\theta) = ELBO + KL(q\|P)$ → Posterior

且 $KL(q\|P) \geq 0 \Leftrightarrow$ 取"=" $q = P$ 即 $q(z) = P(z|x,\theta)$

则：$\boxed{\log P(x|\theta) \geq ELBO}$ → 下界

让 ELBO 尽量大，则 $\log P(x|\theta)$ 也会大。

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

$= \arg\max_\theta \int_z q(z) \log \frac{P(x,z|\theta)}{q(z)} dz$

上一次迭代取 $q(z) = P(z|x,\theta)$ 时，$\log P(x|\theta) = ELBO$，（θ 是自变量，parameter，令为 $\theta^{(t)}$）

$= \arg\max_\theta \int_z P(z|x,\theta^{(t)}) [\log P(x,z|\theta) - \log P(z|x,\theta^{(t)})] dz$

$\theta^{(t)}$ constant，无关，与 θ

$\therefore \hat{\theta} = \arg\max_\theta \int_z \log P(x,z|\theta) \cdot P(z|x,\theta^{(t)}) dz$

$\hat{\theta} \to \theta^{(t+1)}$　Variable　constant

9.3. ELBO + Jensen's Inequality

$\log P(x|\theta) = ELBO + KL(q\|p)$ 9.2.

$q = q(z)$

$p = p(z|x,\theta)$

$$\log P(x|\theta) = \log \int_z P(x,z|\theta)\,dz$$
$$= \log \int_z \frac{P(x,z|\theta)}{q(z)} \cdot q(z)\,dz$$
$$= \log E_{q(z)}\left[\frac{P(x,z|\theta)}{q(z)}\right]$$
$$\geq E_{q(z)}\left[\log \frac{P(x,z|\theta)}{q(z)}\right] \rightarrow ELBO$$

"=" 成立 $\Leftrightarrow \frac{P(x,z|\theta)}{q(z)} = C$

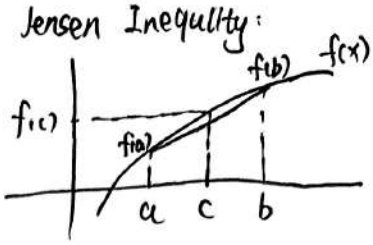
$$q(z) = \frac{1}{C} P(x,z|\theta)$$
$$1 = \int_z q(z)\,dz = \int_z \frac{1}{C} P(x,z|\theta)\,dz$$
$$= \frac{1}{C}\int_z P(x,z|\theta)\,dz$$
$$1 = \frac{1}{C} P(x|\theta)$$
$$C = P(x|\theta)$$

则 $q(z) = \frac{1}{P(x|\theta)} \cdot P(x,z|\theta) = P(z|x,\theta)$ → Posterior

我们随机引入的 $q(z)$ 就是一个 posterior

则有 $\log P(x|\theta^{(t)}) \geq E_{q(z)}\left[\log \frac{P(x,z|\theta)}{P(z|x,\theta^{(t)})}\right] \Rightarrow \hat{\theta} = \arg\max_\theta E_{q(z)}\left[\log \frac{P(x,z|\theta)}{P(z|x,\theta^{(t)})}\right]$ 与 9.2 同理

Jensen Inequality:



$f(x)$ = concave funtion

$t \in [0,1]$

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

$f(c) = f[ta + (1-t)b] \geq tf(a) + (1-t)f(b)$ → Jensen

当 $t = \frac{1}{2}$ 时.

$$f\left[\underbrace{\frac{a}{2}}_{E} + \underbrace{\frac{b}{2}}_{E}\right] \geq \underbrace{\frac{f(a)}{2}}_{E} + \underbrace{\frac{f(b)}{2}}_{E}$$

$f(E) \geq E[f]$

9.4 EM Review 再回首

① 从狭义 EM → 广义 EM

② 狭义 EM 是广义 EM 的一个特例

③ EM 变种

概率生成模型 $P(x|\theta)$

EM 就是求这个 $\hat{\theta}$

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

$x \sim P(x)$ $\quad X = \{x_i\}_{i=1}^N$

9.5. Generalized EM

$\log P(x|\theta) = \underbrace{ELBO}_{\mathcal{L}(q,\theta)} + KL(q\|p)$

$\log P(x|\theta) \geq \mathcal{L}(q,\theta)$

$q = p$ 时 → 有时 p 不一定能算出，q 不一定能取到 p

$q \to p$，KL 越小，ELBO 越大

广义：则有先算 $\hat{q}$，再求 $\hat{\theta}$

固定 θ，$\hat{q} = \arg\min_q KL(q\|p) = \arg\max_q \mathcal{L}(q,\theta)$ → E-step

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

广义 EM：

$$\begin{cases} \text{E-step: } q^{(t)} = \arg\max_q \mathcal{L}(q,\theta^{(t)}) \\ \text{M-step: } \theta^{(t+1)} = \arg\max_\theta \mathcal{L}(q^{(t+1)},\theta) \end{cases}$$

$\arg\max \log P(x|\theta) = \arg\max \mathcal{L}(q,\theta)$ 不易计算，引入辅助函数求极值，故 $\arg\max \mathcal{L}(q,\theta)$

$KL(q\|p) = \log P(x|\theta) - \mathcal{L}(q,\theta)$

$\arg\min KL = \arg\max_q \mathcal{L}(q,\theta)$

解释 $\arg\max = ?\arg\min$ ($\log P(x|\theta)$ 常值)

坐标上升法 (SMO)

梯度上升法

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

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

简化 θ

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

$H[q]$

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

故广义 EM，就是处理当 q 无法取到 p 的情况，$q = q(z)$，$p = p(z|x,\theta)$，而 q 越接近 p，则 KL 就越小，ELBO 越大，故先固定 θ 求最优 $\hat{q}$，在对 $\hat{q}$ 求 $\hat{\theta}$，并且着力到 $\mathcal{L}(q,\theta)$ 上进行计算

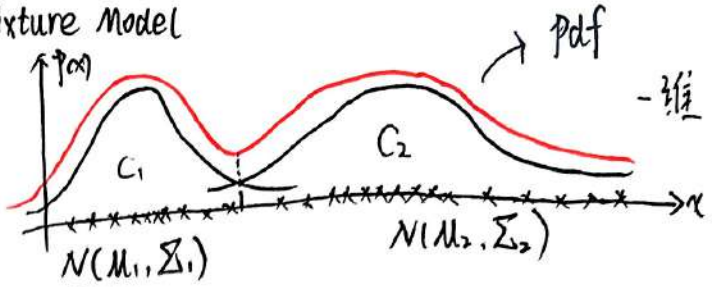
10. Gaussian Mixture Model

10.1 Introduction

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

从几何角度来看：加权平均（多个高斯分布叠加而成）

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



从混合模型角度看：（生成模型）

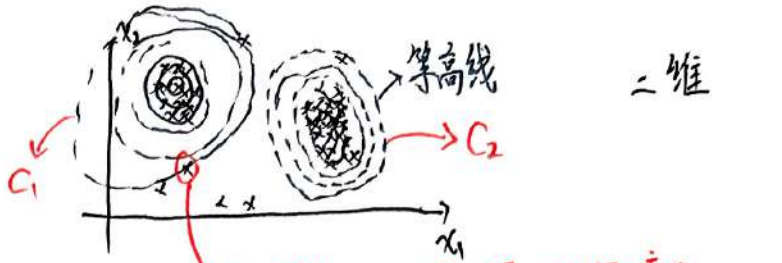
N个样本 $x_1, x_2, \cdots, x_N$

x: observed variable

z: latent variable
↳对应的样本x是属于哪个高斯分布
↳离散的随机变量

z	C_1	C_2	$\cdots$	C_K
	1	2	$\cdots$	K
$p(z)$	p_1	p_2	$\cdots$	p_K

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



① 假设有一个骰子
② 不均匀的骰子
③ 投一个面，假设为p_k
④ 在第K个高斯分布中随机采样，采到即为样本。

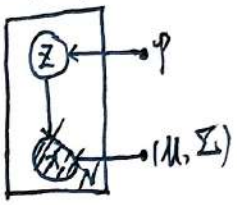
x即$\in C_1$，又$\in C_2$，但$\in C_1$概率大

$x \wedge z$ △

$p=(p_1,p_2,\cdots,p_K)$

$x|z=C_k \sim N(x|\mu_k,\Sigma_k)$

概率图



10.2 MLE

由上节对混合模型：

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

Mixture model，z下x的概率是Gauss distribution

p_k △ 概率值

$p(x,z)=p(z)\cdot p(x|z)$

$=p_z\cdot N(x|\mu_z,\Sigma_z)$

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

X: observed data → $X=(x_1,x_2,\cdots,x_N)$

(X,Z): complete data → $(x_1,z_1),(x_2,z_2),\cdots,(x_N,z_N)$

θ: parameter → $\theta=\{p_1,p_2,\cdots,p_K,\mu_1,\mu_2,\cdots,\mu_K,\Sigma_1,\Sigma_2,\cdots,\Sigma_K\}$

$$\hat{\theta}_{MLE}=\arg\max_\theta \log p(x)=\arg\max_\theta \log\prod_{i=1}^{N} p(x_i)=\arg\max_\theta \sum_{i=1}^{N}\log p(x_i)=\arg\max_\theta \sum_{i=1}^{N}\log\sum_{k=1}^{K} p_k\cdot N(x_i|\mu_k,\Sigma_k)$$

十分复杂，无法得到解析解
连加号，十分难算

∴ 直接用MLE求解GMM无法得出解析解

10.3 EM - E-Step

$$\theta^{(t+1)}=\arg\max_\theta \underbrace{E_{z|x,\theta^{(t)}}[\log p(x,z|\theta)]}_{Q(\theta,\theta^{(t)})}$$

variable / constant

$$Q(\theta,\theta^{(t)})=\int_z \log p(x,z|\theta)\cdot p(z|x,\theta^{(t)})\,dz$$

$$=\sum_z \log\prod_{i=1}^{N} p(x_i,z_i|\theta)\cdot\prod_{i=1}^{N} p(z_i|x_i,\theta^{(t)})$$

$$=\sum_{z_1,z_2,\cdots,z_N}\sum_{i=1}^{N}\log p(x_i,z_i|\theta)\cdot\prod_{i=1}^{N} p(z_i|x_i,\theta^{(t)})$$

$$=\sum_{z_1,z_2,\cdots,z_N}\left[\log p(x_1,z_1|\theta)+\log p(x_2,z_2|\theta)+\cdots+\log p(x_N,z_N|\theta)\right]\prod_{i=1}^{N} p(z_i|x_i,\theta^{(t)})$$

对于：$\sum_{z_1,z_2,\cdots,z_N}\log p(x_1,z_1|\theta)\cdot\prod_{i=1}^{N} p(z_i|x_i,\theta^{(t)})$ → 第一项

$$=\sum_{z_1,z_2,\cdots,z_N}\left(\log p(x_1,z_1|\theta)\cdot p(z_1|x_1,\theta^{(t)})\cdot\prod_{i=2}^{N} p(z_i|x_i,\theta^{(t)})\right)$$

$$=\sum_{z_1}\log p(x_1,z_1|\theta)\cdot p(z_1|x_1,\theta^{(t)})\cdot\sum_{z_2,\cdots,z_N}\prod_{i=2}^{N} p(z_i|x_i,\theta^{(t)})$$

$$=\sum_{z_1}\log p(x_1,z_1|\theta)\cdot p(z_1|x_1,\theta^{(t)})\cdot\sum_{z_2} p(z_2|x_2,\theta^{(t)})\cdot\sum_{z_3} p(z_3|x_3,\theta^{(t)})\cdots\sum_{z_N} p(z_N|x_N,\theta^{(t)})$$

（$\sum_{z_2} p(z_2|x_2,\theta^{(t)})=1$，$\sum_{z_3} p(z_3|x_3,\theta^{(t)})=1$，…，$\sum_{z_N} p(z_N|x_N,\theta^{(t)})=1$，$\theta^{(t)}$常数）

展开：$Q(\theta,\theta^{(t)})=\sum_{z_1}\log p(x_1,z_1|\theta)\cdot p(z_1|x_1,\theta^{(t)})+\cdots+\sum_{z_N}\log p(x_N,z_N|\theta)\cdot p(z_N|x_N,\theta^{(t)})$

$$=\sum_{i=1}^{N}\sum_{z_i}\log p(x_i,z_i|\theta)\cdot p(z_i|x_i,\theta^{(t)})$$

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

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

10.2 $P(x,z)=p_z\cdot N(x|\mu_z,\Sigma_z)$

$P(z|x)=\dfrac{p_{k_z}\cdot N(x|\mu_z,\Sigma_z)}{\sum_{k=1}^{K}p_k\cdot N(x|\mu_k,\Sigma_k)}$

代入. $\Rightarrow Q(\theta,\theta^{(t)})=\sum_{i=1}^{N}\sum_{z_i}\log\left[p_{z_i}\cdot N(\gamma_i|\mu_{z_i},\Sigma_{z_i})\right]\cdot\dfrac{p_{z_i}\cdot N(\gamma_i|\mu_{z_i}^{(t)},\Sigma_{z_i}^{(t)})}{\sum_{k=1}^{K}p_k^{(t)}\cdot N(\gamma_i|\mu_k^{(t)},\Sigma_k^{(t)})}$

由 θ 是 parameter，θ 与 μ,Σ,p 都有关，故代入上式后 θ 可不写是第 t 个参数，带上标

$\rightarrow P(z_i|\gamma_i,\theta^{(t)})$

10.4 EM-M-Step

$$Q(\theta,\theta^{(t)})=\sum_{z_i}\sum_{i=1}^{N}\log\left[p_{z_i}\cdot N(\gamma_i|\mu_{z_i},\Sigma_{z_i})\right]\cdot P(z_i|\gamma_i,\theta^{(t)})$$

$$=\sum_{k=1}^{K}\sum_{i=1}^{N}\log\left[p_k N(\gamma_i|\mu_k,\Sigma_k)\right]\cdot P(z_i=c_k|\gamma_i,\theta^{(t)})$$

$$=\sum_{k=1}^{K}\sum_{i=1}^{N}\left[\log p_k+\log N(\gamma_i|\mu_k,\Sigma_k)\right]\cdot P(z_i=c_k|\gamma_i,\theta^{(t)})$$

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

求 $p_k^{(t+1)}$：$p_k^{(t+1)}=\arg\max_{p_k}\sum_{k=1}^{K}\sum_{i=1}^{N}\log p_k\cdot P(z_i=c_k|\gamma_i,\theta^{(t)})$，s.t. $\sum_{k=1}^{K}p_k=1$ $\rightarrow$ $\begin{cases}\max\limits_{p_k}\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{N}\log p_k\cdot P(z_i=c_k|\gamma_i,\theta^{(t)})\\ \text{s.t.}\ \sum\limits_{k=1}^{K}p_k=1\end{cases}$

$p^{(t+1)}=(p_1^{(t+1)},\cdots,p_K^{(t+1)})^T$ $\quad\mathcal{L}(p,\lambda)=\sum_{k=1}^{K}\sum_{i=1}^{N}\log p_k\cdot P(z_i=c_k|\gamma_i,\theta^{(t)})+\lambda\left(\sum_{k=1}^{K}p_k-1\right)$.

（$P(z_i=c_k|\gamma_i,\theta^{(t)})$ 与 p_k 无关）

对其中一项 c_k 求偏导 $\dfrac{\partial\mathcal{L}}{\partial p_k}=\sum_{i=1}^{N}\dfrac{1}{p_k}\cdot P(z_i=c_k|\gamma_i,\theta^{(t)})+\lambda\overset{\Delta}{=}0$

$\Rightarrow\sum_{i=1}^{N}P(z_i=c_k|\gamma_i,\theta^{(t)})+p_k\lambda=0$. $\rightarrow$ 通项（k 项）

有 $\sum_{i=1}^{N}P(z_i=c_1|\gamma_i,\theta^{(t)})+p_{c_1}\lambda=\cdots=\sum_{i=1}^{N}P(z_i=c_k|\gamma_k,\theta^{(t)})+p_{c_k}\lambda=0$

$\Rightarrow\sum_{i=1}^{N}\sum_{k=1}^{K}P(z_i=c_k|\gamma_i,\theta^{(t)})+\sum_{k=1}^{K}p_k\lambda=0$

$\because\sum_{k=1}^{K}p_k=1$，$\sum_{k=1}^{K}P(z_i=c_k|\gamma_i,\theta^{(t)})=1$ $\rightarrow$ 相对积分

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

$\therefore\quad\lambda=-N$

$\therefore\quad p_k^{(t+1)}=\dfrac{1}{N}\sum_{i=1}^{N}P(z_i=c_k|\gamma_i,\theta^{(t)})=\dfrac{1}{N}\sum_{i=1}^{N}\gamma_{ik}$

（$P(z_i=c_k|\gamma_i,\theta^{(t)})$ 总管对行 $\rightarrow$ 责任；这 $z_i=c_k$ 是 $p^{(t+1)}=(p_1^{(t+1)},\cdots,p_K^{(t+1)})$ 时用）

$\therefore p^{(t+1)}=(p_1^{(t+1)},\cdots,p_K^{(t+1)})$，

若记 $P(z_i=c_k|\gamma_i,\theta^{(t)})=\gamma_{ik}$ 可节省空间（γ_{ik} 责任）

② $\mu_k^{(t+1)}=\dfrac{\sum_{i=1}^{N}P(z_i=c_k|\gamma_i,\theta^{(t)})\cdot\gamma_i}{\sum_{i=1}^{N}P(z_i=c_k|\gamma_i,\theta^{(t)})}=\dfrac{\sum_{i=1}^{N}\gamma_{ik}\cdot\gamma_i}{\sum_{i=1}^{N}\gamma_{ik}}$

③ $\Sigma_k^{(t+1)}=\dfrac{\sum_{i=1}^{N}P(z_i=c_k|\gamma_i,\theta^{(t)})\cdot(\gamma_i-\mu_k)(\gamma_i-\mu_k)^T}{\sum_{i=1}^{N}P(z_i=c_k|\gamma_i,\theta^{(t)})}=\dfrac{\sum_{i=1}^{N}\gamma_{ik}\cdot(\gamma_i-\mu_k)(\gamma_i-\mu_k)^T}{\sum_{i=1}^{N}\gamma_{ik}}$

11. Variational Inference

11.1 Introduction

频率角度 → 优化问题：

回归：$D=\{(x_i,y_i)\}_{i=1}^N$, $x_i\in\mathbb{R}^p$, $y_i\in\mathbb{R}$

$f(w)=w^Tx$ (Model), loss function：$L(w)=\sum_{i=1}^N\|w^Tx_i-y_i\|^2$, $\hat{w}=\arg\min L(w)$ } 无约束

解法：① 解析解：$\frac{\partial L(w)}{\partial w}\triangleq 0\Rightarrow w^*=(X^TX)^{-1}X^TY$（$X$: $N\times p$）

② 数值解：GD → Gradient Descent

SVM（分类）：① $f(w)=\text{sign}(w^Tx+b)$；② loss function：$\begin{cases}\min\frac{1}{2}w^Tw\\ \text{s.t. } y_i(w^Tx_i+b)\geq 1\ (i=1,\cdots,N)\end{cases}$ ⇒有约束，convex优化

③ QP；Lagrange Dual

EM：$\hat{\theta}=\arg\max\log P(X|\theta)\Rightarrow\theta^{(t+1)}=\arg\max_\theta\int_Z\log P(X,Z|\theta)\cdot P(Z|X,\theta^{(t)})\,dZ$

看到X数据，假设有隐变量Z，给定$P(Z)$，求$P(Z|X)$

Inference { 精确推断；近似推断 { 确定性近似 → VI；随机近似 → MCMC, MH, Gibbs } } 近似 → 计算

贝叶斯角度 → 积分问题：

$P(\theta|X)=\frac{P(X|\theta)\cdot P(\theta)}{P(X)}$（似然 $P(X|\theta)$；先验 $P(\theta)$；后验 $P(\theta|X)$；$P(X)=\int_\theta P(X|\theta)\cdot P(\theta)\,d\theta$）

Bayes Inference；Bayes 决策：X - N个样本，$\hat{x}$新的样本，求$P(\hat{x}|X)$

$P(\hat{x}|X)=\int_\theta P(\hat{x},\theta|X)\,d\theta=\int_\theta P(\hat{x}|\theta)\cdot\underbrace{P(\theta|X)}_{\text{后验}}\,d\theta=E_{\theta|X}[P(\hat{x}|\theta)]$

11.2. Formula Deduction

X: observed data

Z: latent variable + parameter

(X,Z): complete data

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

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

左边 $=\int_Z\log P(X)\cdot Q(Z)\,dZ=\log P(X)$ = constant

右边 $=\underbrace{\int_Z^{Q(Z)}\log\frac{P(X,Z)}{Q(Z)}\,dZ}_{\text{evidence lower bound ELBO}}-\underbrace{\int_Z^{Q(Z)}\log\frac{P(Z|X)}{Q(Z)}\,dZ}_{KL(Q\|P)}$

$=\mathcal{L}(Q)+KL(Q\|P)$

$\mathcal{L}(Q)$ 变为 ↓ ≥0

$KL\to 0$ ↑ ↓

$\hat{Q}(Z)=\arg\max_{Q(Z)}\mathcal{L}(Q)\Rightarrow\hat{Q}(Z)\simeq P(Z|X)$

$Q(Z)=\prod_{i=1}^M Q_i(Z_i)$ → mean theory → $Q_j(Z_j)$

$\mathcal{L}(Q)=\int_Z Q(Z)\log P(X,Z)\,dZ-\int_Z Q(Z)\log Q(Z)\,dZ$ （这里的Z是大Z，代表M个隐变量）

① $=\int_Z\prod_{i=1}^M Q_i(Z_i)\log P(X,Z)\,dZ_1dZ_2\cdots dZ_M$

$=\int_{Z_j}Q_j(Z_j)\,dZ_j\left(\int_{Z_1\cdots Z_M(\neq Z_j)}\prod_{i\neq j}^M Q_i(Z_i)\log P(X,Z)\cdot dZ_1\cdots dZ_{j-1}dZ_{j+1}\cdots dZ_M\right)$

$\int\log P(X,Z)\cdot\prod_{i\neq j}^M Q_i(Z_i)\,dZ_i=\log\hat{P}(X,Z_j)$

$=\int_{Z_j}Q_j(Z_j)\cdot\underbrace{E_{\prod_{i\neq j}^M Q_i(Z_i)}[\log P(X,Z)]}_{(\text{关于}\prod_{i\neq j}^M Q_i(Z_i))}\cdot dZ_j$

② $=\int_Z Q(Z)\log Q(Z)\,dZ=\int_Z\prod_{i=1}^M Q_i(Z_i)\cdot\log\prod_{i=1}^M Q_i(Z_i)\,dZ$

$=\int_Z\prod_{i=1}^M Q_i(Z_i)\cdot\sum_{i=1}^M\log Q_i(Z_i)\,dZ=\int_Z\prod_{i=1}^M Q_i(Z_i)\cdot[\log Q_1(Z_1)+\cdots+\log Q_M(Z_M)]\,dZ$

其中一项：$\int_Z\prod_{i=1}^M Q_i(Z_i)\cdot\log Q_1(Z_1)\,dZ=\int_Z Q_1(Z_1)\cdots Q_M(Z_M)\cdot\log Q_1(Z_1)\,dZ$

$=\int_{Z_1Z_2\cdots Z_M}Q_1Q_2\cdots Q_M\log Q_1\,dZ_1dZ_2\cdots dZ_M$

P35 $=\int_{Z_1}Q_1\log Q_1\,dZ_1\cdot\underbrace{\int_{Z_2}Q_2\,dZ_2}_{1}\cdots\underbrace{\int_{Z_M}Q_M\,dZ_M}_{1}=\int_{Z_1}Q_1\log Q_1\,dZ_1$

$\because$ 所有项 $\sum_{i=1}^{M}\int_{Z_i} q_i(Z_i)\log q_i(Z_i)\,dZ_i = \int_{Z_j} q_j(Z_j)\log q_j(Z_j)\,dZ_j + C$ 其它项十第j项

① - ② $= \int_{Z_j} q_j(Z_j)\log\frac{\hat{P}(X,Z_j)}{q_j(Z_j)}\,dZ_j + C = -KL(q_j \| \hat{P}(X,Z_j)) \le 0$，当 $\frac{\hat{P}(X,Z_j)}{q_j(Z_j)}$ 为常数的话，取等号

$\therefore q_j(Z_j) = \hat{P}(X,Z_j)$ 通过 argmax ELBO 作出，它的核心就是 $P(X,Z)$ → 结合 PGM. $\log q_j(Z_j) = \log \hat{P}(X,Z_j) = E_{\prod_{i\neq j} q_i(Z_i)}[\log P(X,Z)]$

11.3 Review (再回首)

X: observed variable → $X = \{x_i\}_{i=1}^N = \{x^{(i)}\}_{i=1}^N$

Z: latent variable → $Z = \{z_i\}_{i=1}^N = \{z^{(i)}\}_{i=1}^N$

(X,Z): complete data

θ: model parameter

EM中: $\log P(X|\theta) = \underbrace{ELBO}_{\mathcal{L}(q,\theta)} + \underbrace{KL(q\|p)}_{\ge 0} \ge \mathcal{L}(q,\theta)$

$E_{q(Z)}\left[\log\frac{P_\theta(x^{(i)},Z)}{q(Z)}\right]$ 而在VI中，是求 q 的分布，故 θ 可假设已求出，也可假设为~~缺省~~参数

$ELBO = E_{q(Z)}\left[\log\frac{P(X,Z|\theta)}{q(Z)}\right] = E_{q(Z)}[\log P_\theta(X,Z)] + H[q(Z)]$ 故VI: $\log P_\theta(X) = \underbrace{ELBO}_{\mathcal{L}(q)} + \underbrace{KL(q\|p)}_{\ge 0} \ge \mathcal{L}(q)$

$KL(q\|p) = \int q(Z)\cdot\log\frac{q(Z)}{P_\theta(Z|X)}\,dZ$

目标函数: $\hat{q} = \arg\min_q KL(q\|p) = \arg\max_q \mathcal{L}(q)$

有用 $x^{(i)}$ 样本来表述

$\log P_\theta(X) = \log\prod_{i=1}^N P_\theta(x^{(i)}) = \sum_{i=1}^N \log P_\theta(x^{(i)})$

VI (mean field) → classical VI

Assumption: $q(Z) = \prod_{i=1}^M q_i(Z_i)$

$\log q_j(Z_j) = E_{\prod_{i\neq j} q_i(Z_i)}[\log P_\theta(X^{(i)},Z)] + \text{const}$

$= \int_{q_1}\int_{q_2}\cdots\int_{q_{j-1}}\int_{q_{j+1}}\cdots\int_{q_M} q_1 q_2\cdots q_{j-1}q_{j+1}\cdots q_M[\log P_\theta(x^{(i)},Z)]\cdot dq_1 dq_2\cdots dq_{j-1}dq_{j+1}\cdots dq_M$

则关于目标函数有: $\hat{q}_1(Z_1) = \int_{q_2}\cdots\int_{q_M} q_2\cdots q_M[\log P_\theta(x^{(i)},Z)]dq_2\cdots dq_M$

$\hat{q}_2(Z_2) = \int_{\hat{q}_1}\int_{q_3}\cdots\int_{q_M}\hat{q}_1 q_3\cdots q_M[\log P_\theta(x^{(i)},Z)]d\hat{q}_1 dq_3\cdots dq_M$

$\vdots$

$\hat{q}_M(Z_M) = \int_{\hat{q}_1}\int_{\hat{q}_2}\cdots\int_{\hat{q}_{M-1}}\hat{q}_1\hat{q}_2\cdots\hat{q}_{M-1}[\log P_\theta(x^{(i)},Z)]d\hat{q}_1 d\hat{q}_2\cdots d\hat{q}_{M-1}$

坐标上升法: Coordinate Ascend

classical VI:

问题: ①假设太多虽: (mean field) 每个变量相互独立 ~~独立~~; ②intractable

11.4 SGVI

目标函数: $\hat{q} = \arg\min_q KL(q\|p) = \arg\max_q \mathcal{L}(q)$

假设 $q(Z) \to q_\phi(Z)$，$q(Z)$ 的参数为 ϕ，通常是一个指数族分布

$\log P_\theta(x^{(i)}) = \underbrace{ELBO}_{\mathcal{L}(\phi)} + \underbrace{KL(q\|p)}_{\ge 0} \ge \mathcal{L}(\phi)$

$\hat{\phi} = \arg\max_\phi \mathcal{L}(\phi)$

$ELBO = E_{q_\phi(Z)}[\log P_\theta(x^{(i)},Z) - \log q_\phi(Z)] = \mathcal{L}(\phi)$

则关于 $\hat{\phi} = \arg\max_\phi \mathcal{L}(\phi)$

$\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi E_{q_\phi}[\log P_\theta(x^{(i)},Z) - \log q_\phi]$ ···①

$= \nabla_\phi \int q_\phi[\log P_\theta(x^{(i)},Z) - \log q_\phi]dZ$

$= \int \nabla_\phi q_\phi \cdot [\log p_\theta(x^{(i)}, z) - \log q_\phi] dz + \int q_\phi \cdot \nabla_\phi [\log p_\theta(x^{(i)}, z) - \log q_\phi] dz$

$= ① + ②$

$= \int \nabla_\phi \cdot q_\phi \cdot [\log p_\theta(x^{(i)}, z) - \log q_\phi] dz$

由：$\nabla_\phi q_\phi = \nabla_\phi \log q_\phi \cdot q_\phi$

得：$\int q_\phi \cdot \nabla_\phi \log q_\phi [\log p_\theta(x^{(i)}, z) - \log q_\phi] dz$

$= E_{q_\phi} [\nabla_\phi \log q_\phi [\log p_\theta(x^{(i)}, z) - \log q_\phi]$

② $= -\int q_\phi \cdot \nabla_\phi \log q_\phi \, dz$

$= -\int q_\phi \cdot \frac{1}{q_\phi} \nabla_\phi \cdot q_\phi dz$

$= -\int \nabla_\phi q_\phi \, dz$

$= -\nabla_\phi \int q_\phi dz$

$= 0$

variance Reduction

需要非常多样本

high variance

方差不稳定，并且在采样过程中得到的 $\nabla_\phi \mathcal{L}(\phi)$ 是近似的，之后使用 SG 仍是近似，二者误差叠加

故：$\nabla_\phi \mathcal{L}(\phi) = E_{q_\phi} [\nabla_\phi \log q_\phi [\log p_\theta(x^{(i)}, z) - \log q_\phi]$ ，$q_\phi = q_\phi(z)$ 可用 MC 采样，

$z^{(l)} \sim q_\phi(z)$，$l = 1, 2, \cdots, L$

$\approx \frac{1}{L} \sum_{l=1}^{L} \nabla_\phi \log q_\phi(z^{(l)}) [\log p_\theta(x^{(i)}, z^{(l)}) - \log q_\phi(z^{(l)})]$ → 缺点：

降低 Variance：Reparameterization Trick（目的：让 $q_\phi \to p_\varepsilon$，与 ∇_ϕ 无关，易求，简化改进）

重参数化技巧

$\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi E_{q_\phi} [\log p_\theta(x^{(i)}, z) - \log q_\phi]$，P36 ①式

假定 $z = g_\phi(\varepsilon, x^{(i)})$，$\varepsilon \sim p(\varepsilon)$ 是一个确定的分布

$z \sim q_\phi(z|x^{(i)})$

$\varepsilon \sim p(\varepsilon)$ → 自己给定的一个分布

可认为 g_ϕ 是一个变换 → z 的随机性转换给 ε

含 $\varepsilon \sim p(\varepsilon)$ 作用：在一开始我们设 $q(z|x^{(i)})$ 的参数为 ϕ（通常为指数族），但我们并不知道是哪一种分布，故有在 ∇_ϕ 时 E_{q_ϕ} 与 ϕ 相关，在迭计算中存在不稳定，high variance 问题，所以我们引入 Reparameterization Trick，使其 $\varepsilon \sim p(\varepsilon)$ 并有一个确定变换 $z = g_\phi(\varepsilon, x^{(i)})$，且 $\varepsilon \sim p(\varepsilon)$ 是一个确定的分布 此时 $E_{p(\varepsilon)}$ 与 ∇_ϕ 就无关了，化简

由 ~~$\int q_\phi(z) dz = \int p(\varepsilon) d\varepsilon$~~

~~$q_\phi(z) dz = p(\varepsilon) d\varepsilon$~~

$\int q_\phi(z|x^{(i)}) dz = 1$

$\int p(\varepsilon) d\varepsilon = 1$

有 $|q_\phi(z|x^{(i)}) dz| = |p(\varepsilon) d\varepsilon|$

由 $\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi E_{q_\phi} [\log p_\theta(x^{(i)}, z) - \log q_\phi]$

$= \nabla_\phi \int [\log p_\theta(x^{(i)}, z) - \log q_\phi] \cdot q_\phi dz$

$= \nabla_\phi \int [\log p_\theta(x^{(i)}, z) - \log q_\phi] \cdot p(\varepsilon) d\varepsilon$

$= \nabla_\phi E_{p(\varepsilon)} [\log p_\theta(x^{(i)}, z) - \log q_\phi]$ 分布 $p(\varepsilon)$ 与 ϕ 无关

$= E_{p(\varepsilon)} \{\nabla_\phi [\log p_\theta(x^{(i)}, z) - \log q_\phi]\}$

$= E_{p(\varepsilon)} \{\nabla_z [\log p_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})] \cdot \nabla_\phi z\}$ 全连式法则

$= E_{p(\varepsilon)} \cdot \{\nabla_z [\log p_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})] \cdot \nabla_\phi \cdot g_\phi(\varepsilon, x^{(i)})\}$

$\frac{\partial f}{\partial \phi} = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial \phi}$ $z = g(\phi)$

此时采样：

$\varepsilon^{(l)} \sim p(\varepsilon)$，$l = 1, 2, \cdots, L$

$\nabla_\phi \mathcal{L}(\phi) \approx \frac{1}{L} \sum_{l=1}^{L} \{\nabla_z [\log p_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})] \cdot \nabla_\phi g_\phi(\varepsilon^{(l)}, x^{(i)})\}$

SGVI：

$\phi^{(t+1)} \leftarrow \phi^{(t)} + \lambda^{(t)} \cdot \nabla_\phi \mathcal{L}(\phi)$ ← 每一步都要算 P37

12. MCMC (Markov chain & Monte Carlo)

12.1 Sampling Method Introduction

Inference { 精确推断 ; 近似推断 { 确定性 → VI ; 随机 → MCMC }

Monte Carlo Method: 基于采样的随机近似方法

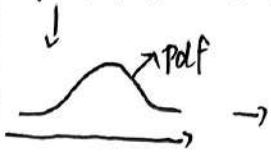
$P(z|x)$ ← observed data ⟶ $E_{z|x}[f(z)] = \int P(z|x)\cdot f(z)\,dz \approx \frac{1}{N}\sum_{i=1}^{N} f(z_i)$

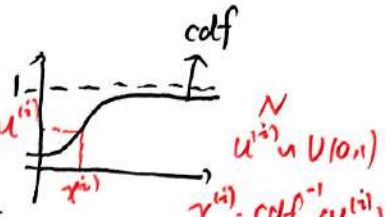
latent variable

$z^{(1)}, z^{(2)}, \cdots, z^{(N)}$ 来自 $\sim P(z|x)$

① 根据概率分布采样:

$p(z)$: PDF ⟶ 求出 CDF





$u \sim U(0,1)$ 伪随机数

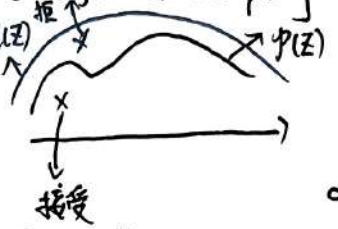
$u^{(i)} \sim U(0,1)$

$z^{(i)} = cdf^{-1}(u^{(i)})$

$z^{(1)}, z^{(2)}, \cdots z^{(N)}$ 样本

(由pdf求cdf很困难)

② Rejection Sampling



$q(z)$: proposal distribution

$\forall z_i, Mq(z_i) \geq p(z_i)$

借助 $U(0,1)$

i: $z^{(i)} \sim q(z)$

ii: $u \sim U(0,1)$

if $u \leq \alpha$, 接收 $z^{(i)}$

else 拒绝

α: 接收率, $\alpha = \frac{p(z^{(i)})}{Mq(z^{(i)})}$, $0 \leq \alpha \leq 1$

$p(z)$ 与 $q(z)$ 相似程度

③ Importance Sampling (期望采样)

$E_{p(z)}[f(z)] = \int p(z)f(z)\,dz = \int \frac{p(z)}{q(z)}\cdot q(z)\cdot f(z)\,dz$

$= \int f(z)\frac{p(z)}{q(z)}\cdot q(z)\,dz \approx \frac{1}{N}\sum_{i=1}^{N} f(z_i)\frac{p(z_i)}{q(z_i)}$

$\frac{p(z_i)}{q(z_i)}$: weight

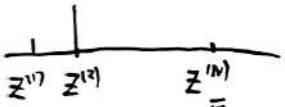
$z_i \sim q(z)$, $i=1,\cdots,N$ 样本从 $q(z)$ 中进

取决于 $p(z)$ 与 $q(z)$ 相似程度

③的变种: Sambling - Importance - Resambling (简要思想)

Importance sampling 的结果: 权重 (weight)

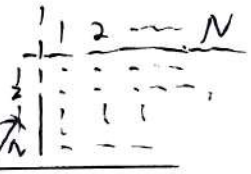
① 采样



② 2次采样, 从 $z^{(1)}, z^{(2)}, \cdots z^{(N)}$ 采样

以①中 weight 为概率直接去采样

采样效率不均匀, 很可能有很高, 很低的



第 i 个状态到第 j 个状态的概率

每行到列概率和为1, $\sum_{j=1}^{\infty} P_{ij} = 1$

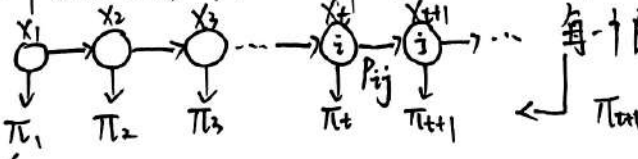
12.2 Markov Chain

Markov chain: 时间和状态都是离散的 $\{X_t\}$, P-转移矩阵 $[P_{ij}]$

有次(一阶) Markov chain: $P(X_{t+1}=x \mid X_1, X_2, \cdots, X_t) = P(X_{t+1}=x \mid X_t)$ 无效性, 未来只依赖于当前, 与过去无关

只依赖于前一个状态

$P_{ij} = P_{ij} = P(X_{t+1}=j \mid X_t = i)$



每一个随机变量都有属于它的概率分布 π_i

平稳分布: $\{\pi(k)\}$

$\pi_{t+1}(x^*) = \int \pi_t(x)\cdot P(x \mapsto x^*)\,dx$

若 $\exists \pi = [\pi^{(1)}, \pi^{(2)}, \cdots, \pi^{(N)}, \cdots]$, $\sum_{i=1}^{\infty}\pi(i) = 1$

使得 $\pi(x^*) = \int \pi(x)\cdot P(x \mapsto x^*)\,dx$ 成立

$x \xrightarrow{P(x\mapsto x^*)} x^*$

t: $\pi_t(x)$ $\quad$ $t+1$: $\pi_{t+1}(x^*)$

则 $\{\pi(k)\}$ 是 $\{X_i\}$ 的平稳分布, 即每一相邻的 π_i 的分布是相同的

细致平衡原理

Detailed Balance: $\pi(x)\cdot P(x \mapsto x^*) = \pi(x^*)\cdot P(x^* \mapsto x)$

这: $P(x \mapsto x^*) = P(x^*|x)$

Detailed Balance $\overset{(DB)}{\rightleftharpoons}$ 平稳分布 (没有基于 0-0-…-0-… 这个结构的 Markov Chain 特性, 而是更 general 的)

证明: $\int \pi(x)\cdot P(x \mapsto x^*)\,dx = \int \pi(x^*)\cdot P(x^* \mapsto x)\,dx$ (DB)

$= \pi(x^*)\int P(x^* \mapsto x)\,dx$

$= \pi(x^*)$

$\int P(x^* \mapsto x)\,dx = 1$ 若为离散 $\sum_{j=1}^{\infty} P_{ij} = 1$ 转移矩阵的一行

12.3 MH Algorithm 接受率

$p(z)\cdot Q(z\to z^*)\cdot\alpha(z,z^*)=p(z^*)\cdot Q(z^*\to z)\cdot\alpha(z^*,z)$ ①，$Q=[Q_{ij}]$ proposal matrix 从 $P=[P_{ij}]$ 中随机选取的

$p(i)\cdot P(i\to j)=p(j)\cdot P(j\to i)$

$\alpha(z,z^*)=\min\left(1,\frac{p(z^*)\cdot Q(z^*\to z)}{p(z)\cdot Q(z\to z^*)}\right)$ → $Q(z|z^*)$，→ $Q(z^*|z)$

①式中，若不加入 $\alpha(z,z^*)$ 时不满足 Detailed Balance 因为 Q 是随机选取的，此时 $p(z)$ 并不满足 DB.

$p(z)\cdot Q(z\mapsto z^*)\neq p(z^*)\cdot Q(z^*\mapsto z)$

Q_{ij} 是任取的

从前往后接后，加上 α 后由后往前接后

证明加入 $\alpha(z,z^*)$ 后满足 DB：$p(z)\cdot P(z\mapsto z^*)$

$\boxed{p(z)\cdot Q(z\mapsto z^*)\cdot\alpha(z,z^*)}=p(z)\cdot Q(z\mapsto z^*)\cdot\min\left(1,\frac{p(z^*)\cdot Q(z^*\to z)}{p(z)\cdot Q(z\to z^*)}\right)$

$=\min\left(p(z)\cdot Q(z\to z^*),p(z^*)\cdot Q(z^*\to z)\right)=\min p(z^*)\cdot Q(z^*\to z)\cdot\min\left(1,\frac{p(z)\cdot Q(z\mapsto z^*)}{p(z^*)\cdot Q(z^*\mapsto z)}\right)$

$=\boxed{p(z^*)\cdot Q(z^*\to z)\cdot\alpha(z^*,z)}=p(z^*)\cdot P(z^*\mapsto z)$

$\alpha(z^*,z)$

Metropolis-Hastings:

$u\sim U(0,1)$

$z^*\sim Q(z|z^{(i-1)})$ ↑ 从 Q 中采集

$\alpha=\min\left(1,\frac{p(z^*)\cdot Q(z^*\mapsto z)}{p(z)\cdot Q(z\mapsto z^*)}\right)$

if $u\le\alpha$, $z^{(i)}=z^*$

else $z^{(i)}=z^{(i-1)}$

$p(z)=\frac{\hat p(z)}{Z_p}$ 归一化因子，常数化因子，即一般不求

$\hat p(z^*),\hat p(z)$

$z^{(i)}\sim\hat p(z)\sim p(z)$

与 Rejection Sampling 区别：$u>\alpha$ 时不拒绝，而是重复上一个样本，这样不会造成样本量减少

$p(z)\to E_{p(z)}[f(z)]=\int_z p(z)f(z)dz\simeq\frac{1}{N}\sum_{i=1}^{N}f(z^{(i)})$

代入这 N 个样本点

α：接收率的作用，把 $Q\to P$

①从 Q 中采集；②利用判断是否接收

转执行 N 次，得到 $z^{(1)},z^{(2)},\dots,z^{(N)}$ 个样本点

12.4 Gibbs Algorithm (special MH Sampling) 一维一维的采

$p(z)=p(z_1,\dots,z_M)$ 高维采样的数据是高维

$z_i\sim p(z_i|z_{-i})$，固定其它样本

$p(z)=p(z_1,z_2,z_3)$

$i=z_1^{(0)},z_2^{(0)},z_3^{(0)}$

$j:t+1$:

$z_1^{(t+1)}\sim p(z_1|z_2^{(t)},z_3^{(t)})$

$z_2^{(t+1)}\sim p(z_2|z_1^{(t+1)},z_3^{(t)})$

$z_3^{(t+1)}\sim p(z_3|z_1^{(t+1)},z_2^{(t+1)})$

$z_{-1}=z_{-1}^*\Rightarrow z_{-i}=z_{-i}^*$

$\alpha=\min\left(1,\frac{p(z^*)\cdot Q(z^*\to z)}{p(z)\cdot Q(z\to z^*)}\right)$

计算 α：

$\frac{p(z^*)\cdot Q(z^*\to z)}{p(z)\cdot Q(z\to z^*)}=\frac{p(z_i^*|z_{-i}^*)\cdot p(z_{-i}^*)\cdot p(z|z_{-i}^*)}{p(z_i|z_{-i})\cdot p(z_{-i})\cdot p(z^*|z)}=\frac{p(z_i^*|z_{-i}^*)\cdot p(z_{-i}^*)\cdot p(z_i|z_{-i}^*)}{p(z_i|z_{-i})\cdot p(z_{-i})\cdot p(z_i^*|z_{-i})}$

$p(z^*)=p(z_i^*,z_{-i}^*)$

由 $z_{-i}=z_{-i}^*$，有：$\frac{p(z_i^*|z_{-i}^*)p(z_{-i}^*)\cdot p(z_i|z_{-i}^*)}{p(z_i|z_{-i}^*)\cdot p(z_{-i}^*)\cdot p(z_i^*|z_{-i}^*)}=1$

代入 MH 中的 $\alpha=\min\left(1,\frac{p(z^*)\cdot Q(z^*\mapsto z)}{p(z)\cdot Q(z\mapsto z^*)}\right)$ 中后，$\alpha=1$，效率更高

12.5 Review

采样的动机：① 采样本身就是常见的任务

② 求和求积分，eg. $E_{p(x)}[f(x)] = \int p(x) f(x)\,dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})$

什么是好的样本：① 样本趋于高概率区域

② 样本之间相互独立

采样是困难的：(the curse of high dimensionality)

① partation function is intractable　eg. $p(x) = \frac{1}{Z}\hat{p}(x) = \frac{\hat{p}(x)}{\int \hat{p}(x)dx}$，$x \in \mathbb{R}^p$（$Z$：归一化常数；$\int \hat{p}(x)dx$：intractable）

② high dimension，eg. $x \in \{1,2,\cdots,k\}^p$

状态空间 k^p

x	1	2	…	k^p
p	p_1	p_2	…	p_{k^p}

→ 考虑样本趋于高概率，把每一个 $p_1 \to p_{k^p}$ 比较不现实

因此有：Rejection Sampling，Important Sampling，$p(x)$ 采样有困难，通过 $q(x)$ 绕过 $p(x)$

MCMC：{ MH；Gibbs（MH特例）

12.6 Stationary Distribution（平稳分布）

Rejection Sampling / Importance Sampling：① $q(x)$ 和 $P(x)$ 接近　② $q(x)$ 简单

核心：构造一条马尔可夫链，使其收敛的 $q(x)$ 与我们的 $P(x)$ 相近，通过采样的方式，在 m 之后的样本取出一些样本（n 个），则就实现了绕过 $P(x)$ 直接从 $P(x)$ 中采样

重要的点是转移矩阵 Q 的构造

状态空间：$\{1,2,\cdots,k\}$

$$q^{(t+1)}(x=j) = \sum_{i=1}^{k} q^{(t)}(x=i)\cdot Q_{ij}$$

$q^{(1)}$：

x	1	2	…	k
$q_x^{(1)}$	q_1^1	q_1^2	…	q_1^k

→ 若有 k 个状态

状态转移矩阵（矩阵）：Q：

$$Q = \begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1k} \\ Q_{21} & Q_{22} & \cdots & Q_{2k} \\ \vdots & & & \vdots \\ Q_{k1} & Q_{k2} & \cdots & Q_{kk} \end{pmatrix}$$

$\sum_{j=1}^{k} Q_{ij} = 1 \quad (i=1,\cdots,k)$

$\sum_{j=1}^{k} Q_{ji} = 1 \quad (i=1,\cdots,k)$

} 每行每列概率都为 1

随机矩阵

令 $q^{(t+1)} = (q^{(t+1)}(x=1),\ q^{(t+1)}(x=2)\ \cdots\ q^{(t+1)}(x=k))_{1\times k}$

而 $q^{(t+1)}(x=j) = \sum_{i=1}^{k} q^{(t)}(x=i)\cdot Q_{ij}$

$\therefore q^{(t+1)} = \left(\sum_{i=1}^{k} q^{(t)}(x=i)\cdot Q_{i1},\ \sum_{i=1}^{k} q^{(t)}(x=i)\cdot Q_{i2},\ \cdots,\ \sum_{i=1}^{k} q^{(t)}(x=i)\cdot Q_{ik}\right)_{1\times k} = q^{(t)}_{1\times k}\cdot Q_{k\times k}$

（Q_{i1}：all 状态；Q_{i2}：所有状态；Q_{ik}：所有状态；→ 即前一个的所有状态 × Q_{ij} → 到 j）

$\sum_{i=1}^{k} q^{(t)}(x=i)\cdot Q_{i1} = q^{(t)}(x=1)Q_{11} + q^{(t)}(x=2)\cdot Q_{21} + \cdots + q^{(t)}(x=k)\cdot Q_{k1}$

$q^{(t)} = (q^{(t)}_{(x=1)}\ q^{(t)}_{(x=2)}\ \cdots\ q^{(t)}_{(x=k)})_{1\times k}$

$q^{(t+1)} = q^{(t)}\cdot Q = q^{(t-1)}\cdot Q^2 = \cdots = q^{(1)}Q^t$

即：$q^{(t+1)} = q^{(1)}\cdot(A\cdot\Lambda A^{-1})^t = q^{(1)}\cdot A\cdot\Lambda^t A^{-1}$

~~当 $t\to\infty$ 时，$\Lambda^t \to \begin{pmatrix} 0 & & \\ & 0 & \\ & & \ddots \end{pmatrix} = A$~~

随机矩阵特征值的绝对值 ≤ 1

$\therefore Q = A\cdot\Lambda A^{-1}$

$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_k \end{pmatrix}$，$|\lambda_i| \le 1$，for $i=1,\cdots,k$，

只有不妨设 λ_i 为 1

有存在足够多大的 m，s.t. $\Lambda^m = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix}$

$q^{(m+1)} = q^{(1)}\cdot A\cdot\Lambda^m A^{-1}$

$q^{(m+2)} = q^{(m+1)}\cdot A\cdot\Lambda A^{-1} = q^{(1)}A\cdot\Lambda^{m+1}\cdot A^{-1} = q^{(m+1)}$

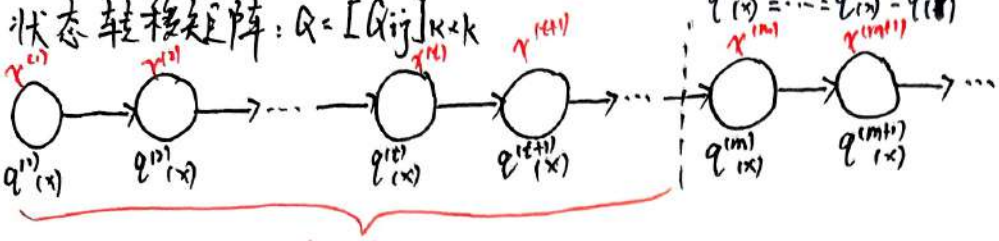
$\Lambda^{m+1} = \Lambda^m\cdot\Lambda = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix}\begin{pmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & \lambda_k \end{pmatrix} = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \end{pmatrix}$

$\therefore q^{(m+2)} = q^{(m+1)}$，当 $t > m$ 时，$q^{(m+1)} = q^{(m+2)} = \cdots = q^{(\infty)}$，都是同一个分布，平稳

12.7 Problem & Thinking

状态空间：{1,2,…,k}

状态转移矩阵：$Q=[Q_{ij}]_{k\times k}$



burn-in

mixing time

MCMC：利用Markov chain 收敛于平稳分布设计Q，使得平稳分布 $q(x)$ 即目标分布 $P(x)$

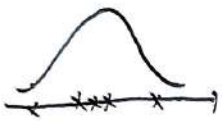
MCMC问题：

①理论只保证收敛性，但无法知道何时收敛

②mixing time 过长 P(x)太复杂，高维及相关性

③样本之间有一定的相关性（每隔100/1000个样本取一个）

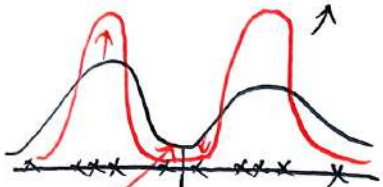
单峰 vs 多峰（1维）



$P(x)=\frac{1}{Z}\exp(-E(x))$

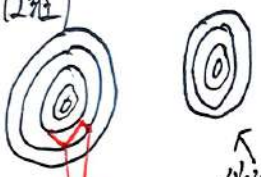
非负 partition 配分 energy function 能量大稳定

陡峭 vs 缓和



即有可能无法到达另一个峰点，无法越过这条低沟，因为算法都是趋向于高概率区域的

（2维）

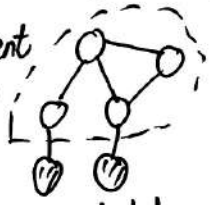


这两点概率相同

做了很长时间还没采样

问题：太复杂

Boltzman Machine



latent data

observed data

RBM（隐变量 latent variable 无关）

h（隐层）

相互独立

latent data.

观测层（V）

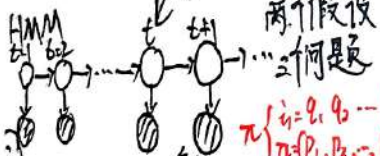
$P(h|v)=P(h_1|v)P(h_2|v)P(h_3|v)$

RBM → Restricted Boltzman machine

---

13. Hidden Markov Model

13.1 Background



一个模型 $\lambda=(\pi,A,B)$

两个假设 三个问题

π $i_1=q_1, q_2 \cdots q_N$ $\pi=(p_1,p_2,\cdots,p_N)$

$\lambda=(\pi,A,B)$

初始prob dist 状态转移矩阵 发射矩阵（从状态到观测）离散+time

概率图 { 有向 - Bayesian Network；无向 - Markov Random Field（Markov Network） }

Dynamic Model（动态 time）{ HMM；kalman Filter；Particle Filter }

HMM 三个问题

①Evaluation $P(O|\lambda)$ → 前向后向

②Learning → λ 如何求 → EM $\lambda=\arg\max P(O|\lambda)$ Baum Welch

③Decoding：找到一个状态序列 $I=\arg\max P(I|O)$

$I=\arg\max P(I|O)$

{ 预测 → $P(i_{t+1}|O_1,O_2,\cdots,O_t)$

滤波 → $P(i_t|O_1,O_2,\cdots,O_t)$

$A=[a_{ij}]$，$a_{ij}=P(i_{t+1}=q_j|i_t=q_i)$

观测变量 $O_1,O_2,\cdots,O_T \to V=\{v_1,v_2,\cdots,v_M\}$

状态变量 $i_1,i_2,\cdots,i_T \Rightarrow Q=\{q_1,q_2,\cdots,q_N\}$

$B=[b_{ij(k)}]$，$b_{ij}(k)=P(O_t=v_k|i_t=q_j)$

两个假设：①齐次Markov假设 ②观测独立假设

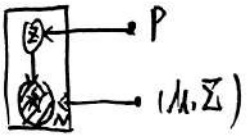
引申 ①后效性：$P(i_{t+1}|i_t,i_{t-1},\cdots,i_1,O_t,O_{t-1},\cdots,O_1)=P(i_{t+1}|i_t)$

② $P(O_t|i_t,i_{t-1},\cdots,i_1,O_{t-1},\cdots,O_1)=P(O_t|i_t)$

观测变量的值域有N个对应 mixture

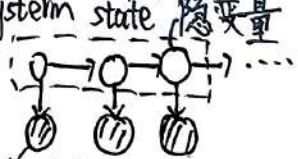
GMM $N, x_1,\cdots,x_N$

x_i 从 $P(x|\theta)$



$P(x|z) \sim N(\mu,\Sigma)$

system state 隐变量

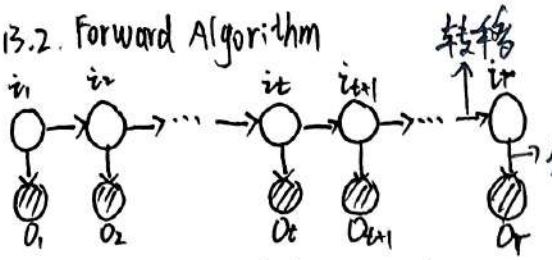


观测变量

state → 离散 HMM

连续 { 线性 → kalman Filter；非线性 → Particle Filter

13.2 Forward Algorithm



$I = i_1 i_2 \cdots i_T$ → 状态序列，$Q = \{q_1, q_2, \cdots, q_N\}$ → 状态值集合

$O = O_1 O_2 \cdots O_T$ → 观测序列，$V = \{v_1, v_2, \cdots, v_M\}$ → 观测值集合

(i) $\lambda = (\pi, A, B)$

π：初始概率分布，$\pi = (\pi_1, \cdots, \pi_N)$，$\sum_{i=1}^{N} \pi_i = 1$，$\pi_i = P(i_1 = q_i)$

A：$[a_{ij}]_{N \times N}$ → 转移矩阵，$a_{ij} = P(i_{t+1} = q_j \mid i_t = q_i)$ $i=1,\cdots,N$ $j=1,\cdots,N$

B：$[b_j(k)]_{N \times M}$ → 发射矩阵，$b_j(k) = P(O_t = v_k \mid i_t = q_j)$ $j=1,\cdots,N$ $k=1,\cdots,M$

(ii) 两个假设：① 齐次 Markov ② 观测独立

1) $P(i_{t+1} \mid i_1, \cdots, i_t, O_1, \cdots, O_t) = P(i_{t+1} \mid i_t)$

2) $P(O_t \mid i_1, \cdots, i_t, O_1, \cdots, O_t) = P(O_t \mid i_t)$

(iii) 三个问题

① Evaluation：Given λ，求 $P(O \mid \lambda)$ Forward-Backward

② Learning：$\hat{\lambda}_{MLE} = \arg\max_{\lambda} P(O \mid \lambda)$ Baum Welch / EM

③ Decoding：$\hat{I} = \arg\max_{I} P(I \mid O, \lambda)$ Viterbi

Evaluation：Given λ，求 $P(O \mid \lambda)$

$$P(O \mid \lambda) = \sum_{I} P(I, O \mid \lambda) = \sum_{I} P(O \mid I, \lambda) \cdot P(I \mid \lambda)$$

$P(I, O \mid \lambda)$ 联合概率

由于 $\sum$ 太复杂，故有 Forward, Backward

$P(i_T \mid i_{T-1}) = a_{i_{T-1}, i_T}$

$$\because P(I \mid \lambda) = P(i_1, i_2, \cdots, i_T \mid \lambda) = P(i_T \mid i_1, i_2, \cdots, i_{T-1}, \lambda) \cdot P(i_1, i_2, \cdots, i_{T-1}, \lambda)$$

$$= a_{i_{T-1}, i_T} \cdot a_{i_{T-2}, i_{T-1}} \cdots a_{i_1, i_2} \cdot \pi(i_1) = \pi(i_1) \cdot \prod_{t=2}^{T} a_{i_{t-1}, i_t}$$

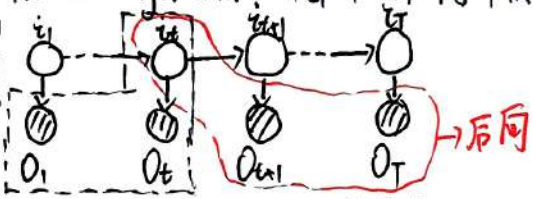
$$P(O \mid I, \lambda) = \prod_{t=1}^{T} b_{i_t}(O_t)$$

$$\therefore P(O \mid \lambda) = \sum_{I} \pi(i_1) \cdot \prod_{t=2}^{T} a_{i_{t-1}, i_t} \prod_{t=1}^{T} b_{i_t}(O_t)$$

$$= \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} \cdots \sum_{i_T=1}^{N} \pi(i_1) \cdot \prod_{t=2}^{T} a_{i_{t-1}, i_t} \cdot \prod_{t=1}^{T} b_{i_t}(O_t)$$

$N^T \Rightarrow O(N^T)$ → 复杂度太高，又往从计算

Forward Algorithm：算出 α_T 再对每一个状态求和



记 $\alpha_t(i) = P(O_1, \cdots, O_t, i_t = q_i \mid \lambda)$

$\alpha_T(i) = P(O, i_T = q_i \mid \lambda)$

$$P(O \mid \lambda) = \sum_{i=1}^{N} P(O, i_t = q_i \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

有 N 个状态值，把它积掉

$$\alpha_{t+1}(j) = P(O_1, \cdots, O_t, O_{t+1}, i_{t+1} = q_j \mid \lambda) = \sum_{i=1}^{N} P(O_1, \cdots, O_t, O_{t+1}, i_{t+1} = q_j, i_t = q_i \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_{t+1} \mid O_1, \cdots, O_t, i_t = q_i, i_{t+1} = q_j, \lambda) \cdot P(O_1, \cdots, O_t, i_t = q_i, i_{t+1} = q_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_{t+1} \mid i_{t+1} = q_j) \cdot P(O_1, \cdots, O_t, O_{t+1}, i_t = q_i, i_{t+1} = q_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_{t+1} \mid i_{t+1} = q_j) \cdot P(i_{t+1} = q_j \mid O_1, \cdots, O_t, i_t = q_i, \lambda) \cdot P(O_1, \cdots, O_t, i_t = q_i \mid \lambda)$$

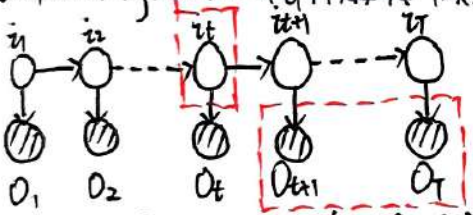
$P(i_{t+1} = q_j \mid i_t = q_i, \lambda)$

$P(O_1, \cdots, O_t, i_t = q_i \mid \lambda) \Rightarrow \alpha_t(i)$

$$= \sum_{i=1}^{N} \underbrace{P(O_{t+1} \mid i_{t+1} = q_j)}_{b_j(O_{t+1})} \cdot \underbrace{P(i_{t+1} = q_j \mid i_t = q_i, \lambda)}_{a_{ij}} \cdot \alpha_t(i)$$

$$\therefore \alpha_{t+1}(j) = \sum_{i=1}^{N} b_j(O_{t+1}) \cdot a_{ij} \cdot \alpha_t(i)$$ Forward

13.3 Backward Algorithm：算出 β_1，再对每一个状态求和



记 $\beta_t(i) = P(O_{t+1}, \cdots, O_T \mid i_t = q_i, \lambda)$

$\beta_1(i) = P(O_2, \cdots, O_T \mid i_1 = q_i, \lambda)$

$$P(O \mid \lambda) = P(O_1, \cdots, O_T \mid \lambda) = \sum_{i=1}^{N} P(O_1, \cdots, O_T, i_1 = q_i \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_1, \cdots, O_T \mid i_1 = q_i, \lambda) \cdot P(i_1 = q_i) \quad \pi_i$$

$$= \sum_{i=1}^{N} P(O_1 \mid O_2, \cdots, O_T, i_1 = q_i, \lambda) \cdot P(O_2, \cdots, O_T \mid i_1 = q_i, \lambda) \cdot \pi_i$$

$$= \sum_{i=1}^{N} P(O_1 \mid i_1 = q_i) \cdot \beta_1(i) \cdot \pi_i$$

$$= \sum_{i=1}^{N} b_i(O_1) \, \beta_1(i) \cdot \pi_i = P(O \mid \lambda)$$

$\beta_t(i) = P(O_{t+1}, \cdots, O_T \mid i_t = q_i, \lambda)$

$$= \sum_{j=1}^{N} P(O_{t+1}, \cdots, O_T, i_{t+1} = q_j \mid i_t = q_i, \lambda)$$

$$= \sum_{j=1}^{N} P(O_{t+1}, \cdots, O_T \mid i_{t+1} = q_j, i_t = q_i, \lambda) \cdot P(i_{t+1} = q_j \mid i_t = q_i, \lambda)$$

$$= \sum_{j=1}^{N} P(O_{t+1}, \cdots, O_T \mid i_{t+1} = q_j, \lambda) \, a_{ij} = \sum_{j=1}^{N} P(O_{t+1} \mid O_{t+2}, \cdots, O_T, i_{t+1} = q_j, \lambda) \cdot P(O_{t+2}, \cdots, O_T \mid i_{t+1} = q_j, \lambda) \cdot a_{ij}$$

$P(O_{t+1} \mid i_{t+1} = q_j, \lambda)$ $\quad \beta_{t+1}(j)$

$$\Rightarrow \sum_{j=1}^{N} b_j(O_{t+1}) \cdot a_{ij} \cdot \beta_{t+1}(j) = \beta_t(i)$$ Backward

$i \to t \to O_{t+1}$，head to tail，若 i_{t+1} 被观测，$i_t \perp O_{t+1}$

13.4 Baum Welch Algorithm - EM

$$\theta^{(t+1)} = \arg\max_{\theta} \int_Z \log P(X,Z|\theta) \cdot P(Z|X,\theta^{(t)})\, dZ$$

X: 观测 → O

Z: 隐变量 → I → 离散

θ: 参数 → λ

$$P(O|\lambda) = \sum_I P(O,I|\lambda) = \sum_{i_1} \cdots \sum_{i_T} \pi(i_1) \prod_{t=2}^{T} a_{i_{t-1},i_t} \prod_{t=1}^{T} b_{i_t}(O_t)$$

$$\pi_i = P(i_1 = q_i)$$

$$\lambda^{(t+1)} = \arg\max_{\lambda} \sum_I \log P(O,I|\lambda) \cdot \underbrace{P(I|O,\lambda^{(t)})}_{\frac{P(I,O|\lambda^{(t)})}{P(O|\lambda^{(t)})}}$$

$P(O|\lambda^{(t)})$ 常数，且 O 与 λ 无关，留下的 $\lambda^{(t)}$ 求迭代式

而 $\lambda^{(t)}$ 是上一次迭代产生的一个

$$\lambda^{(t)} = (\pi^{(t)}, A^{(t)}, B^{(t)})$$

$$= \arg\max_{\lambda} \sum_I \log P(O,I|\lambda) \cdot P(O,I|\lambda^{(t)})$$

$$Q(\lambda,\lambda^{(t)}) = \sum_I \log P(O,I|\lambda) \cdot P(O,I|\lambda^{(t)})$$

$$= \sum_I \log \left[\sum_{i_1} \sum_{i_2} \cdots \sum_{i_T} \pi(i_1) \cdot \prod_{t=2}^{T} a_{i_{t-1},i_t} \cdot \prod_{t=1}^{T} b_{i_t}(O_t) \right] \cdot P(O,I|\lambda^{(t)})$$

$$= \sum_I \left\{ \left[\log \pi_{i_1} + \sum_{t=2}^{T} \log a_{i_{t-1},i_t} + \sum_{t=1}^{T} \log b_{i_t}(O_t) \right] \cdot P(O,I|\lambda^{(t)}) \right\}$$

计算 $\pi^{(t+1)}$.

$$\pi^{(t+1)} = \arg\max_{\pi} Q(\lambda,\lambda^{(t)})$$

$$= \arg\max_{\pi} \sum_I \left[\log \pi_{i_1} \cdot P(O,I|\lambda^{(t)}) \right]$$

$$= \arg\max_{\pi} \sum_{i_1} \cdots \sum_{i_T} \left[\log \pi_{i_1} \cdot P(O,i_1,i_2,\cdots,i_T|\lambda^{(t)}) \right]$$

$$= \arg\max_{\pi} \sum_{i_1} \left[\log \pi_{i_1} \cdot P(O,i_1|\lambda^{(t)}) \right]$$

$$\begin{cases} \arg\max_{\pi} \sum_{i_1}^{N} \left[\log \pi_i \, P(O,i_1=q_i|\lambda^{(t)}) \right] \\ \text{s.t.} \sum_{i_1}^{N} \pi_i = 1 \end{cases}$$

$$\mathcal{L}(\pi,\gamma) = \sum_{i_1}^{N} \left[\log \pi_i \, P(O,i_1=q_i|\lambda^{(t)}) \right] + \gamma \left(\sum_{i_1}^{N} \pi_i - 1 \right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = \frac{1}{\pi_i} P(O,i_1=q_i|\lambda^{(t)}) + \gamma \triangleq 0$$

$$P(O,i_1=q_i|\lambda^{(t)}) + \pi_i \gamma = 0 \Rightarrow \pi_i = -\frac{1}{\gamma} P(O,i_1=q_i|\lambda^{(t)})$$

$$\sum_{i_1}^{N} \left[P(O,i_1=q_i|\lambda^{(t)}) + \sum_{i_1}^{N} \pi_i \gamma \right] = 0$$

$$P(O|\lambda^{(t)}) + \gamma = 0$$

$$\gamma = -P(O|\lambda^{(t)})$$

$$\pi_i^{(t+1)} = \frac{1}{P(O|\lambda^{(t)})} P(O,i_1=q_i|\lambda^{(t)})$$

$$\pi^{(t+1)} = (\pi_1^{(t+1)}, \pi_2^{(t+1)}, \cdots, \pi_N^{(t+1)})$$

$$\alpha_t(i) \cdot \beta_t(i) = P(i_t=q_i, O|\lambda)$$

$$\gamma_t(i) = P(i_t=q_i|O,\lambda) = \frac{P(i_t=q_i,O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}$$

$$\xi_t(i,j) = P(i_t=q_i, i_{t+1}=q_j|O,\lambda)$$

$$\xi_t(i,j) = \frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{P(O|\lambda)} = \frac{P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}{\sum_{i=1}^{N}\sum_{j=1}^{N} P(i_t=q_i,i_{t+1}=q_j,O|\lambda)}$$

$$P(i_t=q_i,i_{t+1}=q_j,O|\lambda) = \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)$$

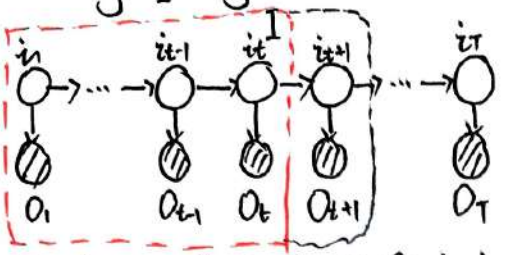
$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}$$

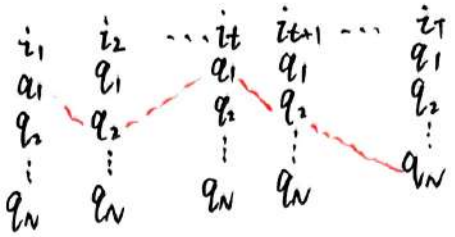
故 $a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$, $b_j(k) = \frac{\sum_{t=1, O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$

$$\pi_i = \gamma_1(i)$$

13.5 Viterbi Algorithm

Decoding: $\hat{I} = \arg\max P(I|O,\lambda)$ → Viterbi





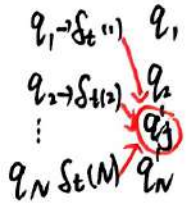
找一个概率最大的序列

$\delta_t(i) = \max_{i_1,i_2,\cdots,i_{t-1}} P(O_1,O_2,\cdots,O_t,i_1,i_2,\cdots,i_{t-1},i_t=q_i)$，condition λ 已知，路径 I 到 t 时刻 q_i 的 t 时刻的最大概率值

概率值　已经确定　即从开始到 t 时刻 q_i 的 max prob

$\delta_{t+1}(j) = \max_{i_1,i_2,\cdots,i_t} P(O_1,O_2,\cdots,O_t,O_{t+1},i_1,i_2,\cdots,i_t,i_{t+1}=q_j)$

最大的一个概率值 $= \max_{1\le i\le N} \delta_t(i)\, a_{ij}\, b_j(O_{t+1})$



从 t 到 $t+1$ 中 $\delta_{t+1}(1\sim N)$ 到 $\delta_{t+1}(j)$ 最大的一个

$\Psi_{t+1}(j) = \arg\max_{1\le i\le N} \delta_t(i)\cdot a_{ij}$　（t 时刻；转移矩阵；转移；最大转移）

→ 这个得到的是 t 时刻到 $t+1$ 时刻 q_j 处 P 最大的点（t 时刻）即 $\delta_t(i)$ 中的 i

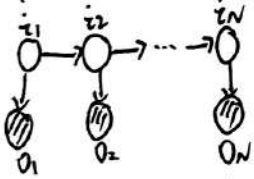
$\delta_t(i) \to \delta_{t+1}(j)$ 最大的路径的 i，即 q_i 的位置

13.6 Summary

Dynamic Model:

HMM (mixture + time)

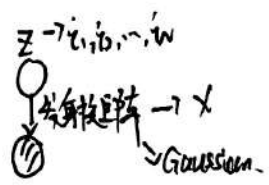
HMM = GMM + time



离散 $i_1,i_2,\cdots,i_N$

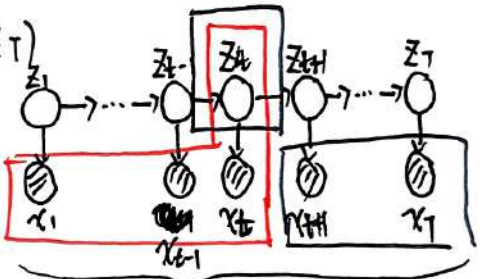
$P(O_j|i_j) \sim N(\mu,\Sigma)$

GMM:



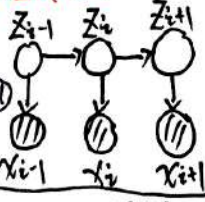
Forward: $\alpha_t = P(Y_1,\cdots,Y_t,Z_t)$

Backward: $\beta_t = P(Y_{t+1},\cdots,Y_T|Z_T)$



Dynamic Model → state space model

EM：使用 Forward-Backward

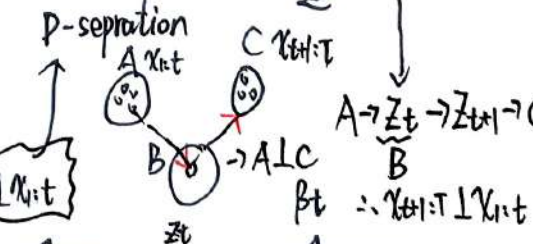


Y 观测值

Z：隐变量

- learning（已知 x）：$\lambda_{MLE} = \arg\max_\lambda P(x|\lambda)$ Baum-Welch ①
- Inference（已知 λ）：
  - decoding: $P(z_1,\cdots,z_t|Y_1,\cdots,Y_t)$ → Viterbi Algorithm ②
  - prob of evidence: $P(x|\theta) = P(Y_1,Y_2,\cdots,Y_T|\theta)$ Forward Algorithm ③、Backward Algorithm ④
  - filtering: $P(Z_t|Y_1,Y_2,\cdots,Y_t)$ → online → $P(Z_1|Y_1) \to P(Z_2|Y_1,Y_2) \to \cdots \to P(Z_t|Y_1,\cdots,Y_t)$ Forward Algorithm
  - smoothing: $P(Z_t|Y_1,Y_2,\cdots,Y_T)$ → offline Forward-Backward Algorithm ⑤
  - prediction: $P(Z_{t+1},Z_{t+2}|Y_1,Y_2,\cdots,Y_t)$ Forward 预测下一段时刻 / $P(Y_{t+1},Y_{t+2}|Y_1,Y_2,\cdots,Y_t)$

看图.



$A \to Z_t \to Z_{t+1} \to C$

$\therefore Y_{t+1:T} \perp Y_{1:t}$

$$\text{Filtering: } P(Z_t|Y_1,\cdots,Y_t) = \frac{P(Z_t,Y_{1:t})}{P(Y_{1:t})} = \frac{P(Y_{1:t},Z_t)}{\sum_{Z_t} P(Y_{1:t},Z_t)} \propto P(Y_{1:t},Z_t) = \alpha_t$$

分母=常数化因子（$Y_{1:t}$ 都已知）

$$\text{Smoothing: } P(Z_t|Y_{1:T}) = \frac{P(Y_{1:T},Z_t)}{P(Y_{1:T})} = \frac{P(Y_{1:T},Z_t)}{\sum_{Z_t} P(Y_{1:T},Z_t)}$$

拆开前后向

$$P(Y_{1:T},Z_t) = P(Y_{1:t},Y_{t+1:T},Z_t) = P(Y_{t+1:T}|Y_{1:t},Z_t)\cdot P(Y_{1:t},Z_t) = P(Y_{t+1:T}|Y_{1:t},Z_t)\cdot\alpha_t = P(Y_{t+1:T}|Z_t)\cdot\alpha_t = \beta_t\alpha_t$$

α_t：$P(Y_{1:t},Z_t)$；β_t：$P(Y_{t+1:T}|Z_t)$

$Y_{t+1:T} \perp Y_{1:t}$

给定 $B(Z_t)$，A.（阻塞）

P-separation

P44

∴ Smoothing 中: $P(Z_t|\gamma_{1:T}) \propto P(\gamma_{1:T}, Z_t) = \alpha_t \beta_t$

Prediction: (利用两个假设)

$$P(Z_{t+1}|\gamma_{1:t}) = \sum_{Z_t} P(Z_{t+1}, Z_t|\gamma_{1:t}) = \sum_{Z_t} \underbrace{P(Z_{t+1}|Z_t, \gamma_{1:t})}_{P(Z_{t+1}|Z_t)} \cdot \underbrace{P(Z_t|\gamma_{1:t})}_{\text{Filtering}} = \sum_{Z_t} P(Z_{t+1}|Z_t) \cdot P(Z_t|\gamma_{1:t})$$

转移矩阵 已知

$$P(\gamma_{t+1}|\gamma_{1:t}) = \sum_{Z_{t+1}} P(\gamma_{t+1}, Z_{t+1}|\gamma_{1:t}) = \sum_{Z_{t+1}} \underbrace{P(\gamma_{t+1}|Z_{t+1}, \gamma_{1:t})}_{P(\gamma_{t+1}|Z_{t+1})} \cdot P(Z_{t+1}|\gamma_{1:t}) \propto \sum_{Z_{t+1}} P(\gamma_{t+1}|Z_{t+1}) \cdot P(Z_{t+1}|\gamma_{1:t})$$

已知 发射矩阵

---

14. Linear Dynamic System (Kalman Filter)

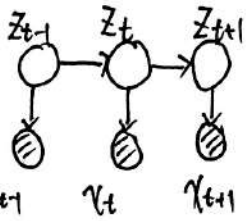
14.1 Background

Dynamic Model (State space model)

- HMM: state is discrete (离散, 转移矩阵, 发射矩阵)
- Linear Dynamic System (Kalman filter): Linear Gaussian Model
- Partical Filter: Non-linear, Non Gaussian → 连续, 状态转移方程, 观测方程

⇒ tasks: Learning, Inference (P44)

HMM: $\lambda = (\pi, A, B)$



$A = [A_{ij}]_{N\times N}$, $A_{ij} = P(Z_{t+1} = q_j | Z_t = q_i)$

$B = [b_j(k)]_{N\times M}$, $b_j(k) = P(O_t = v_k | Z_t = q_j)$

Kalman filter:

$$\begin{cases} Z_t = A \cdot Z_{t-1} + B + \varepsilon \\ \gamma_t = C \cdot Z_t + D + \delta \end{cases} \Leftarrow \text{Gaussian Dist}$$

(Linear)

$\varepsilon \sim N(0, Q)$, $P(Z_t|Z_{t-1}) \sim N(A \cdot Z_{t-1} + B, Q)$

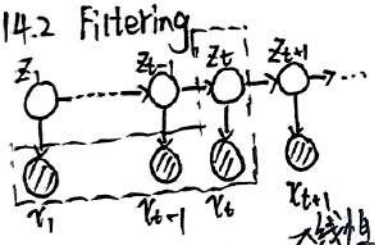
$\delta \sim N(0, R)$, $P(\gamma_t|Z_t) \sim N(C \cdot Z_t + D, R)$

$Z_1 \sim N(\mu_1, \Sigma_1)$

$\theta = (A, B, C, D, Q, R, \mu_1, \Sigma_1)$

关于 μ 与 Σ 做 Prediction 和 update

14.2 Filtering



$$\begin{cases} P(Z_t|Z_{t-1}) = N(A \cdot Z_{t-1} + B, Q) \\ P(\gamma_t|Z_t) = N(C \cdot Z_t + D, R) \\ P(Z_1) = N(\mu_1, \Sigma_1) \end{cases}$$ 特性

Kalman Filter 是 Linear Gaussian Model

$$\begin{cases} Z_t = AZ_{t-1} + B + \varepsilon, & \varepsilon \sim N(0, Q) \\ \gamma_t = C \cdot Z_t + D + \delta, & \delta \sim N(0, R) \end{cases}$$ 线性 / 噪声 Gauss

$\theta = (A, B, C, D, Q, R, \mu_1, \Sigma_1)$

Learning

Inference: $P(Z|\gamma)$

- decoding → HMM (Viterbi)
- prob of evidence → $P(\gamma_t|\gamma_{t-1}) \to$ 求和消元
- marginal posterior:
  - fitering: $P(Z_t|\gamma_1, \cdots, \gamma_t)$
  - Smoothing: $P(Z_t|\gamma_1, \cdots, \gamma_T)$
  - prediction:
    - $P(Z_t|\gamma_1, \cdots, \gamma_{t-1})$
    - $P(\gamma_t|\gamma_1, \cdots, \gamma_{t-1})$

$$\underbrace{P(Z_t|\gamma_1, \cdots, \gamma_t)}_{\text{Fitering}} = \frac{P(\gamma_1, \cdots, \gamma_t, Z_t)}{P(\gamma_1, \cdots, \gamma_t)} \propto P(\gamma_1, \cdots, \gamma_t, Z_t) = \underbrace{P(\gamma_t|\gamma_1, \cdots, \gamma_{t-1}, Z_t)}_{P(\gamma_t|Z_t)} \cdot P(\gamma_1, \cdots, \gamma_{t-1}, Z_t) = P(\gamma_t|Z_t) \cdot P(\gamma_1, \cdots, \gamma_{t-1}, Z_t)$$

$P(\gamma_1, \cdots, \gamma_t)$ 常值 ; 符合两个假设

$$= P(\gamma_t|Z_t) \cdot \underbrace{P(Z_t|\gamma_1, \cdots, \gamma_{t-1})}_{\text{prediction}} \cdot P(\gamma_1, \cdots, \gamma_{t-1}) \propto \underbrace{P(\gamma_t|Z_t)}_{\text{update}} \cdot \underbrace{P(Z_t|\gamma_1, \cdots, \gamma_{t-1})}_{\text{Prediction}}$$

$P(\gamma_t|Z_t)$

$$\int_{Z_{t-1}} P(Z_t, Z_{t-1}|\gamma_1, \cdots, \gamma_{t-1})\, dZ_{t-1}$$

故: $P(Z_t|\gamma_1, \cdots, \gamma_t) = \int_{Z_{t-1}} P(Z_t|Z_{t-1}) \cdot P(Z_{t-1}|\gamma_1, \cdots, \gamma_{t-1})\, dZ_{t-1} = P(\gamma_t|Z_t) \cdot P(Z_t|\gamma_1, \cdots, \gamma_{t-1})$

$$\int_{Z_{t-1}} P(Z_t|Z_{t-1}, \gamma_1, \cdots, \gamma_{t-1}) \cdot P(Z_{t-1}|\gamma_1, \cdots, \gamma_{t-1})\, dZ_{t-1}$$

$$\int_{Z_{t-1}} P(Z_t|Z_{t-1}) \cdot P(Z_{t-1}|\gamma_1, \cdots, \gamma_{t-1})\, dZ_{t-1}$$

连续变量的 Prediction

步骤: $t=1$, $\begin{cases} P(Z_1|\gamma_1) \to \text{update} \\ P(Z_2|\gamma_1) \to \text{prediction} \end{cases}$

$t=T$ $\begin{cases} P(Z_T|\gamma_1, \cdots, \gamma_T) \\ P(Z_{T+1}|\gamma_1, \cdots, \gamma_T) \end{cases}$

online

$t=2$, $\begin{cases} P(Z_2|\gamma_1, \gamma_2) \to \text{update (correction)} \\ P(Z_3|\gamma_1, \gamma_2) \to \text{prediction} \end{cases}$

$p(x) = N(x|u, \Lambda^{-1})$ 已知
$p(y|x) = N(y|Ax+b, L^{-1})$ 已知
$p(y) = N(y|Au+b, L^{-1}+A\Lambda^{-1}A^T)$ 未知

$p(x|y) = N(u+\Lambda^{-1}A^T(L^{-1}+A\Lambda^{-1}A^T)^{-1}(y-Au-b), \Lambda^{-1}-\Lambda^{-1}A^T(L^{-1}+A\Lambda^{-1}A^T)^{-1}A\Lambda^{-1})$

P4推导

→这里有 $\Lambda^{-1}=\Sigma_{t-1}$ 记作 $u=u_{t-1}$ / u_t^*
$L^{-1}=Q/R$, $A=A/C$
$b=B/D$
$y=z_t$ / γ_t

$\begin{cases} P(z_t|z_{t-1}) = N(A\cdot z_{t-1}+B, Q) \\ P(\gamma_t|z_t) = N(C\cdot z_t+D, R) \\ P(z_1) = N(u_1, \Sigma_1) \end{cases}$ } kalman Filter特性

代P45公式

Filtering:

step 1: prediction: (未看到γ_t) 代公式↑ update $p(x)$

$P(z_t|\gamma_1,\cdots,\gamma_{t-1}) = \int_{z_{t-1}} P(z_t|z_{t-1})\cdot P(z_{t-1}|\gamma_1,\cdots,\gamma_{t-1})dz_{t-1}$

$p(y)$ ← $N(u_t^*, \Sigma_t^*) = \int N(z_t|Az_{t-1}+B, Q)\cdot N(u_{t-1}, \Sigma_{t-1})$
$p(y|x)$ $p(x)$

$\begin{cases} u_t^* = A\cdot u_{t-1}+B \\ \Sigma_t^* = Q+A\cdot\Sigma_{t-1}\cdot A^T \end{cases}$ 求 ⇒ step 1 中的参数

由$z_t|z_{t-1}$关系与z_{t-1}此时状态，去测z_t
且$(\gamma_1,\cdots,\gamma_{t-1})$为已观测到

step 2: update (看到γ_t)

$p(x|y)$ $p(y|x)$ $p(x)$

$P(z_t|\gamma_1,\cdots,\gamma_t) \propto P(\gamma_t|z_t)\cdot P(z_t|\gamma_1,\cdots,\gamma_{t-1})$

$N(u_t, \Sigma_t) \propto N(\gamma_t|Cz_t+D, R)\cdot N(u_t^*, \Sigma_t^*)$
$p(y|x)$ $p(x)$

$p(z_t|\gamma_{1:t}) \propto p(\gamma_t|z_t)\cdot p(z_t)$ $\gamma_1,\cdots,\gamma_{t-1}$为观测信息，在step 1中已经看到

$p(x|y)$ $p(y|x)$ $p(x)$

$\begin{cases} u_t = u_t^* + \Sigma_t^* C^T(R+C\Sigma_t^* C^T)^{-1}(\gamma_t - Cu_t^* - D) \\ \Sigma_t = \Sigma_t^* - \Sigma_t^* C^T(R+C\Sigma_t^* C^T)^{-1}C\Sigma_t^* \end{cases}$ ⇒ step 2 中的参数

状态转移矩阵是 $\begin{cases} u_t = u_{t-1}+\Sigma_{t-1}A^T(Q+A\Sigma_{t-1}A^T)^{-1}(z_t - Au_{t-1}-B) \\ \Sigma_t = \Sigma_{t-1}-\Sigma_{t-1}A^T(Q+A\Sigma_{t-1}A^T)^{-1}A\Sigma_{t-1} \end{cases}$

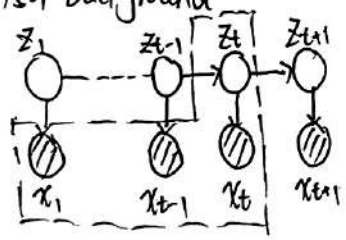
先预测，看到γ_t后再做更新

看到了γ_t去根据 $z_t|\gamma_{1:t-1}$ $(\gamma_t|z_t$ 做 update

---

15. Particle Filter

15.1 Background



Filtering 问题：求解 $P(z_t|\gamma_1,\cdots,\gamma_t)$

$\begin{cases} z_t = g(z_{t-1}, u, \varepsilon) \\ \gamma_t = h(z_t, u, \delta) \end{cases}$ 非线性，噪声非高斯

状态转移方程是 / 观测方程是

step 1: prediction:
$P(z_t|\gamma_1,\cdots,\gamma_{t-1}) = \int_{z_{t-1}} P(z_t|z_{t-1})\cdot P(z_{t-1}|\gamma_1,\cdots,\gamma_{t-1})dz_{t-1}$ → P43. prediction

step 2: update:
$P(z_t|\gamma_1,\cdots,\gamma_t) \propto P(\gamma_t|z_t)\cdot P(z_t|\gamma_1,\cdots,\gamma_{t-1})$ → P43. update

15.2 Importance Sampling & SIS

Monte Carlo Method.

$P(z|x)$ $E_{z|x}[f(z)] = \int f(z)\cdot P(z|x)dz \approx \frac{1}{N}\sum_{i=1}^N f(z^{(i)})$

若我们能够从$P(z|x)$中采样，则有N个样本 $z^{(i)} \sim P(z|x)$: $z^{(1)}, z^{(2)}, \cdots, z^{(N)}$

若无法直接采样，构建到从$q(z)$ $q(z)$：提议分布 proposal distribution

Importance Sampling:

$E[f(z)] = \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}\cdot q(z)dz \cdots ①$

$q(z)$是一个已知，可采样的分布，$z^{(i)} \sim q(z)$, $i=1,2,\cdots,N$

则①式 $= \frac{1}{N}\sum_{i=1}^N f(z^{(i)})\cdot\frac{p(z^{(i)})}{q(z^{(i)})} = \frac{1}{N}\sum_{i=1}^N f(z^{(i)})w^{(i)} = \sum_{i=1}^N f(z^{(i)})\hat{w}^{(i)}$

weight $w^{(i)}$ 不好求 归一化

Filtering

$P(z_t|\gamma_1,\cdots,\gamma_t)$, $w_t^{(i)} = \frac{p(z_t^{(i)}|\gamma_1,\cdots,\gamma_t)}{q(z_t^{(i)}|\gamma_1,\cdots,\gamma_t)}$

$t=1$: $w_1^{(i)}$, $i=1,\cdots,N \to w_1^1, w_1^2, \cdots, w_1^N$

$t=2$: $w_2^{(i)}$, $i=1,\cdots,N \to w_2^1, w_2^2, \cdots, w_2^N$

由于不易求的问题，是否找得到 $w_{t-1}^{(i)} \to w_t^{(i)}$ 存在关系

引出：

Sequential Importance Sampling: SIS

此模型关注的是 $P(z_1,\cdots,z_t|\gamma_1,\cdots,\gamma_t)$

$w_t^{(i)} \propto \frac{P(z_1,\cdots,z_t|\gamma_1,\cdots,\gamma_t)}{q(z_1,\cdots,z_t|\gamma_1,\cdots,\gamma_t)}$

$P(z_1,\cdots,z_t|\gamma_1,\cdots,\gamma_t) = \frac{P(z_{1:t},\gamma_{1:t})}{P(\gamma_{1:t})} = \frac{1}{C}P(z_{1:t},\gamma_{1:t})$

C constant

$= \frac{1}{C}\cdot P(\gamma_t|z_{1:t},\gamma_{1:t-1})\cdot P(z_{1:t},\gamma_{1:t-1})$

$= \frac{1}{C}P(\gamma_t|z_t)\cdot P(z_{1:t},\gamma_{1:t-1})$

$= \frac{1}{C}P(\gamma_t|z_t)\cdot P(z_t|z_{1:t-1},\gamma_{1:t-1})P(z_{1:t-1},\gamma_{1:t-1})$

$= \frac{1}{C}P(\gamma_t|z_t)\cdot P(z_t|z_{t-1})\cdot P(z_{1:t-1},\gamma_{1:t-1})$

$= \frac{1}{C}P(\gamma_t|z_t)\cdot P(z_t|z_{t-1})\cdot P(z_{1:t-1}|\gamma_{1:t-1})\cdot P(\gamma_{1:t-1})$

$= \frac{D}{C}P(\gamma_t|z_t)\cdot P(z_t|z_{t-1})\cdot P(z_{1:t-1}|\gamma_{1:t-1})$

$q(Z_{1:t}|X_{1:t}) = q(Z_t|Z_{1:t-1},X_{1:t}) \cdot q(Z_{1:t-1}|X_{1:t})$

假设有: $q(Z_t|Z_{1:t-1},X_{1:t}) = q(Z_t|Z_{t-1})$ ，q是我们指定的，具有任意性，根据条件独立性，若q分布也符合Markov齐次，则 此步 $\alpha = \frac{P(X_t|Z_t)\cdot P(Z_t|Z_{t-1})}{q(Z_t|Z_{t-1})}$

$q(Z_{1:t}|X_{1:t}) = q(Z_{1:t-1}|X_{1:t-1})$

$$W_t^{(i)} \propto \frac{p(Z_{1:t}|X_{1:t})}{q(Z_{1:t}|X_{1:t})} \propto \frac{P(X_t|Z_t)\cdot P(Z_t|Z_{t-1})\, p(Z_{1:t-1}|X_{1:t-1})}{q(Z_t|Z_{1:t-1},X_{1:t})\, q(Z_{1:t-1}|X_{1:t-1})} = \frac{P(X_t|Z_t)\cdot P(Z_t|Z_{t-1})}{q(Z_t|Z_{1:t-1},X_{1:t})} \cdot W_{t-1}^{(i)} \propto$$

→ 权重计算

假定初始时 $W_1^{(i)}$ 已算好，则 $t=2$, $W_2^{(i)} = \alpha W_1^{(i)}$ 以此类推，解决了W计算难的问题

权值退化，随着t增大，$\lim_{t\to\infty} W_t = 0$，因为每一个$W_t$都<1，而 $\sum_{i=1}^{t} W_i = 1$，就像铺满(0,1)区间，权值退化或小的越小的太大

15.3 Particle Filter ↔ Re-Sampling

Sequential Importance Sampling Filter (SIS)

$$W_t^{(i)} \propto \frac{p(Z_{1:t}|X_{1:t})}{q(Z_{1:t}|X_{1:t})} \propto \frac{p(X_t|Z_t)p(Z_t|Z_{t-1})}{q(Z_t|Z_{1:t-1},X_{1:t})} W_{t-1}^{(i)} \Rightarrow$$

问题：权值退化 $W_t^{(i)}$

比如有100个$W_t^{(i)}$，99个都为0.001，1个接近1

~~resolution~~ 解决：Re-sampling (Basic Particle Filter)

选择一个合适的 proposal dist $q(z)$ (SIR Filter)

algorithm

前提 $t-1$ 时刻采样已完成 $\Rightarrow W_{t-1}^{(i)}$ 已知

t时刻：① Sampling for $i=1,\dots,N$

$$Z_t^{(i)} \sim q(Z_t|Z_{t-1}^{(i)}, X_{1:t})$$

$$W_t^{(i)} \propto W_{t-1}^{(i)} \cdot \frac{p(X_t|Z_t^{(i)})p(Z_t^{(i)}|Z_{t-1}^{(i)})}{q(Z_t^{(i)}|Z_{1:t-1},X_{1:t})}$$

end

② Normalized $W_t^{(i)}$ 归一化，$\sum_{i=1}^{N} \hat{W}_t^{(i)} = 1$ 即 $W_t^{(i)} \to \hat{W}_t^{(i)}$

③ Resampling $\hat{W}_t^{(i)} = \frac{1}{N}$，(所有粒子权重相同，以个数代替权重)

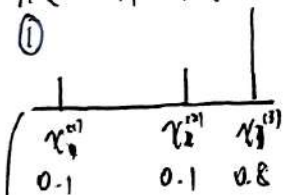
Basic Particle Filter —— 加上的Resampling

Basic Particle Filter：SIS + Resampling 最基本的resampling方法

假设采样完成，得到 x_1, x_2, x_3

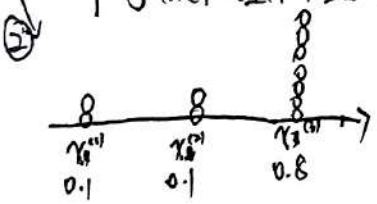
① 对于3个样本 $x^{(1)}, x^{(2)}, x^{(3)}$ 0.1 0.1 0.8 → pdf

0.1 0.2 1 ↔ cdf

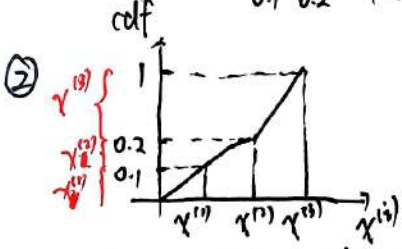


第一次采样，得到3个数据以及各自权重

resampling (以权重作为离散分布去采样)



resampling，多份数据权重是用个数表示

② 

$u \sim U(0,1)$ 采样 在[0,1)采样

$[0, 0.1) = x^{(1)}$

$[0.1, 0.2) = x^{(2)}$

$[0.2, 1] = x^{(3)}$

总结：这样就避免了权值退化，一开始的可采样多个样本，之后停止采样，做cdf图，通过在(0,1)平均分布中采样，落到哪个样本区间就采样哪个

15.4 SIR Filter 选择一个更好的提议分布

Sampling-Importance-Resampling: SIS + Resampling + $q(Z_t|Z_{1:t-1},X_{1:t}) = p(Z_t|Z_{t-1})$

algorithm:

前提 $t-1$ 时刻采样已完成 $\Rightarrow W_{t-1}^{(i)}$ 已知

t时刻

① Sampling: for $i=1,\dots,N$:

$$Z_t^{(i)} \sim p(Z_t|Z_{t-1}^{(i)})$$

$$W_t^{(i)} \sim p(X_t|Z_t^{(i)}) \cdot W_{t-1}^{(i)}$$

end

② Normalized: $W_t^{(i)} \to \hat{W}_t^{(i)}$, $\sum_{i=1}^{N} \hat{W}_t^{(i)} = 1$

② Resampling $\hat{W}_t^{(i)} = \frac{1}{N}$

"generate and test" 从自己身上找 $Z_t^{(i)}$

$Z_t^{(i)} \sim p(Z_t|Z_{t-1}^{(i)})$ $W_t^{(i)} \sim p(X_t|Z_t^{(i)}) \cdot W_{t-1}^{(i)}$

新生成的 上一时刻权重

若 $W_t^{(i)}$ 越大，则说明 $Z_t^{(i)}$ 取的点比较合适，更能体现这份布

IS → SIS → SIR

16. Conditional Random Field

16.1 Background

max margin classifier

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

(凸向间隔)

- SVM
- 感知机 (误分类驱动) $f(w)=\text{sign}(w^Tx+b)$ — PLA
- LDA — 线性判别分析, 类间大, 类内小

硬分类

Logistic Regression (概率判别模型) ⇒ Maximum Entropy Model (log linear 对数线性模型)

Maximum Entropy Principle (给定均值和方差, Gaussian dist 熵最大)

soft max 软分类

(P(y|x)建模)

Maximum Entropy Markov Model → MEMM → CRF (判别)

(P(x,y)建模) Naive Bayes ⇒ Hidden Markov Model

$\lambda=(\pi,A,B)$ 离散

$y_{(0,1,\cdots)} \to seq$

+time

Gaussian Mixture Model

(概率生成模型)

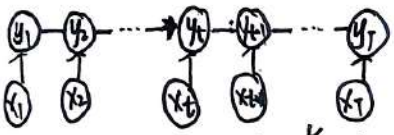
朴素贝叶斯假设

$$P(x|y=V_0)=\prod_{i=1}^{n}P(x_i|y=V_0)$$

Naive Bayes:

$x_i \perp x_j | y, i \neq j$

CRF: 去MEMM箭头, P(Y|X)



全局归一化

$$P(Y|X)=\frac{1}{Z}\sum_{i=1}^{K}\phi_{C_i}(X_{C_i})$$

最后归一化 / 最大团

(John Lafferty)

label bias problem

标注偏差 (原因: 局部归一化)

扬弃了3个HMM的2个假设

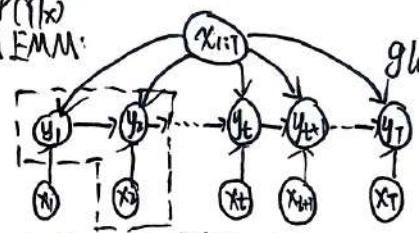
① 齐次Markov

② 观测独立 → P(X,Y)

y → 隐变量 → y ~ 离散分布 1,2,…,K

x → 观测变量 → x|y ~ Gaussian Dist

P(Y|X) MEMM:



global input

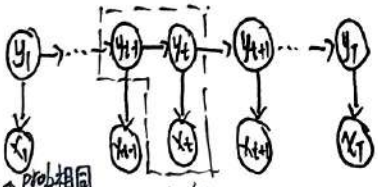
local input

并不符合独立性假设 (HMM中)

会受多因素影响 ②

16.2 HMM vs MEMM

生成模型 (Naive Bayes, GMM) — Naive Bayes 推→ HMM



prob相同

$\lambda=(\pi,A,B)$ 统计得到

HMM中:

$$P(x|\lambda)=\sum_{Y}P(x,Y|\lambda)$$ P42

建模对象

① 齐次一阶Markov

② 观测独立假设

$$P(y_t|y_{1:t-1},x_{1:t-1})=P(y_t|y_{t-1})$$

$$P(x_t|y_{1:t},x_{1:t-1})=P(x_t|y_t)$$

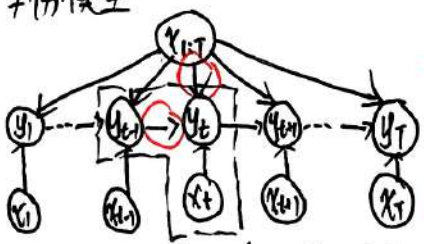
建模对象: $P(X,Y|\lambda)=P(x|Y,\lambda)\cdot P(Y|\lambda)$

$$P(X,Y|\lambda)=\prod_{t=1}^{T}P(x_t,y_t|\lambda)$$

发射 / 转移

$$=\prod_{t=1}^{T}P(y_t|y_{t-1},\lambda)\cdot P(x_t|y_t,\lambda)$$

判别模型



global input

local input

给定 y_t, x_{t-1} 与 x_t 有联系, 并不独立, D-separation

建模对象: $P(Y|X,\lambda)$

$$P(Y|X,\lambda)=\prod_{t=1}^{T}P(y_t|y_{t-1},x_{1:T},\lambda)$$

看 y_t 时指向它的箭头用 global input

打破了观测独立假设 (更加合理)

$P(Y|X,\lambda)$ 比 $P(X,Y|\lambda)$ 相对容易一些

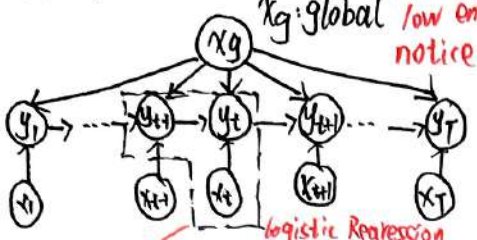
优点: 判别式模型, 计算量更少

打破观测独立性

缺点: Label Bias Problem

16.3 MEMM VS. CRF

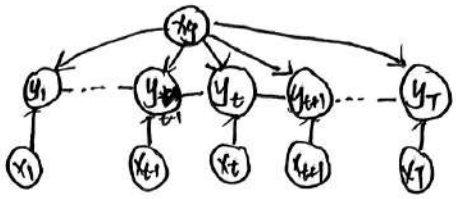
MEMM:



condition dist with low entropy take less notice of observation

有向 ⟹ 无向

Chain-structured CRF



克服了 Label Bias problem

建模对象: $P(Y|x,\lambda) = \prod_{t=1}^{T} P(y_t|y_{t-1}, x_{1:T}, \lambda)$ 为了简化，看成 y 的初始状态的表述

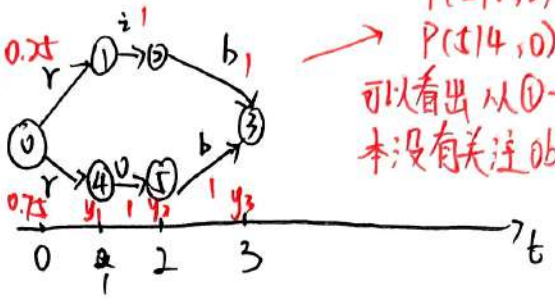
y_T 受两个 y_{t+1} 与 x_t 影响

mass score → 归一化 → 根源因

function > 0

由 y_{t-1}, y_t, x_t 组成

缺点: Label Bias Problem



$P(2|1,i) = 1 = P(2|1)$

$P(5|4,0) = 1 = P(5|4)$

可以看出从 ①→②, ④→⑤, 根本没有关注 observation (i, o)

Decoding: $Y = (y_1, y_2, y_3)$

$\hat{Y} = \arg\max_{y_1,y_2,y_3} P(y_1, y_2, y_3 | x_{rib})$

$= 0 \to 4 \to 5 \to 3$

$=$ rob. 显然不合理 (这就是归一化，忽视 observation 的结果) 就是说，图中 1→2; 2→3, 4→5, 5→3, 上的 observation 没有任何作用，只与 traning 数据相关

eg. 样本 4 个，3 个 rob, 1 个 rib. 则在图上标注概率

traning: $P(1|00, Y) = 0.2$

$P(4|0, Y) = 0.7$

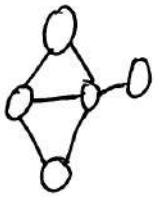
viterbi: Decoding.

16.4 pdf-parameter perspective

MRF 因子分解: Markov Random Field

$$P(x) = \frac{1}{Z}\prod_{i=1}^{K} \psi_i(x_{ci}) = \frac{1}{Z}\prod_{i=1}^{K} \exp[-E_i(x_{ci})] = \frac{1}{Z}\exp\sum_{i=1}^{K} F_i(x_{ci})$$

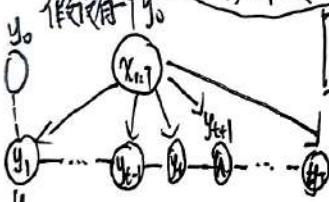
势函数 > 0　　能量函数



K 个最大团

C_i: 最大团

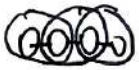
Start 假设有 t 个 y_0



线性链条件随机场 以下讨论图结 LC-CRF

LC-CRF

From here



4 个点 3 个最大团

CRF-PDF

$$P(Y|x) = \frac{1}{Z}\exp\sum_{t=1}^{T}\left[\sum_{k=1}^{K}\lambda_k f_k(y_{t-1}, y_t, x_{1:T}) + \sum_{l=1}^{L}\mu_l g_l(y_t, x_{1:T})\right]$$

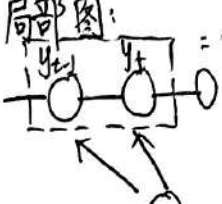
$$P(Y|x) = \frac{1}{Z}\exp\sum_{i=1}^{K} F_i(x_{ci})$$

$$= \frac{1}{Z}\exp\sum_{t=1}^{T} F_t(y_{t-1}, y_t, x_{1:T})$$

由于是线性链，F_t 其实有 4 个团

$$= \frac{1}{Z}\exp\sum_{t=1}^{T} F(y_{t-1}, y_t, x_{1:T})$$

局部图:



$$= \frac{1}{Z}\exp\sum_{t=1}^{T}\left[\sum_{k=1}^{K}\lambda_k f_k(y_{t-1}, y_t, x_{1:T}) + \sum_{l=1}^{L}\mu_l g_{local}(y_t, x_{1:T})\right]$$

↑ CRF-PDF

状态函数　　转移函数

$F(y_{t-1}, y_t, x_{1:T}) = \Delta_{y_{t-1}, x_{1:T}} + \Delta_{y_t, x_{1:T}} + \Delta_{y_{t-1}, y_t, x_{1:T}}$

$= \Delta_{y_{t-1}, x_{1:T}} + \Delta_{y_t, x_{1:T}}$

标准问题: 关于 "小明是中国人" 这一词性问题

$y_{t-1}, y_t \in \{n, v, adj, \dots\}$

$\Delta_{y_{t-1}, y_t, x_{1:T}} = \sum_{k=1}^{K}\lambda_k f_k(y_{t-1}, y_t, x_{1:T})$

$\Delta_{y_t, x_{1:T}} = \sum_{l=1}^{L}\mu_l g_{local}(y_t, x_{1:T})$

f_k: 特征函数

$f_k\{y_{t-1} = n, y_t = v, x_{1:T}\} = 1$

$f_k\{y_{t-1} = n, y_t = \text{助词}, x_{1:T}\} = 0$

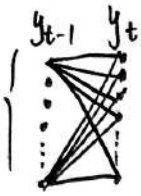
f_k, g_l 一般是给定的特征函数，λ_k, μ_l 参数

16.5 pdf-vector perspective

$$P(Y|x)=\frac{1}{Z}\exp\sum_{t=1}^{T}[\Delta_{转移}+\Delta_{状态}]=\frac{1}{Z}\exp\sum_{t=1}^{T}\Big[\sum_{k=1}^{K}\lambda_k f_k(y_{t-1},y_t,x_{1:T})+\sum_{l=1}^{L}\gamma_l g_l(y_t,x_{1:T})\Big]$$

关于 K,L 的取值

假设 $y_t\in S=\{动,名,副,助,\cdots\}$

而 

共 $|S|$ 个单子
总共 $|S|^2$ 个组合
则 $K\le|S|^2$，且重要程度由 λ_k 决定
λ 的取值与 y_t 的集合个数相关

化简替换：

$$y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_T\end{pmatrix};\ x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_T\end{pmatrix};\ \lambda=\begin{pmatrix}\lambda_1\\\lambda_2\\\vdots\\\lambda_K\end{pmatrix};\ \gamma=\begin{pmatrix}\gamma_1\\\gamma_2\\\vdots\\\gamma_L\end{pmatrix};$$

$$f=\begin{pmatrix}f_1\\f_2\\\vdots\\f_K\end{pmatrix}=f(y_{t-1},y_t,x);\ g=\begin{pmatrix}g_1\\g_2\\\vdots\\g_L\end{pmatrix}=g(y_t,x)$$

则有：$P(Y=y|X=x)=\frac{1}{Z(x,\lambda,\gamma)}\exp\sum_{t=1}^{T}[\lambda^T\cdot f(y_{t-1},y_t,x)+\gamma^T\cdot g(y_t,x)]$

$$=\frac{1}{Z(x,\lambda,\gamma)}\exp\Big[\lambda^T\cdot\sum_{t=1}^{T}f(y_{t-1},y_t,x)+\gamma^T\cdot\sum_{t=1}^{T}g(y_t,x)\Big]$$

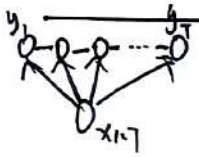
$$\theta=\begin{pmatrix}\lambda\\\gamma\end{pmatrix}_{K+L}\qquad H=\begin{pmatrix}\sum_{t=1}^{T}f\\\sum_{t=1}^{T}g\end{pmatrix}_{K+L}$$

$$\therefore P(Y=y|X=x)=\frac{1}{Z(\theta,H)}\exp[\theta^T\cdot H(y_t,y_{t-1},x)]$$

$P(y|x)$

$$=\frac{1}{Z(\theta,H)}\exp[\langle\theta,H\rangle]$$

16.6. model problems　Given training data $=\{(x^{(i)},y^{(i)})\}_{i=1}^{N}$，$x,y$ 均是 T 维



- Learning：parameter estimation $\hat{\theta}=\arg\max\prod_{i=1}^{N}P(y^{(i)}|x^{(i)})$ ②
- Inference：
  - marginal prob：$p(x_1),\cdots,p(x_n)$ 或 $P(y_t|x)$ ①
  - conditional prob：$P(x|y)$ 生成模型
  - MAP Inference：decoding　$\arg\max_y(Y|x)$ 最大 prob 路径　$\hat{y}=\arg\max_{y=y_1,\cdots y_T}P(y|x)$ ③

16.7 marginal probability

Given $P(Y=y|X=x)$，求 $P(y_t=i|x)$

$$P(y|x)=\frac{1}{Z}\prod_{t=1}^{T}\psi_t(y_{t-1},y_t,x)$$

$y=y_1,\cdots,y_T$

状态 $y_t\in S$，集合则 $O(|S|^T\cdot T)$ 计算大

$$P(y_t=i|x)=\sum_{y_1,y_2\cdots y_{t-1}y_{t+1}\cdots y_T}P(y|x)=\sum_{y_{<t,t-1>}}\sum_{y_{<t+1,T>}}\frac{1}{Z}\prod_{t'=1}^{T}\psi_{t'}(y_{t'-1},y_{t'},x)=\frac{1}{Z}\sum_{y_{<t,t-1>}}\psi_1(y_0,y_1,x)\psi_2(y_1,y_2,x)\cdots\psi_{t-1}(y_{t-2},y_{t-1},x)$$

包括连乘

$$\cdot\psi_t(y_{t-1},y_t=i,x)\cdot\sum_{y_{<t+1,T>}}\psi_{t+1}(y_t=i,y_{t+1},x)\cdots\psi_T(y_{T-1},y_T,x)=\frac{1}{Z}\sum_{y_{t-1}}\psi_t(y_{t-1},y_t=i,x)\sum_{y_{t-2}}\psi_{t-1}(y_{t-2},y_{t-1},x)\cdots$$

$$\Big[\sum_{y_1}\psi_2(y_1,y_2,x)\Big(\sum_{y_0}\psi_1(y_0,y_1,x)\Big)\Big]\sum_{y_T}\cdots\sum_{y_{T-1}}\cdots+\Delta_{右}=\frac{1}{Z}(\Delta_{左}+\Delta_{右})$$

类似这样指数从后往前套求和符号

Δ左 $= \sum_{y_{t-1}} \psi_t(y_{t-1}, y_t=i, x) \sum_{y_{t-2}} \psi_{t-1}(y_{t-2}, y_{t-1}, x) \cdots \sum_{y_1} \psi_2(y_1, y_2, x) \sum_{y_0} \psi_1(y_0, y_1, x)$

求递推：

Δ左 ↘ $\alpha_t(i)$：$y_0\ y_1\ y_2 \cdots y_{t-1}$ $y_t=i$

所有势函数的乘积　　所有势函数

ψ_t 左 ψ_{t+1}

$y_{t-1}\ y_t\ y_{t+1}$

y_{t-2} 0 0 0 0

取 $\psi_t(y_{t-1}, y_t=i^{,x})$ ← 取这个 $y_t=i$

$\psi_{t+1}(y_t=i^{,x}, y_{t+1}^{,x})$

$\alpha_{t-1}(j)$：$y_0\ y_1\ y_2 \cdots y_{t-2}$ $y_{t-1}=j$

所有势函数　　左半部势函数

y_{t-1} 的右半

$\psi_t(y_{t-1}=j, y_t=i, x) \cdot \alpha_{t-1}(j)$ → 所有 + 左半 + 右半 $= \alpha_t(i)$

中间的势函数

左半部：$\psi_{t-1}(y_{t-2}, y_{t-1}=j^{,x})$

右半部 $\psi_t\{y_{t-1}=j, y_t=i^{,x}\}$

y_{t-1} 的右半部 $= y_t$ 的左半部

$0 \to t-2$　y_{t-1}　y_{t-1}

$\alpha_{t-1}(i)$ 0 0 0 0 0

y_{t-2}　y_{t-1}　y_t　y_{t+1}　y_{t+2}

$\alpha_{t-1}(i)$

$\alpha_t(i)$

$\sum_{y_{t-1}=j} \psi_t(y_{t-1}=j, y_t=i, x) \cdot \alpha_{t-1}(j)$

$$\alpha_t(i) = \sum_{j \in S} \psi_t(y_{t-1}=j, y_t=i, x) \cdot \alpha_{t-1}(j) = \alpha_t(i) \rightarrow \text{递推}$$

$\therefore \Delta$左 $= \alpha_t(i)$

Δ右 $= \sum_{<t+1,T>} \psi_{t+1}(y_t=i, y_{t+1}, x) \cdot \psi_{t+2}(y_{t+1}, y_{t+2}, x) \cdots \psi_{T-1}(y_{T-2}, y_{T-1}, x) \cdot \psi_T(y_{T-1}, y_T, x)$

$= \sum_{y_{t+1}} \psi_{t+1}(y_t=i, y_{t+1}, x) \cdot \sum_{y_{t+2}} \psi_{t+2}(y_{t+1}, y_{t+2}, x) \cdots \sum_{y_{T-1}} \psi_{T-1}(y_{T-2}, y_{T-1}, x) \sum_{y_T} \psi_T(y_{T-1}, y_T, x)$

y_t 的右半部

$y_t=i$ 的右半部势函数

求递推：

$\beta_t(i) = y_{t+1}\ y_{t+2} \cdots y_T$

所有势函数

ψ_t ψ_{t+1} ψ_{t+2}

y_{t-1} y_t y_{t+1} y_{t+2} y_T

$\beta_t(i)$ $\beta_{t-1}(i)$ $\beta_{t-1}(j)$

red more clear

即 $\beta_t(i) = y_t=i$ $y_{t+1}\ y_{t+2} \cdots y_T$　从 t 到 T

右半部　所有势函数

$\beta_{t+1}(j) = y_{t+1}=j$ $y_{t+2}\ y_{t+3} \cdots y_T$　从 t+1 到 T

t+1 的左半部　右半部　所有势函数

$\psi_{t+1}(y_t=i, y_{t+1}=j, x) \cdot \beta_{t+1}(j)$

$$\beta_t(i) = \sum_{j \in S} \psi_{t+1}(y_t=i, y_{t+1}=j, x)\ \beta_{t+1}(j)$$

$\therefore \Delta$右 $= \beta_t(i)$

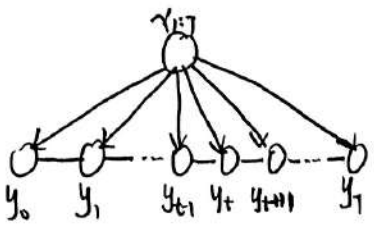
有：$\beta_t(i) =$ y_t 的右半部 + 所有 $(y_{t+1} - y_T)$

$\beta_{t+1}(j) =$ y_{t+1} 的右半部 + 所有 $(y_{t+2} - y_T)$

$\beta_t(j) = \beta_{t+1}(j) +$ y_{t+1} 的左半部

可以用 Inference：变量消除法：Sum-product / Belief propagation

Forward-Backward 也可以理解上述过程



$\alpha_t(i)$ 从左开始，$\beta_t(i)$ 从右开始，从两侧向中间

$\therefore P(y_t=i|x) = \frac{1}{Z} \alpha_t(i) \beta_t(i)$

16.8 Learning

$\hat{\theta} = \arg\max \prod_{i=1}^{N} P(y^{(i)} | x^{(i)})$ N: size of training data

$\hat{\lambda}, \hat{\eta} = \arg\max_{\lambda,\eta} \prod_{i=1}^{N} P(y^{(i)} | x^{(i)})$

$$P(y|x) = \frac{1}{Z(x,\lambda,\eta)} \exp \sum_{t=1}^{T} [\lambda^T \cdot f(y_{t-1}, y_t, x) + \eta^T \cdot g(y_t, x)]$$

$$= \arg\max_{\lambda,\eta} \log \prod_{i=1}^{N} P(y^{(i)} | x^{(i)})$$

$$= \arg\max_{\lambda,\eta} \sum_{i=1}^{N} \log P(y^{(i)} | x^{(i)})$$

$$= \arg\max_{\lambda,\eta} \sum_{i=1}^{N} \left\{ -\log Z(x^{(i)},\lambda,\eta) + \sum_{t=1}^{T} [\lambda^T \cdot \underbrace{f(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)})}_{\text{充分统计量}} + \eta^T \cdot g(y_t^{(i)}, x^{(i)})] \right\}$$

$$\triangleq \arg\max_{\lambda,\eta} L(\lambda, \eta, x^{(i)})$$

方法：梯度上升 $\nabla_\lambda L$, $\nabla_\eta L$

$$\nabla_\lambda L = \sum_{i=1}^{N} [\sum_{t=1}^{T} f(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)}) - \cancel{\nabla_\lambda \log Z(x^{(i)},\lambda,\eta)}]$$

log-Partition Function

$$E_{P(y|x^{(i)})} [\sum_{t=1}^{T} f(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)})]$$

$$= \sum_{y} P(y|x^{(i)}) \cdot \sum_{t=1}^{T} f(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)})$$

$$= \sum_{t=1}^{T} [\sum_{y} P(y|x^{(i)}) \cdot f(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)})]$$

$$= \sum_{t=1}^{T} \sum_{y_{<1,t-2>}} \sum_{y_{t-1}} \sum_{y_t} \sum_{y_{<t+1,T>}} P(y|x^{(i)}) \cdot f(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)})$$

$$= \sum_{t=1}^{T} \sum_{y_{t-1}} \sum_{y_t} \left(\sum_{y_{<1,t-2>}} \sum_{y_{<t+1,T>}} P(y|x^{(i)}) \cdot f(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)}) \right)$$

边缘概率

$$= \sum_{t=1}^{T} \sum_{y_{t-1}} \sum_{y_t} \underbrace{P(y_{t-1}, y_t | x^{(i)})}_{\text{与 P30 计算 } P(y|x) \text{ 的方法雷同}} f(y_{t-1}, y_t, x^{(i)})$$

假设已求出，$P(y_{t-1}, y_t | x^{(i)}) = A(y_{t-1}, y_t)$

$$\nabla_\lambda L = \sum_{i=1}^{N} [\sum_{t=1}^{T} [f(y_{t-1}^{(i)}, y_t^{(i)}, x^{(i)}) - \sum_{y_{t-1}} \sum_{y_t} A(y_{t-1}, y_t) \cdot f(y_{t-1}, y_t, x^{(i)})]]$$

梯度上升算法：

$$\begin{cases} \lambda^{(t+1)} = \lambda^{(t)} + step \cdot \nabla_\lambda L(\lambda^{(t)}, \eta^{(t)}) \\ \eta^{(t+1)} = \eta^{(t)} + step \cdot \nabla_\eta L(\lambda^{(t)}, \eta^{(t)}) \end{cases}$$

$P(x|\eta) = h(x) \exp\{\eta^T \phi(x) - A(\eta)\}$

$A(\eta) = E_{P(x|\eta)}[\phi(x)]$

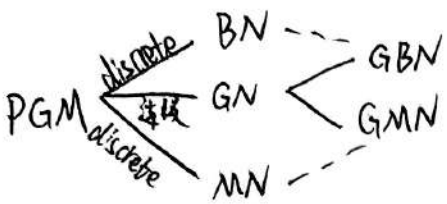
P23. 7.3.

HMM：decoding → Viterbi → 动态规划

CRF：decoding

17. Gaussian Network

17.1 Background

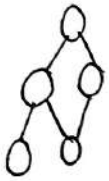


PGM: probabilistic graphical model

BN: bayesian network

MN: markov network

GN: Gaussian network



$x_i \sim N(\mu_i, \Sigma_i)$

$x = (x_1, x_2, \cdots, x_p)^T$

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

$$\Sigma = (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}_{p\times p}$$

$x_i \perp x_j \Longleftrightarrow \sigma_{ij} = 0$

条件独立性: $x_A \perp x_B \mid x_C$

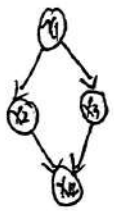
$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1p} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pp} \end{pmatrix}_{p\times p}$$

Λ: precision matrix → information matrix

$x_i \perp x_j \mid \underbrace{-\{x_i, x_j\}}_{\text{除 } x_i, x_j \text{ 外}} \Longleftrightarrow \lambda_{ij} = 0$

17.2 Gaussian Bayesian Network

连续型的PGM. 有向GBN



$x_{pa(4)} = \text{set}\{x_2, x_3\}$

$$P(x) = \prod_{i=1}^{p} P(x_i \mid \underbrace{x_{pa(i)}}_{\text{一个集合(父集)}}) \quad \text{BN的因子分解} \cdots ①$$

GBN is based on linear Gaussian Model

GBN → global model

linear Gaussian Model → local Model →

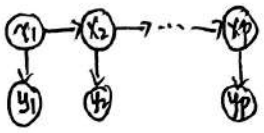
$$\begin{cases} P(x) = N(x \mid \mu_x, \Sigma_x) \\ P(y \mid x) = N(y \mid Ax + b, \Sigma_y) \end{cases}$$

$x \sim N(\mu, \Sigma)$
$x = (x_1, x_2, \cdots, x_p)^T$
$\mu = (\mu_1, \mu_2, \cdots, \mu_p)^T$
$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p)^T$ } ※
$S = \text{diag}(\sigma_i)$
$W = [w_{ij}]$ → 元parents处补0

属于GBN (chain structure)

Kalman Filter (HMM). $\lambda = (\pi, A, B)$

$P(x_t \mid x_{t-1})$ ↑ $P(y_t \mid x_t)$



Review. kalman filter.

$$\begin{cases} P(x_t \mid x_{t-1}) = N(x_t \mid x_{t-1} \cdot A + B, Q) \\ P(y_t \mid x_t) = N(y_t \mid Cx_t + D, R) \end{cases}$$

$$\begin{cases} x_t = Ax_{t-1} + B + \varepsilon, \ \varepsilon \sim N(0, Q) \\ y_t = Cx_t + D + \delta, \ \delta \sim N(0, R) \end{cases}$$

$$P(x) = \prod_{i=1}^{p} P(x_i \mid x_{pa(i)}) \qquad x_{pa(i)} = (x_1, x_2, \cdots, x_k)^T$$

$$P(x_i \mid x_{pa(i)}) = N(x_i \mid \mu_i + W_i^T \cdot x_{pa(i)}, \sigma_i^2)$$

↓ 线性组合

$$x_i = \mu_i + \sum_{j \in x_{pa(i)}} w_{ij} \cdot (x_j - \mu_j) + \sigma_i \cdot \varepsilon_i \quad , \ \varepsilon_i \text{ is r.v} \ , \ \varepsilon_i \sim N(0,1)$$

$$x_i - \mu_i = \sum_{j \in x_{pa(i)}} w_{ij}(x_j - \mu_j) + \sigma_i \cdot \varepsilon_i$$

$x_i \sim N(\mu_i, \sigma_i^2)$

$x_i - \mu_i \sim N(0, \sigma_i^2)$ → 简化模型

P53

根据※式进行化简

$x - \mu = W \cdot (x - \mu) + S \cdot \varepsilon$

$(I - W) \cdot (x - \mu) = S \cdot \varepsilon$

$x - \mu = (I - W)^{-1} \cdot S\varepsilon$

$\Sigma = \text{cov}(x) = \text{cov}(x - \mu) = \text{cov}[(I - W)^{-1} \cdot S\varepsilon]$

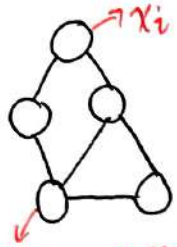
17.3 Gaussian MRF learning：结构参数

$$p(x) = \frac{1}{(2\pi)^{\frac{p}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \cdot \exp\left\{-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1}(x-\mu)\right\} \rightarrow \text{pdf (Gaussian dist)}$$

$$p(x) = \frac{1}{Z} \prod_{i=1}^{p} \underbrace{\psi_i(x_i)}_{\text{node potential}} \prod_{ij \in \varepsilon} \underbrace{\psi_{ij}(x_i, x_j)}_{\text{edge potential}}$$

Λ: precision matrix

$\Lambda\mu$: potention vector $h = \begin{pmatrix} h_1 \\ \vdots \\ h_p \end{pmatrix}$



$(2\pi)^{\frac{p}{2}}$, $|\Sigma|^{\frac{1}{2}}$ 都是 parameters

$\Lambda = \Sigma^{-1}$

$$\therefore p(x) \propto \exp\left\{-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1}(x-\mu)\right\} = \exp\left\{-\frac{1}{2}(x^T\Lambda - \mu^T\Lambda)(x-\mu)\right\}$$

$$= \exp\left\{-\frac{1}{2}(x^T\Lambda x - x^T\Lambda\mu - \mu^T\Lambda x + \mu^T\Lambda\mu)\right\}$$

$$= \exp\left\{-\frac{1}{2}(x^T\Lambda x - 2\mu^T\Lambda x + \mu^T\Lambda\mu)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}x^T\Lambda x + \mu^T\Lambda x\right\} = \exp\left\{\underbrace{-\frac{1}{2}x^T\Lambda x}_{\text{二次项}} + \underbrace{(\Lambda\mu)^T x}_{\text{一次项}}\right\}$$

x_i, x_j 无边，$\lambda_{ij} = 0$

$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}; \Lambda = (\lambda_{ij})_{p \times p}$ → 对称

相同：x_i：$-\frac{1}{2}x_i \cdot \lambda_{ii} + h_i x_i$（二次 / 一次）

把 x_i 看作 node potential

不同：x_i, x_j：$-\frac{1}{2}(\lambda_{ij}x_i x_j + \lambda_{ji}x_j x_i)$，$\Lambda$ 对称矩阵 → 交叉项

x_i, x_j 看作 edge potential

除掉自己二次 $= -\lambda_{ij}x_i x_j$

不然自 $\mu^T\Lambda x$

此处把 Gaussian Dist 与概率图联系

→ 综合：$x_i \perp x_j \mid -\{x_i, x_j\}$，$\lambda_{ij} = 0$

① $x_i \perp x_j$，$(\Sigma = (\sigma_{ij})) \Leftrightarrow \sigma_{ij} = 0$ → marginal independence 全局/绝对独立

② $x_i \perp x_j \mid -\{x_i, x_j\}$ $(\Lambda = \Sigma^{-1} = (\lambda_{ij})) \Leftrightarrow \lambda_{ij} = 0$ → 条件独立

$\lambda_{ij} \neq 0$，图中有边 (x_i, x_j)

$\lambda_{ij} = 0$，图中不有边 (x_i, x_j)

③ $\forall x_i$，$x_i \mid -\{x_i\} \sim N\left(\sum_{j \neq i} \frac{\lambda_{ij}}{\lambda_{ii}} x_j, \lambda_{ii}^{-1}\right)$

斜概率分布

系数

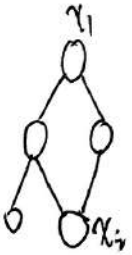
由 P26 全局 Markov property

若两个节点 i 和 j 被集合 S 分离（即 i 和 j 之间所有路径都经过 S）则 $x_i \perp x_j \mid x_S$

由此有：有边时 $x_i \perp x_j \mid -\{x_i, x_j\}$

x_i 可以看作跟它有连接的 x_j 的线性组合

$x = \begin{pmatrix} x_i \\ -\{x_i\} \end{pmatrix} = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$



③补充知识

$\Sigma^{-1} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1}$ 分块矩阵的求逆

$(A - BCD)^{-1}$ → Woodbury formula

Schur complementary P3 1.4

18. Bayesian Linear Regression

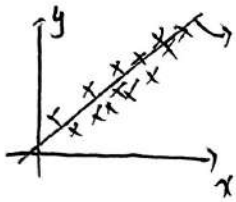
18.1 Background

$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

$X = (x_1, x_2 \cdots x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \cdots x_{1p} \\ x_{21} & x_{22} \cdots x_{2p} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \cdots x_{Np} \end{pmatrix}_{N \times p}$

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



$f(x) = w^T x = x^T w$

$y = f(x) + \varepsilon$ (noise)

x, y, ε are r.v

$x \in \mathbb{R}^p$, $y \in \mathbb{R}$, $\varepsilon \sim N(0, \sigma^2)$

$f(x) = w^T x$, $y = f(x) + \varepsilon = w^T x + \varepsilon$

频率派 (w is unknown constant) → 优化问题 → 点估计

- LSE → Regularized LSE (Lasso, Ridge)
- MLE (noise is Gaussian) ⇑ LSE
- MAP (noise is Gaussian) ⇓ Regularized LSE

$w_{MLE} = \arg\max_w P(Data|w)$

贝叶斯派 (w is r.v) → $P(w|Data)$

$P(w|Data) \propto P(Data|w) \cdot P(w)$

$w_{MAP} = \arg\max_w P(w|Data) = \arg\max_w P(Data|w) P(w)$

$P(w) \sim N(\mu_0, \Sigma_0)$ ↔ Ridge

Laplace ↔ Lasso

PS. (Linear Regression)

18.2 Inference

Data: $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$

$X = (x_1, x_2, \cdots, x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \cdots x_{1p} \\ x_{21} & x_{22} \cdots x_{2p} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \cdots x_{Np} \end{pmatrix}_{N \times p}$

$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1}$, $P(w) \sim N(0, \Sigma_p)$

Model

$\begin{cases} f(x) = w^T x = x^T w \\ y = f(x) + \varepsilon \end{cases}$

$f(x) \sim N(w^T x, 0)$, $\varepsilon + f(x) \sim N(w^T x_i, \sigma^2)$

x, y, ε are r.v

$\varepsilon \sim N(0, \sigma^2)$ 已知

Bayesian Method:

$\begin{cases} \text{Inference}: \text{posterior}(w) \\ \text{Prediction}: x^* \to y^* \end{cases}$

Likelihood - Gaussian

Inference ←

$P(Y|X,w) \sim N(w^T x, \sigma^2)$, $P(w) \sim N(0, \Sigma_p)$

$= \prod_{i=1}^N N(w^T x_i, \sigma^2)$

从 $y_i \to Y$, $x_i \to X$

Likelihood:

$P(Y|X,w) = \prod_{i=1}^N \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right\}$

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

$\sum_{i=1}^N (y_i - w^T x_i)^2 = (y_1 - w^T x_1, y_2 - w^T x_2, \cdots y_N - w^T x_N)$

$= (y_1, y_2 \cdots y_N) - w^T (x_1, x_2 \cdots x_N)$

$\begin{pmatrix} y_1 - w^T x_1 \\ \vdots \\ y_N - w^T x_N \end{pmatrix}$

Inference: $P(w|Data) = \frac{P(w, Y|X)}{P(Y|X)} = \frac{P(Y|w,X)P(w|X)}{\int P(Y|w,x)P(w|x)dw}$

$= P(w|X,Y)$

$w \perp X$

$\because P(w|X) = P(w)$ (likelihood, prior)

$\therefore P(w|X,Y) = \frac{P(Y|w,X)P(w)}{\int P(Y|w,x)P(w)dw} \propto P(Y|w,X) \cdot P(w)$

$P(Y|w,X) = \prod_{i=1}^N P(y_i|w, x_i) = \prod_{i=1}^N N(y_i|w^T x_i, \sigma^2)$

$P(Y|X,w) \sim N(Xw, \sigma^2), P(w) \sim N(0, \sigma^2\Sigma_p)$

$P(w|Data) \propto P(Y|w,X) \cdot P(w)$

Gaussian, Gaussian, Gaussian

两个 Gaussian 相乘，根据共轭性质后验也是 Gaussian

共轭：Gaussian 分布是自共轭的 ☆

likelihood + prior → Gaussian ⇒ Posterior 也为 Gaussian

assume: $P(w|Data) \sim N(\mu_w, \Sigma_w)$

$P(w|Data) \propto \prod_{i=1}^N N(y_i|w^T x_i, \sigma^2) \cdot N(0, \Sigma_p)$

$P(w|Data) \sim N(\mu_w, \Sigma_w)$

求：μ_w, Σ_w 两个方法，先算出 $P(Y|X,w)$

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

$P(Y|X,w) = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N} \exp\left\{-\frac{1}{2}(Y - Xw)^T \cdot \sigma^{-2} I \cdot (Y - Xw)\right\}$

$\sim N(Xw, \sigma^2 I)$

$\Sigma^{-1}, \Sigma = \sigma^2$

$P(w)$

故 $P(w|Data) \propto N(Xw, \sigma^2 I) \cdot N(0, \Sigma_p)$

$\propto \exp\left\{-\frac{1}{2}(Y - Xw)^T \sigma^{-2} I (Y - Xw)\right\} \cdot \exp\left\{-\frac{1}{2} w^T \Sigma_p^{-1} w\right\}$

$= \exp\left\{-\frac{1}{2\sigma^2}(Y^T - w^T X^T)(Y - Xw) - \frac{1}{2} w^T \Sigma_p^{-1} w\right\}$

$= \exp\left\{-\frac{1}{2\sigma^2}(Y^T Y - 2Y^T X w + w^T X^T X w) - \frac{1}{2} w^T \Sigma_p^{-1} w\right\}$

由一个 $P(x)=N(\mu,\Sigma)$，变量为 X　　这是讨论 $P(x)=N(\mu,\Sigma)$ 的一般情况

指数部分：$\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$

$=\exp\left\{-\frac{1}{2}(x^T\Sigma^{-1}-\mu^T\Sigma^{-1})(x-\mu)\right\}$

☆ $=\exp\left\{-\frac{1}{2}[\underbrace{x^T\Sigma^{-1}x}_{二次项}-\underbrace{2\mu\Sigma^{-1}x}_{一次项}+\mu^T\Sigma^{-1}\mu]\right\}$，一次项系数"-2"与外部"$-\frac{1}{2}$"可约掉 $=\mu^T\Sigma^{-1}x$

此是要

$P(w|Data)$ 的二次项（w的）：$-\frac{1}{2\sigma^2}w^Tx^Txw-\frac{1}{2}w^T\Sigma_p^{-1}w=-\frac{1}{2}[w^T(\sigma^{-2}x^Tx+\Sigma_p^{-1})w]$，变量为 w

$\Sigma_w^{-1}=A$ → precision metrix

$P(w|Data)$ 的一次项：$-\frac{1}{2\sigma^2}\cdot(-2Y^Txw)=\underbrace{\sigma^{-2}Y^Tx}_{\mu_w^TA=\mu_w^T\Sigma_w^{-1}=\sigma^{-2}Y^Tx}w$

$A^T=A$（precision matrix）

$\therefore \mu_w^TA=\sigma^{-2}Y^Tx$

$A\mu_w=\sigma^{-2}x^TY$

$\mu_w=\sigma^{-2}A^{-1}x^TY$

故 $P(w|Data)\sim N(\sigma^{-2}A^{-1}x^TY,A^{-1})$ 且 $A=\sigma^{-2}x^Tx+\Sigma_p^{-1}$

18.3 Prediction.

Given x^*, get y^*.

Model：$\begin{cases}f(x)=w^Tx=x^Tw\\ y=f(x)+\varepsilon\end{cases}$　$\varepsilon\sim N(0,\sigma^2)$

① $f(x^*)$：noise-free

$f(x)=x^Tw$

$f(x^*)=x^{*T}w$　$P(w|Data)$ ← 已知 Data 与 w，做 Prediction

$P(f(x^*))=N(\mu_w,\Sigma_w)$

$w\sim N(\mu_w,\Sigma_w)$

$x^{*T}w\sim N(x^{*T}\mu_w,x^{*T}\Sigma_wx^*)$，$\Sigma_{w\,p\times p}$，$x^*_{p\times 1}$

$\therefore P(f(x^*)|Data,x^*)=N(x^{*T}\mu_w,x^{*T}\Sigma_wx^*)$

② y^*：$y^*=f(x^*)+\varepsilon$，noise

$\therefore P(y^*|Data,x^*)=N(x^{*T}\mu_w,x^{*T}\Sigma_wx^*+\sigma^2)$

$\begin{cases}\mu_w\\ \Sigma_w\end{cases}$ 参考上节 $\begin{cases}\mu_w=\sigma^{-2}A^{-1}x^TY\\ \Sigma_w=A^{-1}\end{cases}$，$A=\sigma^{-2}x^Tx+\Sigma_p^{-1}$

总结：先求 Inference 求出 Σ_w 与 μ_w，再利用它们做 Prediction

19.1 Background

19. Gaussian distribution Process 随机过程

Gaussian Dist ——> Multivariate GaussianDist ——> 无限维 Gaussian Dist <=> Gaussian Process

Gaussian Dist：一维，$P(x)=N(\mu,\sigma^2)$

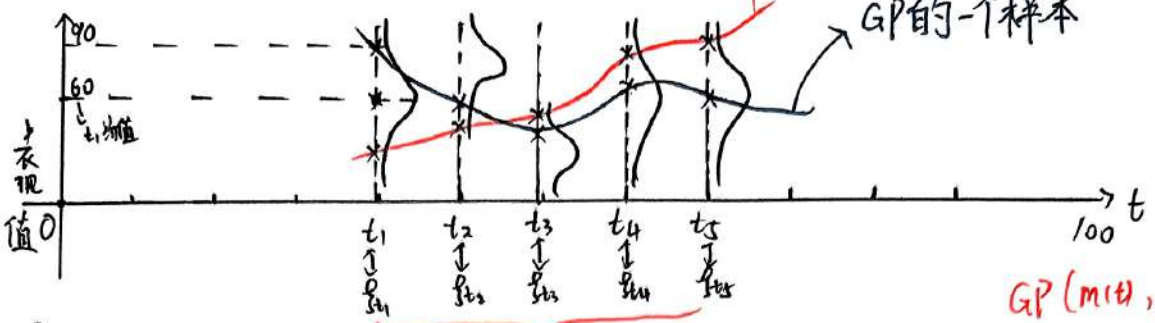
⇕

Gaussian Network：$x\in\mathbb{R}^p$，$P(x)=N(\mu,\Sigma)$，$\Sigma_{p\times p}$，$p<+\infty$（可数有限维）

Process：无限维

$\{\xi_t\}_{t\in T}$，$T\to$连续域，if $\forall n\in\mathbb{N}^+, t_1,t_2,\cdots,t_n\in T$，s.t. $\{\xi_{t_1},\xi_{t_2},\cdots,\xi_{t_n}\}\triangleq \xi_{t_1\sim t_n}\sim N(\mu_{t_1\sim t_n}, \Sigma_{t_1\sim t_n})$

那么$\{\xi_t\}_{t\in T}$就是一个Gaussian progess



人的一生：$[0,100]$

$t\in[0,100]$，$\xi_t\sim N(\mu_t,\sigma_t^2)$ ← 这似在t时刻他的表现(0-100)

$\xi_{t_1\sim t_5}$各自为Gaussian Dist 并且$\xi_{t_1\sim t_5}$联合起来也是，$N(\mu,\Sigma)$

$t=0$，他的一生基本定，$t>0$，每一时刻，μ_t,σ_t^2是已经右确定

$GP(m(t),k(s,t))$

$m(t)=E[\xi_t]$

$k(s,t)=E[(\xi_s-m(s))(\xi_t-m(t))]$

$m(t)\to$ mean function

$k(s,t)\to$ covariance function / kernel function

19.2 weight space perspective　非线性Bayesian Rregression / Gauss process Regression

Recall Baysian Linear Regression

Model: $\begin{cases} f(x)=w^Tx=x^Tw \\ y=f(x)+\varepsilon \quad [\varepsilon\sim N(0,\sigma^2)] \end{cases}$

① $P(w|Data)=N(w|\mu_w,\Sigma_w)$

$\begin{cases}\mu_w=\sigma^{-2}A^{-1}X^TY \\ \Sigma_w=A^{-1}\end{cases}$ $(A=\sigma^{-2}X^TX+\Sigma_p^{-1})$

② Given x^*, Prediction

1) noise-free

$P(f(x^*)|Data,x^*)=N(x^{*T}\mu_w, x^{*T}\Sigma_w x^*)$

2) noise

$P(y^*|Data,x^*)=N(x^{*T}\mu_w, x^{*T}\Sigma_w x^*+\sigma^2)$

接下来介绍kernel Trick:

若 $y=f(x)+\varepsilon$，$f(x)$ is not Linear function

对于Non-linear问题：① Non-linear Transformation ② Bayesian LR → Kernel → kernel BLR

△Pts结论，可以Kernel Trick，通解

↓求出$\phi(x)$，直接定义 kernel function $=k(x,x')$

x,w,ε均升到q维.

$p(w)\sim N(0,\Sigma_q)$

$\varepsilon\sim N(0,\sigma^2)$，此$\sigma$为q维.

Noise-free:

$X=(x_1,x_2\cdots x_N)^T$，$Y=(y_1,y_2\cdots y_N)^T$

$f(x^*)|X,Y,x^*\sim N(x^{*T}(\sigma^{-2}A^{-1}X^TY), x^{*T}A^{-1}x^*)$，$A=\sigma^{-2}X^TX+\Sigma_p^{-1}$ （$p\times p$）

If $\phi: x\longmapsto z$，$x\in\mathbb{R}^p$，$z=\phi(x)\in\mathbb{R}^q$，$q>p$，$p(w)\sim N(0,\Sigma_p)$；$\phi(w)=w'\sim N(0,\Sigma_q)$（$q\times q$），noise升维.

Define: $\Phi=\phi(x)=(\phi(x_1)\ \phi(x_2)\cdots\phi(x_N))^T_{N\times q}$

此处ε也为q维 记$\varepsilon\sim(0,\sigma^2)$

但此σ为$q\times q$ 之前为$p\times p$

Then: Model: $f(x)=\phi(x)^Tw'$，w变为q维

故有：$f(x^*)|X,Y,x^*\sim N(\sigma^{-2}\phi(x^*)^TA^{-1}\Phi^TY,\ \phi(x^*)^TA^{-1}\phi(x^*))$，$A=\sigma^{-2}\Phi^T\Phi+\Sigma_q^{-1}$，$X$与$x^*$有大区别

How to compute A^{-1}? → woodbury formula: $(A+UCV)^{-1}=A^{-1}-A^{-1}U(C^{-1}+VA^{-1}U)^{-1}VA^{-1}$

后面Proof this is kernel function，可用 kernel trick

在不计算A^{-1}的情况下，求均值

$A = \sigma^{-2}\Phi^T\Phi + \Sigma_q^{-1}$

$A\Sigma_q = \sigma^{-2}\Phi^T\Phi\Sigma_q + I$

$A\Sigma_q\Phi^T = \sigma^{-2}\Phi^T\Phi\Sigma_q\Phi^T + \Phi^T$

$= \sigma^{-2}\Phi^T(K + \sigma^2 I)$, $K = \Phi\Sigma_q\Phi^T$

$\Sigma_q\Phi^T = \sigma^{-2}A^{-1}\Phi^T(K+\sigma^2 I)$

$\sigma^{-2}A^{-1}\Phi^T = \Sigma_q\Phi^T(K+\sigma^2 I)^{-1}$

$\sigma^{-2}\phi(x^*)^T A^{-1}\Phi^T Y = \phi(x^*)\Sigma_q\Phi^T(K+\sigma^2 I)^{-1}Y$ → 避开A^{-1}

$f(x^*)|X,Y,x^*$'s expectation 均值

$f(x^*)|X,Y,x^* \sim N(\mu^*, \Sigma^*)$

$\mu^* = \phi(x^*)\Sigma_q\Phi^T(K+\sigma^2 I)^{-1}Y$

$\Sigma^* = \phi(x^*)^T\Sigma_q\phi(x^*) - \phi(x^*)^T\Sigma_q\Phi^T(K+\sigma^2 I)^{-1}\Phi\Sigma_q\phi(x^*)$ $K = \Phi\Sigma_q\Phi^T$

Likewise: $f(x^*)|X,Y,x^*$'s covariance: $\phi(x^*)^T\Sigma_q\phi(x^*) - \phi(x^*)^T\Sigma_q\Phi^T(K+\sigma^2 I)^{-1}\Phi\Sigma_q\phi(x^*)$

$\therefore f(x^*)|X,Y,x^* \sim N\left(\underbrace{\phi(x^*)\Sigma_q\Phi^T}_{A}(K+\sigma^2 I)^{-1}Y,\ \phi(x^*)^T\Sigma_q\phi(x^*) - \underbrace{\phi(x^*)^T\Sigma_q\Phi^T}_{A}(K+\sigma^2 I)^{-1}\Phi\Sigma_q\phi(x^*)\right)$ ☆

$K = \Phi\Sigma_q\Phi^T$, $\phi(x^*)\Sigma_q\Phi^T$, $\phi(x^*)^T\Sigma_q\phi(x^*)$, $\phi(x^*)^T\Sigma_q\Phi^T$, $\Phi\Sigma_q\phi(x^*)$ 形式类似

并且 $\Phi = \phi(X) = (\phi(x_1), \phi(x_2) \cdots \phi(x_N))^T$ 展开后相乘仍为 $\phi^T(x)\Sigma_q\phi(x')$ 这种形式

Φ: $N \times q$

X: $N \times p$

判断：$k(x,x') = \phi^T(x)\Sigma_q\phi(x') \overset{?}{\longleftrightarrow}$ kernel function 参考P22. Definition

$\because \Sigma_q$: positive definite，并且对称，$\Sigma_q = (\Sigma_q^{\frac{1}{2}})^2$

$\therefore k(x,x') = \phi^T(x)\Sigma_q^{\frac{1}{2}}\Sigma_q^{\frac{1}{2}}\phi(x') = (\Sigma_q^{\frac{1}{2}}\phi(x))^T\cdot\Sigma_q^{\frac{1}{2}}\phi(x') = \underline{\langle\psi(x), \psi(x')\rangle}$ 内积

$\psi(x) = \Sigma_q^{\frac{1}{2}}\phi(x)$

$\therefore$ 令 $k(x,x') = \phi^T(x)\Sigma_q\phi(x')$ is kernel function

即说明不用找到具体的$\phi(x)$，运用 kernel Trick 定义 kernel function 即可

Gaussian process Regression = kernel trick (Non-linear Transformation inner product) + Bayesian Linear Regression

Two views:

① weight-space view: $f(x) = \phi^T(x)w$, w → weight; $y = f(x) + \varepsilon$

② function-space view

① 和 ② equal result

关注 $f(x)$

$f(x)$ is r.v, $f(x) \sim GP(m(x), k(x,x'))$

Gauss Process Regression is the extension of Bayesian Linear Regression with kernel Trick.

19.3 from weight space to function space

Recall Gaussian Process:

$\{\xi_t\}_{t\in T}$, T: continuous time/space. $\forall n\in N^+(n\geq 1)$, $t_1, t_2, \cdots, t_n \to \xi_1, \xi_2, \cdots, \xi_n$ (index process; $\xi_1, \xi_2, \cdots, \xi_n$ r.v), 令 $\xi_{1:n}=(\xi_1, \xi_2 \cdots \xi_n)^T$,

If $\xi_{1:n} \sim N(\mu_{1:n}, \Sigma_{1:n})$, Then $\{\xi_t\}_{t\in T}$ is Gaussian process. $\xi_t \sim GP(m(t), k(t,s))$

mean function, covariance function, kernel function

为什么GP由m(t), k(t,s)来表达 (高斯过程存在性定理)

回到weight-space view (关注对象为w)

$\begin{cases} f(x)=\phi(x)^T w \\ y=f(x)+\varepsilon,\ \varepsilon \sim N(0,\sigma^2) \end{cases}$

Bayesian Method:

给定 prior: $w\sim N(0,\Sigma_q)$ 此处w已做从p→q升维

$\because f(x)=\phi(x)^T w$

$\therefore E[f(x)] = E[\phi(x)^T w] = \phi(x)^T E[w] = 0$

$\forall x, x' \in \mathbb{R}^q$

$cov(f(x), f(x')) = E[(f(x)-E[f(x)])(f(x')-E[f(x')])] = E[f(x)f(x')] = E[\phi^T(x)w \cdot \underbrace{\phi^T(x')w}_{\text{转置}}] = E[\phi^T(x)ww^T\phi(x')]$

$= \phi^T(x) E[w\cdot w^T]\phi(x') = \phi^T(x) E[(w-0)(w^T-0)]\phi(x')$

$\because w\sim N(0,\Sigma_q)$

$\therefore cov(f(x), f(x')) = \underbrace{\phi^T(x)\Sigma_q\phi(x')}_{\text{kernel function, } \psi(x)=\Sigma_q^{\frac{1}{2}}\phi(x)} = \langle \psi(x), \psi(x')\rangle = k(x,x')$

for GPR

↑ process

Gaussian function:

$f(x) \sim GP(m(x), k(x,x'))$

mean function, covariance function

① $f(x)$ is function

② $f(x)$ is r.v Gaussian Dist

$t \to \xi_t$, $\{\xi_t\}_{t\in T} \sim GP$

$x \to f(x)$, $\{f(x)\}_{x\in\mathbb{R}^q} \sim GP$

$T=\mathbb{R}^q$ 稠密的

$t \leftrightarrow x$

$\xi_t \leftrightarrow f(x)$

启发: $f(x)$是否可看作一个高斯过程

$\{f(x)\}_{x\in\mathbb{R}^q}$ 一族 就是说，这里推出来的 $E[f(x)]$ 与 $cov(f(x),f(x'))$ 与P5对Gaussian Process的定义有相同处，同为kernel function

GPR:

① weight-space view: 关注的是w ② function-space view 关注的是 f(x)

$x^* \to y^*$

$P(y^*|Data, x^*) = \int_w P(y^*|w, x^*)\cdot P(w)dw$ $P(y^*|Data, x^*) = \int_f P(y^*|f, x^*)\cdot P(f)df$

19.4 function space perspective

$\{f(x)\}_{x\in\mathbb{R}^q} \sim GP(m(x), k(x,x'))$ 已知的GP, m(x), k(x,x') 均已知

$m(x) = E[f(x)]$, $k(x,x') = E[(f(x)-m(x))(f(x')-m(x'))^T]$

Regression:

Data: $\{(x_i, y_i)\}_{i=1}^N$, $y=f(x)+\varepsilon$, $\varepsilon\sim N(0,\sigma^2)$

$X=(x_1, x_2 \cdots x_N)^T_{N\times p}$

$Y=(y_1, y_2 \cdots y_N)^T_{N\times 1}$

$f(x) \sim N(\mu(x), k(X,X))$

$Y = f(x)+\varepsilon \sim N(\mu(x), k(x,x)+\sigma^2 I)$

$f(x)$ is normal Dist

Prediction: Given $X^* = (x_1^* \ x_2^* \cdots x_m^*)$

$Y^* = f(X^*)+\varepsilon$

$\begin{pmatrix} Y \\ f(X^*) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(x) \\ \mu(X^*) \end{pmatrix}, \underbrace{\begin{pmatrix} k(X,X)+\sigma^2 I & k(X,X^*) \\ k(X^*,X) & k(X^*,X^*) \end{pmatrix}}_{\text{分四个块}}\right)$

$\begin{pmatrix} cov(Y,Y) & cov(Y, f(X^*)) \\ cov(f(X^*), Y) & cov(f(X^*), f(X^*)) \end{pmatrix}$

$cov(Y, f(X^*)) = cov(f(x), f(X^*)) = k(x, X^*)$

已知联合高斯分布，求条件概率

$$\begin{pmatrix} Y \\ f(X^*) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(X) \\ \mu(X^*) \end{pmatrix}, \begin{pmatrix} k(X,X)+\sigma^2 I & k(X,X^*) \\ k(X^*,X) & k(X^*,X^*) \end{pmatrix} \right)$$

求 $P(f(X^*) \mid Y, X, X^*)$ → 条件概率 $\sim N(\mu^*, \Sigma^*)$

（X^* 含在 X 中）

公式：$\chi \sim N(\mu, \Sigma)$

$$\chi = \begin{pmatrix} \chi_a \\ \chi_b \end{pmatrix}^p_q ; \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}; \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$$\chi_b \mid \chi_a \sim N(\mu_{b|a}, \Sigma_{b|a})$$

$$\mu_{b|a} = \Sigma_{ba}\Sigma_{aa}^{-1}(\chi_a - \mu_a) + \mu_b$$

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

$$\mu^* = k(X^*, X)\cdot[k(X,X)+\sigma^2 I]^{-1}(Y - \mu(X)) + \mu(X^*)$$

$$\Sigma^* = k(X^*,X^*) - k(X^*,X)[k(X,X)+\sigma^2 I]^{-1}\cdot k(X,X^*)$$

① noise free:

$$P(f(X^*) \mid Y, X, X^*) = N(\mu^*, \Sigma^*)$$

② noise ($\varepsilon \sim N(0, \sigma^2)$)

$$P(f(X^*) \mid Y, X, X^*) = N(\mu^*, \Sigma^* + \sigma^2 I)$$

function space 与 weight space 结果一样，但更加简单

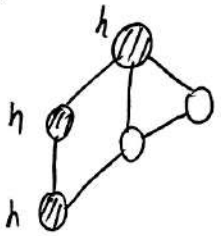
一开始就认为 $f(x) \sim GP(m(x), k(x,x'))$ → 求出 μ^*, Σ^* 才证明 $\sim GP(m(x), k(x,x'))$

---

20. Restricted Boltzmann Machine

20.1 Background

Boltzmann machine：Markov Random field with hidden nodes



nodes → r.v
- observed variable：v
- hidden variable：h

因子分解：(Hammersly clifford Theorem) 基于最大团（C：最大团，$\psi_i(\chi_{ci})$：势函数 potential function，Z：归一化因子，partition function）

$$P(x) = \frac{1}{Z}\prod_{i=1}^{K}\psi_i(\chi_{ci}) \quad \text{s.t. } \psi_i \text{ 严格大于 } 0,\ Z = \sum_{\text{all}}\prod_{i=1}^{K}\psi_i(\chi_{ci}) = \sum_{x_1}\sum_{x_2}\cdots\sum_{x_p}\prod_{i=1}^{K}\psi_i(\chi_{ci})$$

$$\text{令 } \psi_i(\chi_{ci}) = \exp\{-E(\chi_{ci})\} \quad E: \text{energy function}$$

$$\therefore P(x) = \frac{1}{Z}\prod_{i=1}^{K}\psi_i(\chi_{ci}) = \underbrace{\frac{1}{Z}\exp\left\{-\sum_{i=1}^{K}E(\chi_{ci})\right\}}_{\text{指数型分布族}} \Rightarrow \text{简化 } \underbrace{P(x) = \frac{1}{Z}\exp\{-E(x)\}}_{\text{Boltzmann Dist (Gibbs Dist)}} \rightarrow \text{pdf}$$

history：Boltzmann Dist：统计物理：一个物理系统 → 粒子/原子/分子

物理解释
- 每一个粒子的 v 会引起他
- 粒子的影响（碰撞）
- 即 E（energy function）与粒子相关

系统状态：state 表达 → energy

$$P(\text{state}) \propto \exp\left\{-\frac{E}{k\cdot T}\right\}$$

k：Boltzmann constant；T：temperature

假设一共有 M 个状态

s	1	2	…	i	…	M
p	p_1	p_2	…	p_i	…	p_M

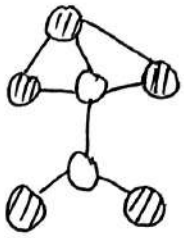
E 和 P(state) 成反比

能量越大，越能够挣脱束缚，不稳定，越可能从当前状态变为其它状态，会向稳态发展

20.2 Representation

$X=\{x_1,x_2,\cdots,x_p\}=\{h,v\}$ → Boltzmann 把 latent 与 observed 分开

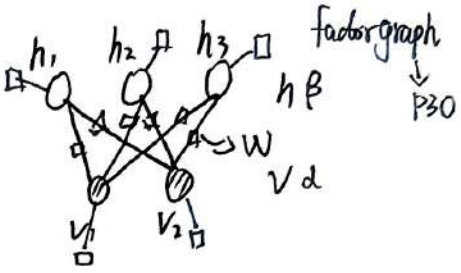
$h=\{h_1,h_2,\cdots,h_m\}$; $v=\{v_1,v_2,\cdots,v_n\}$ $m+n=p$



Boltzmann Machine：问题：Inference {精确 → untractable / 近似 → 计算量太大

↓简化

Restricted Boltzmann Machine：(h,v之间有连接，h,v内部无连接)



⊗ $P(x)=\frac{1}{Z}\exp\{-E(x)\} \Leftrightarrow P(v,h)=\frac{1}{Z}\exp\{-E(v,h)\}$

$E(v,h)=-(h^TWv+\alpha^Tv+\beta^Th)$

$P(x)=P(v,h)=\frac{1}{Z}\exp\{h^TWv+\alpha^Tv+\beta^Th\}=\frac{1}{Z}\exp\{h^TWv\}\cdot\exp\{\alpha^Tv\}\cdot\exp\{\beta^Th\}$

（边 / 点）

factor graph view

$\exp\{h^TWv\}=\exp\{\sum_{i=1}^{m}\sum_{j=1}^{n}h_iw_{ij}v_j\}$

edge / node v

$\exp\{\alpha^Tv\}=\prod_{j=1}^{n}\exp\{\alpha_jv_j\}$

$\exp\{\beta^Th\}=\prod_{i=1}^{m}\exp\{\beta_ih_i\}$

factor / node h

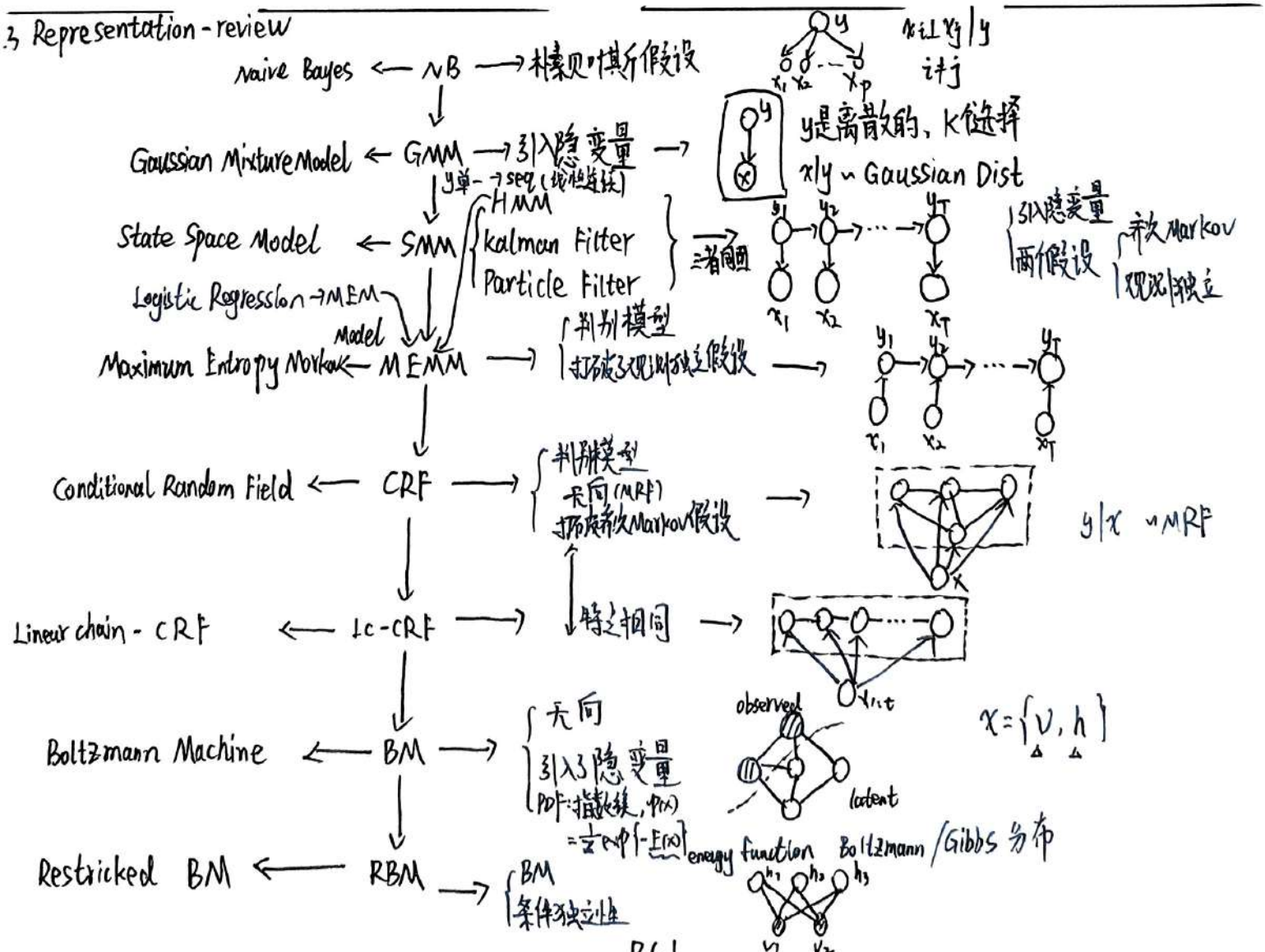
↓拆为vector：

$x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_p\end{pmatrix}=\begin{pmatrix}h\\v\end{pmatrix}$; $h=\begin{pmatrix}h_1\\h_2\\\vdots\\h_m\end{pmatrix}$; $v=\begin{pmatrix}v_1\\v_2\\\vdots\\v_n\end{pmatrix}$ $p=m+n$

∴ RBM's pdf → $P(x)=P(v,h)=\frac{1}{Z}\exp\{h^TWv\}\cdot\exp\{\alpha^Tv\}\cdot\exp\{\beta^Th\}=\frac{1}{Z}\prod_{i=1}^{m}\prod_{j=1}^{n}\exp\{h_iw_{ij}v_j\}\prod_{j=1}^{n}\exp\{\alpha_jv_j\}\cdot\prod_{i=1}^{m}\exp\{\beta_ih_i\}$

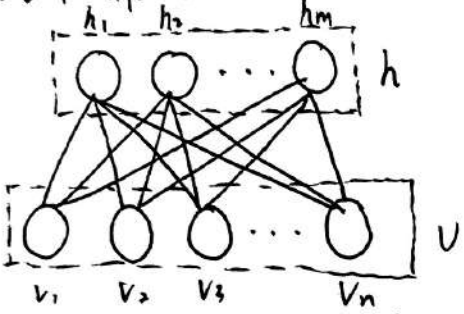
$W_{m\times n}$

20.3 Representation - review



概率图模型 {
- 方向(有向/无向/混合)→边
- 离散/连续/混合 →点
- 条件独立性 →边
- 隐变量 →点
- 指数族分布 →结构
}

20.4 Inference - Posterior Distribution



local Markov. - - - - - -

$P(v|h)=\prod_{i=1}^{n}P(v_i|h)$

目的：Inference → posterior → $P(h|v)$、$P(v|h)$

求 $P(h|v)$　　v：neighbour (h_i)

$h_{-i}=\{h_i\}_{i\neq i}$

$P(h|v)=\prod_{i=1}^{m}P(h_i|v)$

这里假设 h_i：binary RBF　h_i 只有 $\{0,1\}$ 可选 简单例子

$P(h_i=1|v)=P(h_i=1|h_{-i},v)=\frac{P(h_i=1,h_{-i},v)}{P(h_{-i},v)}$

$=\frac{P(h_i=1,h_{-i},v)}{P(h_i=1,h_{-i},v)+P(h_i=0,h_{-i},v)}$ ①　$P(h_i,h_{-i},v)=P(h_{-i},v)$

$P(x)=\frac{1}{Z}\exp\{-E(x)\}$

$P(h,v)=\frac{1}{Z}\exp\{-E(h,v)\}$ 边点点

$E(h,v)=-(h^TWv+a^Tv+\beta^Th)=-(\sum_{i=1}^{m}\sum_{j=1}^{n}h_iw_{ij}v_j+\sum_{j=1}^{n}a_jv_j+\sum_{i=1}^{m}\beta_ih_i)$

P6 具体推导 $\rightarrow=\frac{1}{Z}\exp\{h^TWv+a^Tv+\beta^Th\}$

$E(h,v)=-(\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n}h_iw_{ij}v_j}_{\Delta_1}+\underbrace{h_i\sum_{j=1}^{n}w_{ij}v_j}_{\Delta_2}+\underbrace{\sum_{j=1}^{n}a_jv_j}_{\Delta_3}+\underbrace{\sum_{\substack{i=1\\i\neq i}}^{n}\beta_ih_i}_{\Delta_4}+\underbrace{\beta_ih_i}_{\Delta_5})$

令 $\bar{H}_i(h_{-i},v)=\Delta_1+\Delta_3+\Delta_4$

$\Delta_2+\Delta_5=h_i(\sum_{j=1}^{n}w_{ij}v_j+\beta_i)=h_i\cdot H_i(v)$

$\Delta_1+\Delta_2+\Delta_3+\Delta_4+\Delta_5=h_i\cdot H_i(v)+\bar{H}_i(h_{-i},v)=-E(h,v)$

计算 $P(h_i,h_{-i},v)$ 自然通项，把 $h_i=1.0$ 代入即可

$\therefore E(h,v)=-[h_i\cdot H_i(v)+\bar{H}_i(h_{-i},v)]=-E(h,v)$

$\therefore$ 分子 $=P(h_i=1,h_{-i},v)=\frac{1}{Z}\exp\{\cdot h_i\cdot H_i(v)+\bar{H}_i(h_{-i},v)\}\overset{h_i=1}{=}\frac{1}{Z}\exp\{\cdot H_i(v)+\bar{H}_i(h_{-i},v)\}$

① 式

分母 $=\frac{1}{Z}\exp\{\cdot H_i(v)+\bar{H}_i(h_{-i},v)\}+\frac{1}{Z}\exp\{\cdot\bar{H}_i(h_{-i},v)\}$

$\therefore P(h_i=1|v)=\frac{1}{1+\exp\{-H_i(v)\}}=\sigma(H_i(v))=\sigma[(\cdot\sum_{j=1}^{n}w_{ij}v_j+\cdot\beta_i)]$ ☆　Sigmoid function

$\sigma\rightarrow$ Sigmoid. function $\rightarrow\sigma(x)=\frac{1}{1+e^{-x}}$

Likewise, $P(h_i=0|v)=\frac{1}{1+\exp\{H_i(v)\}}=\sigma(-H_i(v))=\sigma[-(\sum_{j=1}^{n}w_{ij}v_j+\beta_i)]$ ☆

分子 $=P(h_i=0,h_{-i},v)=\frac{1}{Z}\exp\{h_i\cdot H_i(v)+\bar{H}_i(h_{-i},v)\}=\frac{1}{Z}\exp\{\bar{H}_i(h_{-i},v)\}$

分母 $=P(h_i=0,h_{-i},v)+P(h_i=1,h_{-i},v)=\frac{1}{Z}\exp\{\bar{H}_i(h_{-i},v)\}+\frac{1}{Z}\{H_i(v)+\bar{H}_i(h_{-i},v)\}$

$P(v|h)$ 与上述过程类似同理

可以求闭解 $P(h_i|v)$

$x=\begin{pmatrix}x_1\\x_2\\\vdots\\x_p\end{pmatrix}=\begin{pmatrix}h\\v\end{pmatrix}$；$h=\begin{pmatrix}h_1\\h_2\\\vdots\\h_m\end{pmatrix}$；$v=\begin{pmatrix}v_1\\v_2\\\vdots\\v_n\end{pmatrix}$　$p=m+n$

RBM—神经网络 sigmoid → $\sigma(x)=\frac{1}{1+e^{-x}}$

20.5. Inference - Marginal Distribution

目的：Inference → marginal → $P(v)$

$h=\begin{pmatrix}h_1\\h_2\\\vdots\\h_m\end{pmatrix}$ $h_i\in\{0,1\}$ ↗行向量

$W=[w_{ij}]_{m\times n}=\begin{pmatrix}-w_1-\\-w_2-\\\vdots\\-w_m-\end{pmatrix}$

$$P(v)=\sum_h P(h,v)=\sum_h \frac{1}{Z}\exp\{-E(h,v)\}=\sum_h \frac{1}{Z}\exp\{h^TWv+\alpha^Tv+\beta^Th\}$$

$$=\sum_{h_1}\cdots\sum_{h_m}\frac{1}{Z}\exp\{\underbrace{h^TWv}_{\Delta}+\underbrace{\alpha^Tv}+\underbrace{\beta^Th}_{\Delta}\}=\frac{1}{Z}\exp\{\alpha^Tv\}\cdot\sum_{h_1}\cdots\sum_{h_m}\exp\{h^TWv+\beta^Th\}$$

$\sum_{i=1}^m(h_iW_iv+\beta_ih_i)$ $(h_1h_2\cdots h_m)\begin{pmatrix}-w_1-\\-w_2-\\\vdots\\-w_m-\end{pmatrix}v$ $(\beta_1,\cdots,\beta_m)\begin{pmatrix}h_1\\\vdots\\h_m\end{pmatrix}$

$$=\frac{1}{Z}\exp\{\alpha^Tv\}\cdot\sum_{h_1}\cdots\sum_{h_m}\exp\{\sum_{i=1}^m(h_iW_iv+\beta_ih_i)\}$$

$$=\frac{1}{Z}\exp\{\alpha^Tv\}\cdot\sum_{h_1}\cdots\sum_{h_m}\exp\{h_1W_1v+\beta_1h_1\}\cdot\exp\{h_2W_2v+\beta_2h_2\}\cdots\cdot\exp\{h_mW_mv+\beta_mh_m\}$$

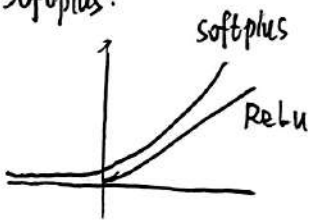
$$=\frac{1}{Z}\exp\{\alpha^Tv\}\sum_{h_1}\exp\{h_1W_1v+\beta_1h_1\}\cdot\cdots\cdot\sum_{h_m}\exp\{h_mW_mv+\beta_mh_m\}\quad\because h_i\in\{0,1\}$$

$$=\frac{1}{Z}\exp\{\alpha^Tv\}(1+\exp\{W_1v+\beta_1\})\cdot\cdots\cdot(1+\exp\{W_mv+\beta_m\})$$

$$=\frac{1}{Z}\exp\{\alpha^Tv\}\cdot\exp\{\log(1+\exp\{W_1v+\beta_1\})\cdot\cdots\cdot(1+\exp\{W_mv+\beta_m\})\}$$

$$=\frac{1}{Z}\exp\{\alpha^Tv+\sum_{i=1}^m\underbrace{[\log(1+\exp\{W_iv+\beta_i\})]}_{\text{Softplus}}\}$$

Softplus：$\log(1+e^x)$



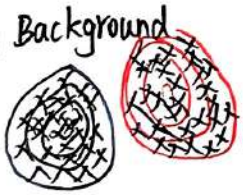
$\therefore P(v)=\frac{1}{Z}\exp\{\alpha^Tv+\sum_{i=1}^m[\text{softplus}(W_iv+\beta_i)]\}$ 其中W_i是W的行向量

Learning 放在后面. P68

---

21. Spectral Clustering

21.1 Background



GMM/ k-means 不好做

kernel + k-means 先映射高维

- compactness：K-means，GMM
- connectivity：spectral clustering

21.2 Model Introduction

Graph-based（带权重的无向图）

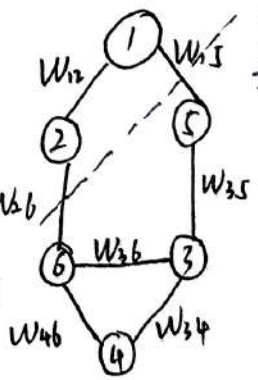
$X=(x_1,x_2\cdots x_N)^T=\begin{pmatrix}x_1^T\\x_2^T\\\vdots\\x_N^T\end{pmatrix}=\begin{pmatrix}\cdots\\\cdots\\\cdots\end{pmatrix}_{N\times p}$

$V=\bigcup_{k=1}^K A_k$

$A_i\cap A_j=\phi,\ \forall i,j\in\{1,2,\cdots,k\}$

$G=\{V,E\}$；$V=\{1,2,\cdots,N\}\Leftrightarrow X$；$E=W=[w_{ij}]_{N\times N}$，$i\le i,j\le N$

W：similarity matrix（affinity matrix），其权重 $w_{ij}=\begin{cases}K(x_i,x_j)=\exp\{-\frac{\|x_i-x_j\|_2^2}{2\sigma^2}\}, & (i,j)\in E\ \text{（边）}\\0, & \text{otherwise}\to(i,j)\notin E\end{cases}$



要切K刀，这是其中一刀

定义：$A\subset V$，$B\subset V$，$A\cap B=\phi$

$$W(A,B)=\sum_{\substack{i\in A\\j\in B}}w_{ij}=\sum_{i\in A}\sum_{i\in B}w_{ij}$$

如图①②刻画其余各点的权重合，无边的w_{ij}为零.

假如：一共 K 个类别，$cut(v)=cut(A_1,\cdots A_K)=\sum_{i=1}^K W(A_i,\bar{A}_i)$

$=\sum_{i=1}^K[W(A_i,v)-W(A_i,A_i)]$

目标：$\min Cut(V)$ $\{A_k\}_{k=1}^{K}$，$Cut(V)=\sum_{i=1}^{k}\frac{W(A_i,\bar{A}_i)}{\Delta}$，$\Delta=degree(A_k)$ 有向图：出度、入度

作约束

无向图：度

如 $d_1=\sum_{j=1}^{N}w_{1j}$；$d_2=\sum_{j=1}^{N}w_{2j}$

$\therefore \Delta=degree(A_k)=\sum_{i\in A_k}d_i$，$d_i=\sum_{j=1}^{N}w_{ij}$

Model

$\therefore Ncut=\sum_{k=1}^{K}\frac{W(A_k,\bar{A}_k)}{\sum_{i\in A_k}d_i}$，$d_i=\sum_{j=1}^{N}w_{ij}$

目标

$\arg\min Ncut$ - 优化问题 $\{A_k\}_{k=1}^{K}$ A_k是J顶点的一个组合

$\{\hat{A}_k\}_{k=1}^{K}=\arg\min_{\{A_k\}_{k=1}^{K}} Ncut(V)$

2.3 Matrix - Indicator Vector / Diagnal Matrix

$1\le i\le N$ 样本个数 N

$1\le j\le K$ 类别个数 K

indicator vector: K维 $y_i=(\ )_{K\times 1}$

$\begin{cases} y_{ij}\in\{0,1\} \\ \sum_{j=1}^{K}y_{ij}=1 \rightarrow \begin{pmatrix} y_{i1}\\ y_{i2}\\ \vdots\\ y_{iK}\end{pmatrix}$ 其中只有一个1

$\{A_k\}_{k=1}^{K}\rightarrow Y=(y_1,y_2\cdots y_N)^T_{N\times K}=\begin{pmatrix} y_1^T\\ y_2^T\\ \vdots\\ y_N^T\end{pmatrix}$

$y_{ij}=1 \Leftrightarrow$ 第i个样本属于第j个类别。目标：$\hat{Y}=\arg\min Ncut(V)$

$O_{K\times K}\Leftrightarrow 0$，P65

Diagnal Matrix

$Ncut(V)=\sum_{k=1}^{K}\frac{W(A_k,\bar{A}_k)}{\sum_{i\in A_k}d_i}=tr\begin{pmatrix} \frac{W(A_1,\bar{A}_1)}{\sum_{i\in A_1}d_i} & & 0\\ & \frac{W(A_2,\bar{A}_2)}{\sum_{i\in A_2}d_i} & \\ & & \ddots\ \frac{W(A_K,\bar{A}_K)}{\sum_{i\in A_K}d_i}\\ 0 & & \end{pmatrix}=tr\begin{pmatrix} W(A_1,\bar{A}_1) & & 0\\ & W(A_2,\bar{A}_2) & \\ & & \ddots\ W(A_K,\bar{A}_K)\\ 0 & & \end{pmatrix}\cdot$

求对角线的和

$\begin{pmatrix} \sum_{i\in A_1}d_i & & 0\\ & \sum_{i\in A_2}d_i & \\ & & \ddots\ \sum_{i\in A_K}d_i\\ 0 & & \end{pmatrix}^{-1}$

$P_{K\times K}$

$\therefore Ncut(V)=tr\,O\cdot P^{-1}$

已知 W、Y，求 O、P

$Y^TY=(y_1,y_2\cdots y_N)\begin{pmatrix} y_1^T\\ y_2^T\\ \vdots\\ y_N^T\end{pmatrix}=\sum_{i=1}^{N}y_i\cdot y_i^T=\begin{pmatrix} N_1 & & \\ & N_2 & \\ & & \ddots\ N_K\end{pmatrix}_{K\times K}$

A_k中的个数

N_k：在N个样本中，属于类K的样本个数，$\sum_{k=1}^{K}N_k=N$，$N_k=|A_k|=\sum_{i\in A_k}1$

$Y^TY=\begin{pmatrix} \sum_{i\in A_1}1 & & \\ & \sum_{i\in A_2}1 & \\ & & \ddots\ \sum_{i\in A_K}1\end{pmatrix}_{K\times K}$

$\therefore \boxed{P=Y^T\cdot D\cdot Y}$，$D=\begin{pmatrix} d_1 & & \\ & d_2 & \\ & & \ddots\ d_N\end{pmatrix}_{N\times N}$ 有N个样本点

$K\times N$ $N\times N$ $N\times K$

$=diag(W\cdot 1_N)$

$i\rightarrow d_i\rightarrow$ 到N个点，参考 $d_i=\sum_{j=1}^{N}w_{ij}$

$W\cdot 1_N=\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1N}\\ w_{21} & w_{22} & \cdots & w_{2N}\\ \vdots & & & \vdots\\ w_{N1} & w_{N2} & \cdots & w_{NN}\end{pmatrix}_{N\times N}\begin{pmatrix} 1\\ 1\\ \vdots\\ 1\end{pmatrix}_{N\times 1}=\begin{pmatrix} d_1\\ d_2\\ \vdots\\ d_N\end{pmatrix}_{N\times 1}$

$\therefore P=Y^T\cdot diag(W1_N)\cdot Y$

Lapalacian Matrix:

$$O = \begin{pmatrix} W(A_1,\bar{A}_1) & & \\ & \ddots & \\ & & W(A_K,\bar{A}_K) \end{pmatrix}_{K\times k}$$

$$W(A_k,\bar{A}_k) = \underbrace{W(A_k,V)}_{\sum_{i\in A_k} d_i} - \underbrace{W(A_k,A_k)}_{\sum_{i\in A_k}\sum_{j\in A_k} W_{ij}}$$

$$\therefore O = \begin{pmatrix} \sum_{i\in A_1} d_i & \cdots & 0 \\ \vdots & \ddots & \\ 0 & \cdots & \sum_{i\in A_k} d_i \end{pmatrix} - \underbrace{\begin{pmatrix} W(A_1,A_1) & \cdots & 0 \\ \vdots & \ddots & \\ 0 & \cdots & W(A_k,A_k) \end{pmatrix}}_{A}$$

$$O = Y^T\cdot D\cdot Y - A$$

猜测 $Y^T W Y = (y_1 \cdots y_N)\begin{pmatrix} W_{11} & \cdots & W_{1N} \\ \vdots & & \vdots \\ W_{N1} & \cdots & W_{NN} \end{pmatrix}\begin{pmatrix} y_1^T \\ \vdots \\ y_N^T \end{pmatrix}$

($K\times N$, $N\times N$, $N\times k$)

$$= \left(\sum_{i=1}^N y_i W_{i1} \cdots \sum_{i=1}^N y_i W_{iN}\right)\begin{pmatrix} y_1^T \\ \vdots \\ y_N^T \end{pmatrix}$$

$$= \sum_{i=1}^N\sum_{j=1}^N y_i y_j^T W_{ij}$$

$$= \begin{pmatrix} \sum_{i\in A_1}\sum_{j\in A_1} W_{ij} & \sum_{i\in A_1}\sum_{j\in A_2} W_{ij} & \cdots & \sum_{i\in A_1}\sum_{j\in A_k} W_{ij} \\ \vdots & & & \\ \sum_{i\in A_k}\sum_{j\in A_1} W_{ij} & & \cdots & \sum_{i\in A_k}\sum_{j\in A_k} W_{ij} \end{pmatrix}$$

$W' = Y^T W Y$

总和同但与P作用后的trace相同

$O' = Y^T D Y - Y^T W Y$, $tr(OP^{-1}) = tr(O'P^{-1})$

注: $Ncut(V) = tr\, O\cdot P^{-1} = tr(Y^T\cdot D\cdot Y - A)\cdot P^{-1} \Leftrightarrow tr(Y^T\cdot D\cdot Y - W')\cdot P^{-1}$

why? : 对角线, trace只要对角线, 且O'P, P为对角阵, 只影响对角

$$\therefore = tr(Y^T\cdot D\cdot Y - Y^T W Y)\cdot P^{-1}$$

$$= tr\left(Y^T(D-W)Y\cdot (Y^T\cdot D\cdot Y)^{-1}\right)$$, 且 $L = D - W$, 为Laplacian Matrix

$$\therefore \hat{Y} = \underset{Y}{\arg\min}\, Ncut(V) = \underset{Y}{\arg\min}\, tr\left[Y^T(D-W)Y\cdot (Y^T\cdot D\cdot Y)^{-1}\right]$$

---

22. Feed-forward Neural Network

21.1 From ML to DL

Machine Learning

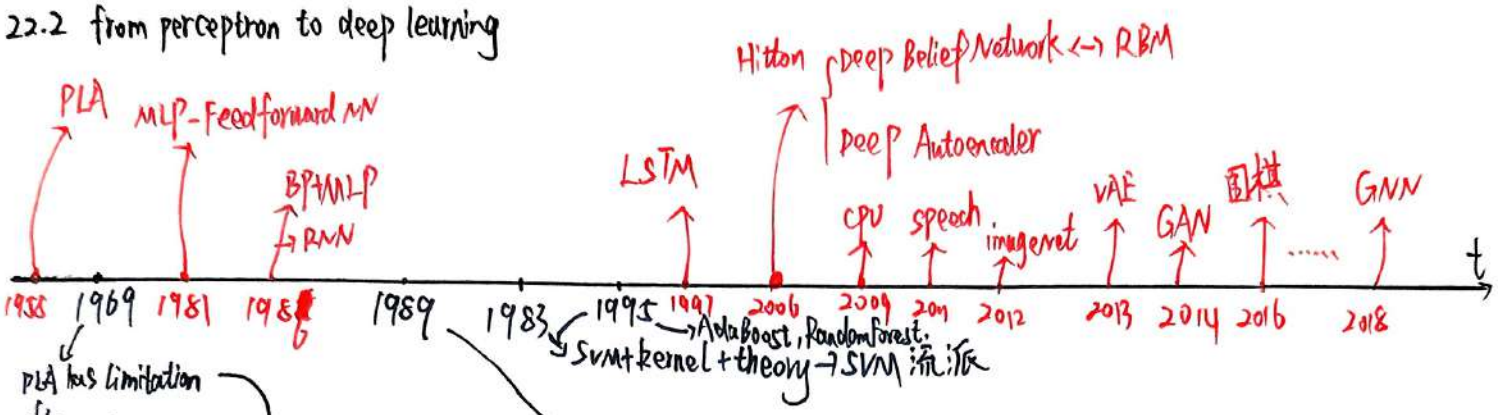
- 频率派 / 统计学习
  - 正则化: loss function + regularizer (惩罚项)
  - 核化: kernel SVM
  - 集成化: AdaBoost, Random Forest
  - 层次化: Neural Network
    - MLP (Mutilayer Perceptron)
    - Autoencoder
    - CNN
    - RNN

  → Deep Neural Network ①
- 贝叶斯派 / PGM
  - 有向: Bayesian Network → Deep Directed Network
    - sigmoid Believe Network
    - variational Autoencoder (VAE)
    - GAN ~~Generative~~ Generative Adversarial network
  - 无向: Markov Network → Deep Boltzmann Machine
  - 有向+无向: Mixed Network → Deep Believe Network

  → Deep Generative Model ②

① 与 ② ⇒ ① ② } Deep Learning

22.2 from perceptron to deep learning



PLA has limitation
↓
解决了 Non-linear
↓ XOR Problem

→ 马文·明斯基 证智能这个AI → 图灵奖

Universal Approximation Theorem
1层 hidden layer → 连续函数
1 layer is good → why deep?
BP → 梯度消失

22.3 non-linear transformation

Solution:

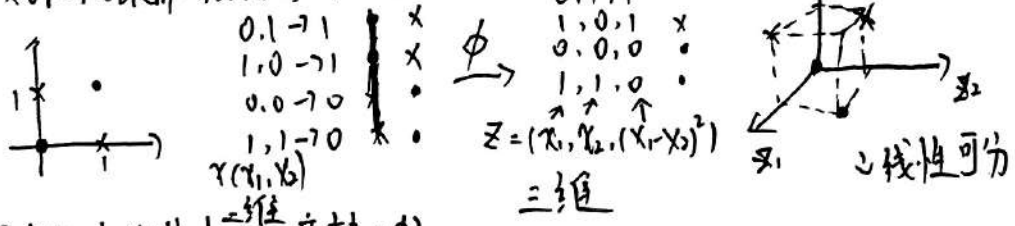
① Non-Transformation: 明转(ϕ)

$\phi: x \longmapsto z$

input space → feature space

Cover's Theorem: 高维比低维更易线性可分

XOR Problem: 相同为0，不同为1



② kernel Method 二维 → 暗转(ϕ)

$K(x, x') = \langle \phi(x), \phi(x') \rangle$ 隐藏了一个 ϕ

$x, x' \in X$

③ 神经网络 (XOR Prob) → 自转(ϕ)

XOR　OR　AND　NOT

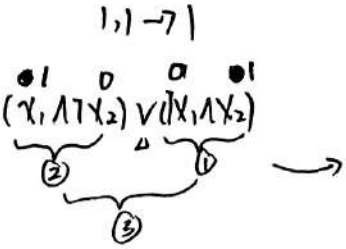
$\oplus$　$\vee$　$\wedge$　$\neg$

复合运算　基本运算

XOR → 相同为0，不同为1

$x_1 \oplus x_2 = (\neg x_1 \wedge x_2) \vee (x_1 \wedge \neg x_2)$

x_1 XOR $x_2 \to 1 = \begin{cases} 1,0 & ① \\ 0,1 & ② \end{cases} = ① \vee ② = (x_1 \wedge \neg x_2) \vee (\neg x_1 \wedge x_2)$



复合运算 → 复合表达式

$\vee$ OR: x_1 OR $x_2 \to 1$

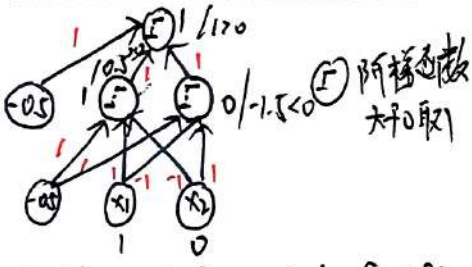
1,0 → 1
0,1 → 1
1,1 → 1

$\wedge$ AND: x_1 AND $x_2 \to 1$

1,1 → 1

$\neg$ Not: x_1 Not $x_2 \to 1$

0 → 1



有向无环不图
↓
神经网络

① $\wedge$
② $\wedge$
③ $\vee$

复合运算 → 复合表达式 → 复合函数

23. Confronting Partition function

23.1 The log-Likelihood gradient

动机: Learning, evaluation

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

$p(x|\theta) = \frac{1}{Z(\theta)} \hat{P}(x|\theta)$, $Z(\theta) = \int \hat{P}(x|\theta)dx$ $\int \hat{P}(x|\theta)dx \neq 1$, $\hat{P}(x|\theta)$未归一化, $P(x|\theta)$归一化后一个有效的概率分布

Given $X = \{x_i\}_{i=1}^N$, estimate: θ

ML上: ~~$\theta = \arg\max \prod_{i=1}^N$~~

$$\hat{\theta} = \arg\max_\theta P(x|\theta) = \arg\max_\theta \prod_{i=1}^N P(x_i|\theta) = \arg\max_\theta \log \prod_{i=1}^N P(x_i|\theta) = \arg\max_\theta \sum_{i=1}^N \log P(x_i|\theta)$$

$$= \arg\max_\theta \sum_{i=1}^N [\log \hat{P}(x_i|\theta) - \log Z(\theta)] = \arg\max_\theta \sum_{i=1}^N \log \hat{P}(x_i|\theta) - \log(Z(\theta)) \cdot N$$

↳ 代入 $P(x|\theta) = \frac{1}{Z(\theta)} \cdot \hat{P}(x|\theta)$

$$= \arg\max_\theta \frac{1}{N} \sum_{i=1}^N [\log \hat{P}(x_i|\theta)] - \log Z(\theta)$$

$L(\theta)$

正相 负相分解

Positive phase negative phase

$$\nabla_\theta L(\theta) = \underbrace{\frac{1}{N} \sum_{i=1}^N \nabla_\theta \log \hat{P}(x_i|\theta)}_{①} - \underbrace{\nabla_\theta \log Z(\theta)}_{②}$$

相对后面积分为常数.

$$② = \frac{1}{Z(\theta)} \cdot \nabla_\theta \cdot Z(\theta) = \frac{P(x|\theta)}{\hat{P}(x|\theta)} \cdot \nabla_\theta \int \hat{P}(x|\theta)dx = \boxed{\frac{P(x|\theta)}{\hat{P}(x|\theta)}} \cdot \underbrace{\int \nabla_\theta \hat{P}(x|\theta)dx}_{\nabla_\theta Z(\theta)} = \int \frac{P(x|\theta)}{\hat{P}(x|\theta)} \cdot \nabla_\theta \hat{P}(x|\theta)dx$$

$\frac{1}{Z(\theta)}$与θ相关, 不积.

参考: $Z(\theta) = \int \hat{P}(x|\theta)dx$, x被积了

$\because \nabla_\theta \log \hat{P}(x|\theta) = \frac{1}{\hat{P}(x|\theta)} \cdot \nabla_\theta \hat{P}(x|\theta)$

$\therefore ② = \int P(x|\theta) \cdot \nabla_\theta \log \hat{P}(x|\theta)dx = \underset{P(x|\theta)}{E}[\nabla_\theta \log \hat{P}(x|\theta)]$ → 求近似 (MC)

① 可以使用梯度上升法

$$\nabla_\theta L(\theta) = \frac{1}{N} \sum_{i=1}^N \nabla_\theta \log \hat{P}(x_i|\theta) - \underset{P(x|\theta)}{E}[\nabla_\theta \log \hat{P}(x|\theta)]$$

23.2 Stochastic Maximum Likelihood

data distribution: P_{data}, 经验分布

model distribution: $P_{model} \triangleq P(x|\theta)$ ← 自己的假设

$\therefore \nabla_\theta L(\theta) = \underbrace{E_{P_{data}}[\nabla_\theta \log \hat{P}(x_i|\theta)]}_{\text{positive phase}} - \underbrace{E_{P_{model}}[\nabla_\theta \log \hat{P}(x|\theta)]}_{\text{negative phase}}$ ← 真实的

$E_{P_{data}}[\nabla_\theta \log \hat{P}(x_i|\theta)]$

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

$L(\theta) = \frac{1}{N} \sum_{i=1}^N \log \hat{P}(x_i|\theta) - \log Z(\theta)$

$$\theta^{(t+1)} = \theta^{(t)} + \eta [(\sum_{i=1}^m \nabla_\theta \log \hat{P}(x_i|\theta^{(t)})) - \sum_{i=1}^m \nabla_\theta \log \hat{P}(\hat{x}_i|\theta^{(t)})]$$

Gradient Ascent based on MCMC

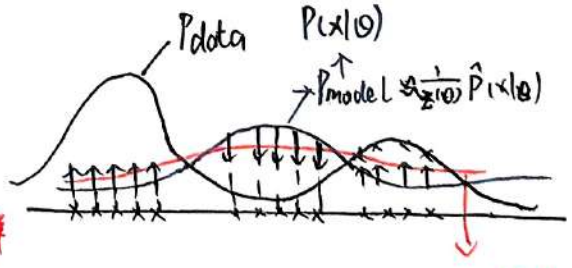
相同时 $\nabla_\theta L(\theta) = 0$, 停止 stop

Gradient Ascend:

$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta L(\theta^{(t)})$

Gibbs Sampling from P_{model} $= P(x|\theta^{(t)})$, $i = 1, 2, \cdots, m$. $\{\hat{x}_i\}_{i=1}^m$ 从上一次迭代采样

$\hat{x}_1 \sim P(x|\theta^{(t)})$
$\vdots$
$\hat{x}_m \sim P(x|\theta^{(t)})$
} fantacy particles



$Z(\theta) = \int \hat{P}(x|\theta)dx$ 减小

23.3 What is Contrastive Divergence

由上述两节总结：

$x \in \mathbb{R}^n$ 或 $\{0,1\}^n$，$P(x|\theta) = \frac{1}{Z(\theta)}\hat{P}(x|\theta)$，$Z_\theta = \int \hat{P}(x|\theta)dx$

MLE：Given $X = \{x_i\}_{i=1}^N$，Estimate θ

$$\begin{cases} \hat{\theta} = \arg\max_\theta L(\theta) = \arg\min_\theta KL(P_{data} \| P_{model}) = \arg\min_\theta KL(P^{(0)} \| P^{(\infty)}) \\ L(\theta) = \frac{1}{N}\sum_{i=1}^N \log \hat{P}(x_i|\theta) - \log Z(\theta) \end{cases}$$

$$\nabla_\theta L(\theta) = \underbrace{E_{P_{data}}[\nabla_\theta \log \hat{P}(x_i|\theta)]}_{\text{postive phase}} - \underbrace{E_{P_{model}}[\nabla_\theta \log \hat{P}(x|\theta)]}_{\text{negative phase}}$$

Gradient Ascend:

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla_\theta L(\theta^{(t)})$$

when $t+1$:

① sampling for positive phase from P_{data}

$x_1, x_2, \cdots, x_m$ themself are training data

② Sampling for negative phase from $P_{model} = P(x|\theta^{(t)})$ (CD Learning)

$\hat{x}_i = x_i$：指直接从 training data 中取 (k-CD, k=1,2,…; 1-CD; 2-CD)

- Initialization $\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_m$ (any distribution)

mixing time：0-step, 1-step, …, k-step（后为平稳态）$\hat{x}_1$

CD-k 采用 Gibbs Sampling

$$\theta^{(t+1)} = \theta^{(t)} + \eta \cdot \left[\sum_{i=1}^m \nabla_\theta \log \hat{P}(x_i|\theta^{(t)}) - \sum_{i=1}^m \nabla_\theta \log \hat{P}(\hat{x}_i|\theta^{(t)})\right]$$

CD-k：在 Markov 链中在第 k 点停止，不用达到平稳分布

23.4 The Name of Contrastive Divergence

$$\hat{\theta} = \arg\max_\theta \frac{1}{N}\sum_{i=1}^N \log P(x_i|\theta) = \arg\max_\theta E_{P_{data}}[\log P_{model}(x_i|\theta)] = \arg\max_\theta \int P_{data} \log P_{model}\, dx$$

$$= \arg\max_\theta \int P_{data} \cdot \log \frac{P_{model}}{P_{data}} dx = \arg\max_\theta -KL(P_{data} \| P_{model}) = \arg\min_\theta KL(P_{data} \| P_{model})$$

$P^{(t)}$ / $P^{(0)}$ $P^{(\infty)}$

CD-k ($k=1,2,\cdots$)

$$CD: KL(P^{(0)} \| P^{(\infty)}) - KL(P^{(k)} \| P^{(\infty)})$$

$$\hat{\theta} = \arg\min \left[\underbrace{KL(P^{(0)} \| P^{(\infty)}) - KL(P^{(k)} \| P^{(\infty)})}_{\text{Contrastive Divergence}}\right]$$

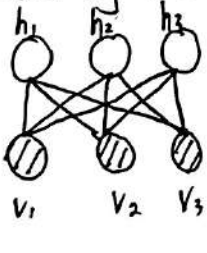
在 23.3 中，修改 Initialization 变为 CD-Learning

$P_{data} = P^{(0)}$

$P_{model} = P^{(\infty)}$

对 $KL(P^{(0)} \| P^{(\infty)}) - KL(P^{(k)} \| P^{(\infty)})$ 求梯度为 $\sum_{i=1}^m \nabla_\theta \log \hat{P}(x_i|\theta^{(t)}) - \sum_{i=1}^m \nabla_\theta \log \hat{P}(\hat{x}_i|\theta^{(t)})$

23.5 log-likehood gradient of energy-based model / RBM Learning



$$\begin{cases} P(h,v) = \frac{1}{Z}\exp\{-E(h,v)\} \\ E(h,v) = -(h^T W v + \alpha^T v + \beta^T h) \end{cases}$$

$$\begin{cases} h = (h_1, h_2 \cdots h_m)^T \\ v = (v_1, v_2 \cdots v_n)^T \\ W = [w_{ij}]_{m \times n} \\ \alpha = (\alpha_1, \alpha_2 \cdots \alpha_n)^T \\ \beta = (\beta_1, \beta_2 \cdots \beta_m)^T \end{cases}$$

$$\log P(v) = \log \frac{1}{Z}\sum_h \exp\{-E(h,v)\}$$

$$= \log \sum_h \exp\{-E(h,v)\} - \log Z$$

$$= \log \sum_h \exp\{-E(h,v)\} - \log \sum_{h,v} \exp\{-E(h,v)\}$$

log-likelihood：training set $v \in S$，$|S| = N$

$$\frac{1}{N}\sum_{v \in S} \log P(v)$$

log-likelihood gradient：

$$\frac{\partial}{\partial \theta}\frac{1}{N}\sum_{v \in S} \log P(v)$$

$$\log P(v) = \log \sum_h P(h,v) = \log \sum_h \cdot \frac{1}{Z}\exp\{-E(h,v)\}$$

$$= \log \sum_h \exp\{-E(h,v)\} - \log \sum_{h,v} \exp\{-E(h,v)\}$$

$$= ① - ②$$

假设 $\theta = \{w, \alpha, \beta\}$

$$\therefore \frac{\partial}{\partial \theta} \log P(v) = \frac{\partial}{\partial \theta} ① - \frac{\partial}{\partial \theta} ②$$

$$\frac{\partial}{\partial \theta} ① = \frac{\partial}{\partial \theta} \log \sum_h \exp\{-E(h,v)\} = \frac{1}{\sum_h \exp\{-E(h,v)\}} \sum_h \cdot \exp\{-E(h,v)\} \cdot \frac{\partial}{\partial \theta} \cdot (-E(h,v))$$

$$= \frac{-1}{\sum_h \exp\{-E(h,v)\}} \cdot \sum_h \cdot \exp\{-E(h,v)\} \cdot \frac{\partial}{\partial \theta} E(h,v)$$

P68

$$\frac{\partial}{\partial\theta}①=\sum_h\cdot\frac{-1\cdot\exp\{-E(h,v)\}}{\sum_h\exp\{-E(h,v)\}}\cdot\frac{\partial}{\partial\theta}E(h,v)\quad=(-1)\sum_h\cdot\frac{\frac{1}{Z}\exp\{-E(h,v)\}\cdot\frac{\partial}{\partial\theta}E(h,v)}{\sum_h\cdot\frac{1}{Z}\exp\{-E(h,v)\}}=-\sum_h\frac{P(h,v)}{\sum_h\cdot P(h,v)}\cdot\frac{\partial}{\partial\theta}E(h,v)$$

$$=-\sum_h\frac{P(h,v)}{P(v)}\cdot\frac{\partial}{\partial\theta}E(h,v)=-\sum_h P(h|v)\cdot\frac{\partial}{\partial\theta}E(h,v)$$

$$\frac{\partial}{\partial\theta}②=\frac{\partial}{\partial\theta}\log\sum_{h,v}\exp\{-E(h,v)\}=\frac{1}{\sum_{h,v}\exp\{-E(h,v)\}}\sum_{h,v}\exp\{-E(h,v)\}\cdot(-1)\cdot\frac{\partial E(h,v)}{\partial\theta}$$

$$=-\sum_{h,v}\underbrace{\frac{\exp\{-E(h,v)\}}{\sum_{h,v}\exp\{-E(h,v)\}}}_{Z}\cdot\frac{\partial E(h,v)}{\partial\theta}=-\sum_{h,v}P(h,v)\cdot\frac{\partial E(h,v)}{\partial\theta}$$

$$\therefore\frac{\partial}{\partial\theta}\log P(v)=\sum_{h,v}P(h,v)\cdot\frac{\partial E(h,v)}{\partial\theta}-\sum_h P(h|v)\cdot\frac{\partial}{\partial\theta}E(h,v)$$ 基于energy-based model，a log-likelihood

---

23.6 log-likelihood gradient of RBM

Objective: Given training set: $v\in S$, $|S|=N$

log-likelihood gradient for RBM:

$$\frac{\partial}{\partial\theta}\cdot\frac{1}{N}\cdot\sum_{v\in S}\log P(v)$$

$\theta=\{w,a,\beta\}$ S: training set

$$\frac{\partial}{\partial\theta}\log P(v)=\sum_{h,v}P(h,v)\cdot\frac{\partial E(h,v)}{\partial\theta}-\sum_h P(h|v)\cdot\frac{\partial}{\partial\theta}E(h,v)$$

求 $W=[w_{ij}]_{m\times n}$:

$$\frac{\partial}{\partial w_{ij}}\log P(v)=-\sum_h P(h|v)\cdot\frac{\partial E(h,v)}{\partial w_{ij}}+\sum_{h,v}P(h,v)\cdot\frac{\partial E(h,v)}{\partial w_{ij}}$$

$$\because E(h,v)=-(h^TW\cdot v+a^Tv+\beta^Th)=-(h^TWv+\Delta)=-\left(\sum_{i=1}^{m}\sum_{j=1}^{n}h_iw_{ij}v_j+\Delta\right)$$

$$\therefore\frac{\partial}{\partial w_{ij}}\log P(v)=-\sum_h P(h|v)\cdot(-h_iv_j)+\sum_{h,v}P(h,v)\cdot(-h_iv_j)$$

$$=\underbrace{\sum_h P(h|v)\cdot h_iv_j}_{①}-\underbrace{\sum_{h,v}P(h,v)\cdot h_iv_j}_{②}$$

RBM: 二值，$v_i\in\{0,1\}$, $h_i\in\{0,1\}$

$h_i\neq0$，=0就约掉

$$①=\sum_{h_1}\sum_{h_2}\cdots\sum_{h_i}\cdots\sum_{h_m}P(h_1,h_2,\cdots,\underbrace{h_i}_{\Delta},\cdots,h_m|v)\cdot h_iv_j=\sum_{h_i}P(h_i|v)\cdot h_iv_j=\boxed{P(h_i=1|v)\cdot v_j}$$

P62有求解公式

$$②=\sum_{h_0}\sum_v\cdot P(v)\cdot P(h|v)\cdot h_iv_j=\sum_v P(v)\cdot\sum_h P(h|v)\cdot h_iv_j=\boxed{\sum_v P(v)\cdot P(h_i=1|v)\cdot v_j}$$

同①

$v\in\{0,1\}^n$ 指loof, 2^{100}. intractable

CD来解决.

23.7 CD-K Algorithm for RBM

log-likelihood gradient for RBM:

$$\frac{\partial}{\partial w_{ij}} \log P(v) = P(h_i=1|v)\, v_j - \sum_v P(v)\, P(h_i=1|v)\cdot v_j = E_{P(v)}[P(h_i=1|v) v_j]$$

$$\frac{\partial}{\partial w_{ij}} \cdot \frac{1}{N} \cdot \sum_{v\in S} \log P(v) = \frac{1}{N} \sum_{v\in S} \frac{\partial}{\partial w_{ij}} \log P(v) \approx \frac{1}{N} \Delta w_{ij}$$

CD-K for RBM

For each $v \in S$

$v^{(0)} \leftarrow v$

For $l = 0, 1, 2, \cdots, k-1$:

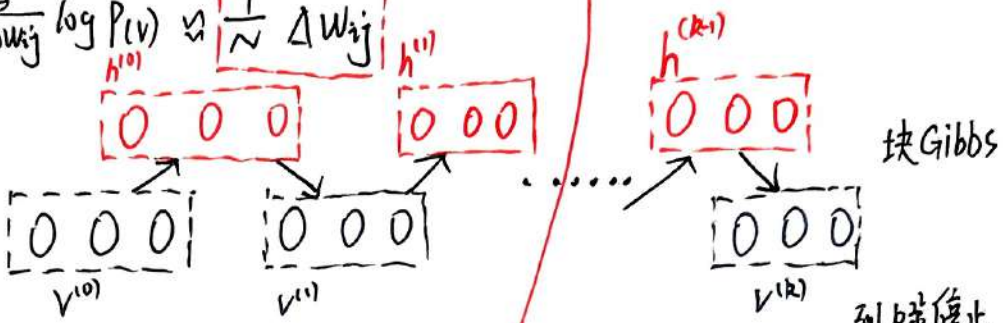
For $i = 1, 2, \cdots, m$: Sample $h_i^{(l)} \sim P(h_i|v^{(l)})$

For $j = 1, 2, \cdots, n$: Sample $v_j^{(l+1)} \sim P(v_j|h^{(l)})$

} k-step Block Gibbs sampling

For $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, n$:

$$\Delta w_{ij} \leftarrow \Delta w_{ij} + \frac{\partial}{\partial w_{ij}} \log P(v) \approx P(h_i=1|v^{(0)})\cdot v_j^{(0)} - P(h_i=1|v^{(k)})\cdot v_j^{(k)}$$



---

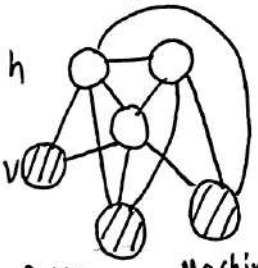
24. Approximate Inference

24.1. Introduction

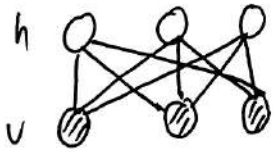
① 推断的动机 { 推断本身 ($P(h|v)$) ; 学习需要

② 推断是困难的 { 有向图: explain away ; 无向图: mutual interaction

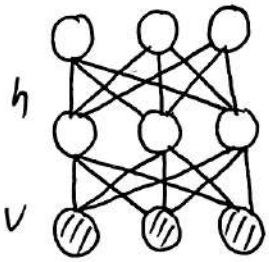
explain away: head to head, 给v后, h不是相互独立



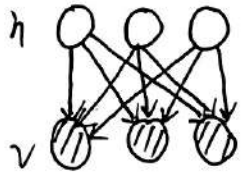
Boltzmann Machine



RBM



Deep BM



Sigmoid Belief network

24.2 Inference is optimization

$V = \{v\}$ 样本

log-likelihood: $\sum_{v\in V} \log P(v)$

$-\int q(h|v) \log q(h|v) = H[q(h|v)]$

$$\log P(v) = \log \frac{P(v,h)}{P(h|v)} = \log \frac{P(v,h)}{q(h|v)} \cdot \frac{q(h|v)}{P(h|v)} = \log \frac{P(v,h)}{q(h|v)} + \log \frac{q(h|v)}{P(h|v)}$$

$$\int \log P(v)\cdot q(h|v)\, dh = \int \log \frac{P(v,h)}{q(h|v)} \cdot q(h|v)\, dh + \int \log \frac{q(h|v)}{P(h|v)} \cdot q(h|v)\, dh$$

$\log P(v) \int q(h|v) dh \rightarrow \log P(v) = E_{q(h|v)}\left[\log \frac{P(v,h)}{q(h|v)}\right] + E_{q(h|v)}\, KL(q(h|v)\,||\,P(h|v))$

$$= E_{q(h|v)}[\log P(v,h) - \log q(h|v)] + KL(q(h|v)||P(h|v))$$

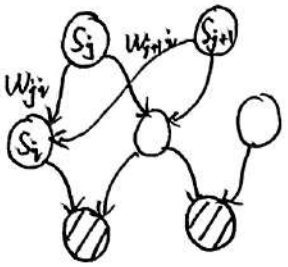
$$= \underbrace{E_{q(h|v)}[\log P(v,h)]}_{\text{ELBO} = \mathcal{L}(v,h,q)} + \underbrace{H[q(h|v)]}_{\text{P70}} + \underbrace{KL(q(h|v)||P(h|v))}_{KL(q\|p)}$$

25. Sigmoid Belief Network

25.1 Introduction

Neal → a student of Hinton

$S=\{s_1,s_2,\cdots,s_n\}=\{v,h\}=\{v,h^{(1)},h^{(2)}\}$，$v,h$为$\{0,1\}$一值



$$\begin{cases} P(S_i=1)=\sigma\left(\sum_{j<i} W_{ji}S_j\right) & \text{对}\ j<i \\ P(S_i=0)=1-P(S_i=1)=\sigma\left(-\sum_{j<i} W_{ji}S_j\right) & \end{cases}$$

$$\sigma(x)=\frac{1}{1+\exp(-x)}$$

Neal → [1990]

$$1-\sigma(x)=\frac{\exp(-x)}{1+\exp(-x)}=\frac{1}{1+\exp(x)}=\sigma(-x)$$

$$\sigma'(x)=\sigma(x)\cdot\sigma(-x)$$

$$\begin{cases} P(S_i \mid S_j: j<i)=\sigma\left(S_i^*\sum_{j<i} W_{ji}S_j\right) & \\ S_i^*=2S_i-1 & \end{cases}$$

一个式子表达

$\begin{cases} S_i=1, S_i^*=1 \\ S_i=0, S_i^*=-1 \end{cases}$

25.2 gradient of log-likelihood

省一个偏置v

$P(S_i \mid S_j: j<i)=\sigma\left(S_i^*\sum_{j<i} W_{ji}S_j\right)$　$P(S)=\prod_i P(S_i \mid S_j: j<i)=P(v,h)$

log-likelihood: $\sum_{v\in V}\log P(v)$

$$\frac{\partial \log P(v)}{\partial W_{ji}}=\frac{1}{P(v)}\cdot\frac{\partial P(v)}{\partial W_{ji}}=\frac{1}{P(v)}\cdot\frac{\partial\sum_h P(v,h)}{\partial W_{ji}}=\sum_h\frac{1}{P(v)}\frac{\partial P(v,h)}{\partial W_{ji}}=\sum_h\frac{P(h\mid v)}{P(h,v)}\cdot\frac{\partial P(v,h)}{\partial W_{ji}}$$

$$=\sum_h P(h\mid v)\cdot\frac{1}{P(s)}\cdot\frac{\partial P(s)}{\partial W_{ji}}$$

$$\frac{1}{P(s)}\cdot\frac{\partial P(s)}{\partial W_{ji}}=\frac{1}{\prod_k P(S_k\mid S_j:j<k)}\cdot\frac{\Delta b\cdot\partial P(S_i\mid S_j:j<i)}{\partial W_{ji}}=\frac{1}{P(S_i\mid S_j:j<i)\cdot\Delta b}\cdot\frac{\Delta b\cdot\partial P(S_i\mid S_j:j<i)}{\partial W_{ji}}$$

$\prod_k$项中有1项与W_{ji}相关

$$=\frac{1}{P(S_i\mid S_j:j<i)}\cdot\frac{\partial\,\sigma\left(S_i^*\sum_{j<i}W_{ji}S_j\right)}{\partial W_{ji}}$$

$$\sigma(x)=\frac{1}{1+\exp(-x)}=\frac{\exp\{x\}}{1+\exp\{x\}}$$

$$\sigma'(x)=\frac{\exp(-x)}{[1+\exp(-x)]^2}=\frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}}=\sigma(x)\cdot\sigma(-x)$$

$$=\frac{1}{\sigma\left(S_i^*\sum_{j<i}W_{ji}S_j\right)}\cdot\sigma\left(S_i^*\textstyle\sum_{j<i}W_{ji}S_j\right)\cdot\sigma\left(-S_i^*\sum_{j<i}W_{ji}S_j\right)\cdot S_i^*\cdot S_j$$

$$=\sigma\left(-S_i^*\sum_{j<i}W_{ji}S_j\right)\cdot S_i^*\cdot S_j$$

$$\therefore\frac{\partial\log P(v)}{\partial W_{ji}}=\sum_h P(h\mid v)\cdot\sigma\left(-S_i^*\sum_{j<i}W_{ji}S_j\right)\cdot S_i^*\cdot S_j$$

$P(h\mid v)=P(h,v\mid v)=P(s\mid v)$

$P(h\mid v)$算不出　explain away

$$\therefore\frac{\partial}{\partial W_{ji}}\sum_{v\in V}\log P(v)=\sum_{v\in V}\sum_h P(h\mid v)\cdot\sigma\left(-S_i^*\sum_{j<i}W_{ji}S_j\right)\cdot S_i^*\cdot S_j=\sum_{v\in V}\sum_s P(s\mid v)\cdot\sigma\left(-S_i^*\sum_{j<i}W_{ji}S_j\right)\cdot S_i^*\cdot S_j$$

$$=E_{(v,h)\sim P(s\mid v)}\left[\sigma\left(-S_i^*\sum_{k<i}W_{ki}S_k\right)\cdot S_i^*\cdot S_j\right]$$

$v\sim P_{data}=P_{v1}$

$$E_{v,s}[f(v,s)]=E_v[E_{s\mid v}[f(v,s)\mid v]]$$

$$\text{记}=E_{v\sim P_{data}}\left[E_{s\sim P(s\mid v)}\left[\sigma\left(-S_i^*\sum_{k<i}W_{ki}S_k\right)\cdot S_{k\neq i}^*\cdot S_j\right]\right]$$

全其期望式

↳ MCMC采样(Neal 小规模)

P11

$$E_{v\sim P_{data}}[E_{s\sim P(s|v)}[f(v,s)|v]] \approx \frac{1}{M}\sum_{m=1}^{M}[\frac{1}{N}\sum_{n=1}^{N} f(v^{(m)}, s^{(m,n)})]$$

$f(v,s) = s(-s_i\sum_{j<i} W_{ji}s_j)\cdot s_i\cdot s_j$ (目标函数)

$v^{(m)}$ 是从 $P_{data}(v)$ 采样的观测样本

$s^{(m,n)}$ 是对每个 $v^{(m)}$, 从 $P(s|v^{(m)})$ 采样的隐变量样本

外层期望 $v\sim P_{data}$, 直接采

假设我们有训练数据集 $D=\{v^{(1)}, v^{(2)}, \cdots, v^{(|D|)}\}$, 则 $P_{data}(v) \approx \frac{1}{|D|}\sum_{v\in D}\delta(v-\hat{v})$ (经验分布)

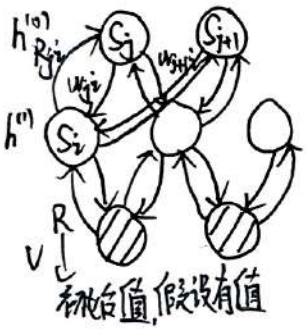
每次训练从 D 中随机抽取 M 个样本 $\{v^{(1)}, v^{(2)}, \cdots, v^{(M)}\}$

内层期望 $s\sim P(s|v)$, Gibbs 高维

内层的 $P(s|v)$ 是高维依赖分布, 用 Gibbs 采样

25.3 Wake-sleep Algorithm - Introduction (Hitton)

Gradient of log-likelihood: $\sum_{v\in V}\sum_{s} P(s|v)\cdot s_i^*\cdot s_j\, \delta(-s_i^*\sum_{k<i} s_k W_{ki})$



$\theta = W$

Model: $P_\theta(v,h)$

W ↓ Generative Connection

R ↑ Recognization Connection (假设存在)

Model: $q_\phi(h|v)$

$\phi = R$

Wake Phase:

① Bottom-up 激活 neuron (获得各层样本)

② Learning Generative Connection (求 W)

sleep Phase:

① Top-down 激活 neuron (获得各层样本)

② Learning Recognization Connection (求 R)

$P(h|v)$ 求不出, 近似一个 $q(h|v)$, $q(h|v)$ 的参数就是 R

25.4 Wake-sleep Algorithm - KL Divergence

Wake Phase: 固定 R 求 W EM => M-step

$$E_{q_\phi(h|v)}[\log P_\theta(v,h)] \approx \frac{1}{N}$$

$$\hat{\theta} = \arg\max_\theta E_{q_\phi(h|v)}[\log P_\theta(h,v)], \text{ with } \phi \text{ fixed}$$

$H[q]$ 的参数是 ϕ, ϕ 是固定的, 与 θ 无关

$$\hat{\theta} = \arg\max_\theta L(\theta) \Leftarrow [ELBO]_{max} \Leftarrow [KL]_{min} \Leftarrow \arg\min_\theta KL[q_\phi(h|v)\|P_\theta(v,h)]$$

Sleep Phase: 固定 W 求 R EM => E-step

$$\hat{\phi} = \arg\max_\phi E_{P_\theta(h,v)}[\log q_\phi(h|v)], \text{ with } \theta \text{ fixed}$$

$$= \arg\max_\phi \int P_\theta(h,v)\cdot \log q_\phi(h|v)\, dh$$

$$= \arg\max_\phi \int P_\theta(v)\cdot P_\theta(h|v)\cdot \log q_\phi(h|v)\, dh$$

只与 θ 有关, 与 h, θ 无关, 提出去对后者略

$$\log P(v) = ELBO + KL(q\|P)$$

$$ELBO = \mathcal{L} = E_{q(h|v)}[\log \frac{P(v,h)}{q(h|v)}]$$

$$= E_{q(h|v)}[\log P(v,h)] + H[q]$$

$$= \arg\max_\phi P_\theta(v)\int P_\theta(h|v)\cdot \log q_\phi(h|v)\, dh$$

$$= \arg\max_\phi \int P_\theta(h|v)\cdot \log q_\phi(h|v)\, dh$$

$$= \arg\max_\phi \int P_\theta(h|v)\cdot \log[\frac{q_\phi(h|v)}{P_\theta(h|v)}\cdot P_\theta(h|v)]\, dh$$

$$= \arg\max_\phi \int P_\theta(h|v)\cdot \log \frac{q_\phi(h|v)}{P_\theta(h|v)} + P_\theta(h|v)\cdot P_\theta(h|v)\, dh$$

与 ϕ 无关

$$= \arg\max_\phi \int P_\theta(h|v)\cdot \log \frac{q_\phi(h|v)}{P_\theta(h|v)}\, dh$$

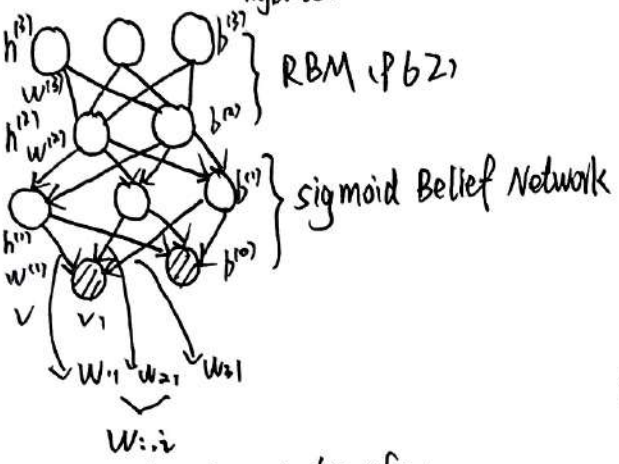
$$= \arg\max_\phi \{-KL[P_\theta(h|v)\|q_\phi(h|v)]\}$$

$$= \arg\min_\phi \{KL[P_\theta(h|v)\|q_\phi(h|v)]\}$$

26. Deep Belief Network

26.1 Introduction

Hinton (2006) hybrid model



$W_{:,i}$: ith column vector of W

$\theta = \{W^{(1)}, W^{(2)}, W^{(3)}, b^{(0)}, b^{(1)}, b^{(2)}, b^{(3)}\}$

有向图 Model

Belief Network ↔ Bayesian Network

$$P(v, h^{(1)}, h^{(2)}, h^{(3)}) = P(v|h^{(1)}, h^{(2)}, h^{(3)}) \cdot P(h^{(1)}, h^{(2)}, h^{(3)})$$

$$= P(v|h^{(1)}) \cdot P(h^{(1)}, h^{(2)}, h^{(3)})$$

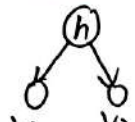
$$= P(v|h^{(1)}) \cdot P(h^{(1)}|h^{(2)}, h^{(3)}) \cdot P(h^{(2)}, h^{(3)})$$

$$= P(v|h^{(1)}) \cdot P(h^{(1)}|h^{(2)}) \cdot P(h^{(2)}, h^{(3)})$$

$$= \prod_i P(v_i|h^{(1)}) \cdot \prod_j P(h_j^{(1)}|h^{(2)}) \cdot P(h^{(2)}, h^{(3)})$$

给定 h 求 v 为 tail-tail

给定 v 求 h explain away



给定 h, $v_1 \perp v_2$

$$P(v_i|h^{(1)}) = \text{sigmoid}(W_{:,i}^{(1)T} \cdot h^{(1)} + b_i^{(0)})$$

到 v_i 这点的所有边 $\rightarrow W_{:,i}^{(0)T}$

利用 P62

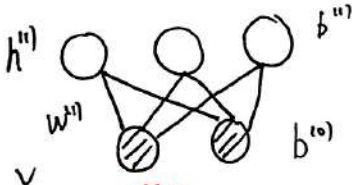
$$P(h_j^{(1)}|h^{(2)}) = \text{sigmoid}(W_{:,j}^{(2)T} \cdot h^{(2)} + b_j^{(1)})$$

$$P(h^{(2)}, h^{(3)}) = \frac{1}{Z}\exp\{\underbrace{h^{(2)T} W^{(3)} \cdot h^{(3)}}_{\text{边}} + \underbrace{h^{(2)} \cdot b^{(2)}}_{\text{点}} + \underbrace{h^{(3)} \cdot b^{(3)}}_{\text{点}}\}$$

$P(h,v) = \frac{1}{Z}\exp\{-E(h,v)\}$

$E(h,v) = -$ 边 + 点 + 点

26.2 Stacking RBM



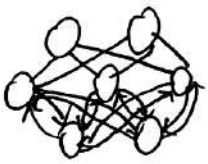
图①

将 RBM 由无向变为有向，需加上蓝色箭头，而假设中固定 $P(v|h^{(1)})$，则需擦去蓝色，留下黑色箭头，变成 DBN

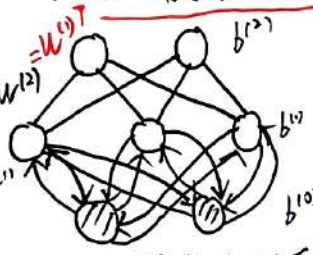
固定 $P(v|h^{(1)})$ 即学习出 $W^{(1)}$，为改进 $W^{(1)}$，则为 $h^{(1)}$ 增加一层或多层

$$P(v) = \sum_{h^{(1)}} P(v, h^{(1)}) = \sum_{h^{(1)}} \underbrace{P(h^{(1)})}_{\text{Prior}} \cdot \underbrace{P(v|h^{(1)})}_{\text{fixed} \rightarrow \text{确定}W\text{后就确定}}$$

为提高 $P(h^{(1)})$ 加上一层 RBM



填加一层 RBM 改善 $P(h^{(1)})$ 提高 $W^{(1)}$



图②

固定 $P(v|h^{(1)})$ 后，向上的蓝色箭头擦去就变成了 DBN

$W^{(1)}$ 的初始点直接赋予 $W^{(1)}$，模型至少与 $W^{(1)}$ 相同，在 $W^{(1)}$ 基础上可以进

$$P(h^{(1)}) = \sum_{h^{(2)}} P(h^{(1)}, h^{(2)})$$

P70，通过 V 学出 $W^{(1)}$，固定 $W^{(1)}$ 加一层为何更好，下一回分解

26.3 Extra layers improve ELBO

关于上节图①

$$\log P(v) = \log \sum_{h^{(1)}} P(v, h^{(1)})$$

$$= \log \sum_{h^{(1)}} q(h^{(1)}|v) \cdot \frac{P(v, h^{(1)})}{q(h^{(1)}|v)}$$

$$= \log \sum_{h^{(1)}} q(h^{(1)}|v) \cdot \frac{P(v, h^{(1)})}{q(h^{(1)}|v)}$$

$$= \log E_{q(h^{(1)}|v)}\left[\frac{P(v, h^{(1)})}{q(h^{(1)}|v)}\right]$$

Jessen

$$\geq E_{q(h^{(1)}|v)}\left[\log \frac{P(v, h^{(1)})}{q(h^{(1)}|v)}\right]$$

$$= \sum_{h^{(1)}} q(h^{(1)}|v)\left[\log P(v, h^{(1)}) - \log q(h^{(1)}|v)\right] \Leftrightarrow \text{ELBO}$$

$$= \sum_{h^{(1)}} q(h^{(1)}|v)\left[\log P(h^{(1)}) + \underbrace{\log P(v|h^{(1)})}_{\text{都确定}} - \log q(h^{(1)}|v)\right]$$

improve $P(h^{(1)})$

$$= \sum_{h^{(1)}} q(h^{(1)}|v) \cdot \log P(h^{(1)}) + C$$

※不加 layer 的话全部跌

关于图②. 2nd layer RBM Learning $\Leftrightarrow$ maximum log-likelihood over $P(h^{(1)})$

$\Leftrightarrow$ maximum ELBO of $P(v)$

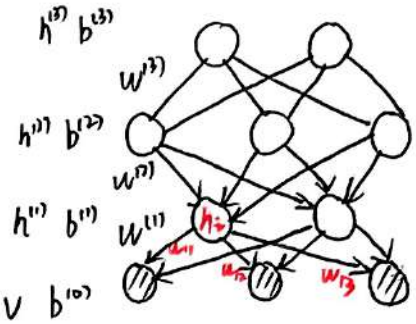
Learning 是求出最好的 $W^{(2)}$ 和 $b^{(1)}$, $b^{(2)}$，使 $P(h^{(1)})$ 的 likeli-hood 达到最大 ($P(h^{(1)})$)，从而在其余者为已知的基础上增大 ELBO，从而增大 $\log P(v)$

26.4 Pre-training

$\log P(v) \geq ELBO = \sum_{h^{(1)}} q(h^{(1)}|v) \cdot \log P(v, h^{(1)}) - \sum_{h^{(1)}} q(h^{(1)}|v) \cdot \log q(h^{(1)}|v)$

$P(h^{(1)}) \cdot P(v|h^{(1)})$

求出总能近似 (explain away)



Generative process 生成样本容易

Gibbs Sampling

当作RBM来算，避免explain away，用RBM近似，则$h^{(1)}$中两个0独立，

$q(h^{(1)}|v) = \prod_i q(h_i^{(1)}|v) = \prod_i sigmoid(W_{i,:}^{(1)} \cdot v + b_i^{(1)})$ （$W_{i,:}^{(1)}$ 行向量，$b_i^{(1)}$ 偏置）

得到 $W_{1,:}$ 与 b_1

以此逐层递进，求 $W_{2,:}, b_2$，$W_{3,:}, b_3$ …

$q(h^{(1)}|v)$ 是一个近似后验，因子分解后的一个近似分布

first, second layer 当作 RBM 求 $q(h^{(1)}|v)$

与真实的 $P(h^{(1)}|v)$ 差距大，$P(h^{(1)}|v)$ 并不是可因子分解的

故ELBO比较松 两层 与P78 PBM呼应

DBN → ELBO is relatively loose 比较松

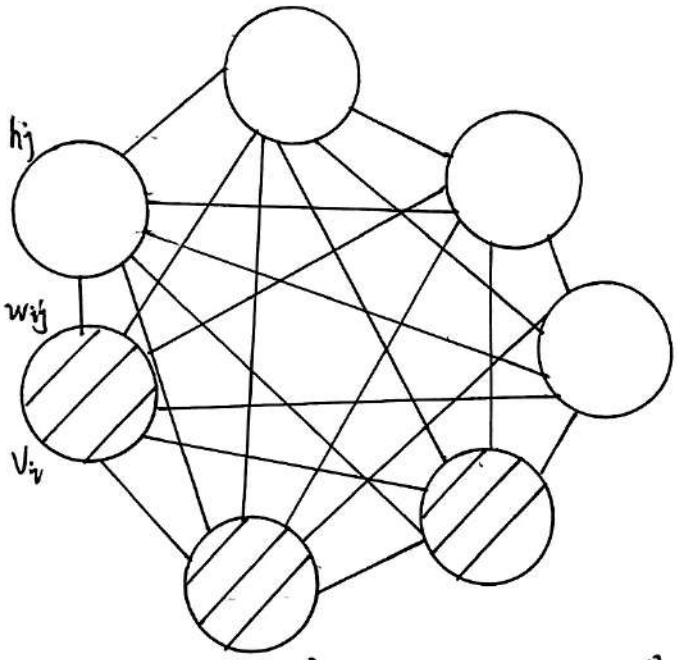
~~此法忽略了 $W^{(2)}$ 对 $h^{(1)}$ 的影响~~

$P(h^{(1)}|v)$ 学出 $W^{(1)}$，再抽样出 $h^{(1)}$，然后再用 $h^{(1)}$，$P(h^{(2)}|h^{(1)})$ 学出 $W^{(2)}$，此时忽略 $W^{(2)}$ 对 $h^{(1)}$ 影响，合并时忽略 $W^{(2)}$ 用 $P(h^{(1)}; W^{(1)})$ 近似 $P(h^{(1)}|W^{(1)}, W^{(2)})$

27. Boltzmann Machine

$\Theta = \{W, L, J\}$；V：样本集合，$|V| = N = D+P$

27.1 Introduction.



$v = \{0,1\}^D$，$h = \{0,1\}^P$，分量0或1，大D大P纵列向量

$L = [L_{ij}]_{D\times D}^{(v与v)}$ $J = [J_{ij}]_{P\times P}^{(h与h)}$，$W = [W_{ij}]_{D\times P}$（h与v的权重）

$v_1 v_1, \cdots, v_3 v_3$ 连

	v_1	v_2	v_3
v_1	0	0	△
v_2		0	△
v_3			0

上三角+下角 无向图，乘上1/2倍

暴力：$\sum_{i=1}^{3}\sum_{j=1}^{3} v_i W_{ij} v_j = 3\times 3 = 9$ 项

实际上用 $(9-3)\times\frac{1}{2} = 3$ 项。

L与J对角线上为0，自己与自己不会相连

$P(v,h) = \frac{1}{Z}\exp\{-E(h,v)\}$

$E(h,v) = -(v^T \cdot W \cdot h + \frac{1}{2} v^T \cdot L \cdot v + \frac{1}{2} h^T \cdot J \cdot h)$ ← L

$Z = \int_h \int_v \exp\{-E(h,v)\}\, dv\, dh$

27.2 Gradient of log-likelihood

$P(v) = \sum_h P(v,h)$

$\frac{1}{N}\sum_{v\in V} \log P(v)$ ← log-likelihood

$\frac{\partial}{\partial\theta}\cdot\frac{1}{N}\sum_{v\in V}\log P(v) = \frac{1}{N}\sum_{v\in V}\frac{\partial \log P(v)}{\partial\theta}$ ← gradient of log-likelihood

$$\frac{\partial \log P(v)}{\partial\theta} = \sum_v\sum_h P(v,h)\cdot\frac{\partial E(v,h)}{\partial\theta} - \sum_h P(h|v)\frac{\partial E(h,v)}{\partial\theta} \quad P68$$

$$\frac{\partial \log P(v)}{\partial W} = \sum_v\sum_h P(v,h)\cdot(-vh^T) - \sum_h P(h|v)\cdot(-vh^T)$$

$$= \sum_h P(h|v)\cdot vh^T - \sum_v\sum_h P(v,h)\cdot vh^T$$

$$\therefore \frac{1}{N}\sum_{v\in V}\frac{\partial \log P(v)}{\partial W} = \frac{1}{N}\sum_{v\in V}\sum_h P(h|v)\cdot vh^T - \frac{1}{N}\sum_{v\in V}\sum_v\sum_h P(v,h)\cdot vh^T \quad *V是样本集合$$

看作是关于v,h的function，v and h已经被积掉，与$\sum_{v\in V}$来说只是一个常数

L,J求法与W类似

$\frac{1}{N}\cdot\sum_{v\in V}$ = $P_{data}(v)$ = $\frac{1}{N}\sum_{v\in V}\sum_h P(h|v)\cdot vh^T - \sum_v\sum_h P(v,h)\cdot vh^T$ → $E_{v\sim P_{data},h\sim P_{model}(h|v)}[vh^T] = E_{P_{data}}[vh^T]$

从数据来数据分布

$= E_{P_{data}}[vh^T] - E_{P_{model}}[vh^T]$

梯度只与vh^T有关，有MCMC sampling比较干净

这是P_{data}隐式包含$P_{model}(h|v)$，但h的分布由模型给定

当样本V iid于 $P_{data}(v)$时，$P_{data} = P_{data}(v)\cdot P_{model}(h|v)$

$\frac{1}{N}\sum_{v\in V}$ 当 $E_{v\sim P_{data}}$

$P_{model} = P_{model}(v,h)$

能直接得到 P_{data}

需要利用模型分布 P_{model}

观测+隐变量

$P(v,h)$联合可用 P_{model}

并且 $P(v,h)$ 为 $P_{model}(v,h)$

而因果判断 $P_{model}(v|h) = \frac{P(v,h)}{P(h)}$ 可知 $P(v,h)$, $P(v|h)$, $P(h|v)$

即 $P(v|h)$是 $P_{model}(v,h)$模型内含的条件分布

27.3 Gradient ascend based on MCMC

$$\Delta W = \alpha(E_{P_{data}}[vh^T] - E_{P_{model}}[vh^T])$$

$$\Delta L = \alpha(E_{P_{data}}[vv^T] - E_{P_{model}}[vv^T])$$

$$\Delta J = \alpha(E_{P_{data}}[hh^T] - E_{P_{model}}[hh^T])$$

$P_{data} = P_{data}(v,h) = P_{data}(v)\cdot P_{model}(h|v)$

$P_{model} = P_{model}(v,h)$，$\alpha$：步长

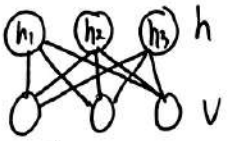
上述ΔW, ΔL, ΔJ都为矩阵

$$\Delta W_{ij} = \alpha(E_{P_{data}}[v_i h_j] - E_{P_{model}}[v_i h_j])$$

positive phase　negative phase

Both hard, intractable

MCMC (1983-1985 Hinton)



RBM下情况利用公式

RBM: $P(h|v) = \prod_{j=1}^{p} P(h_j|v)$

$\therefore P(h_j=1|v) = P(h_j=1|v,h_{-j}) = \sigma(\sum_{i=1}^{n} W_{ij}$

其中，$\sum_{m=1|j}^{p} J_{jm}h_m = 0$，即h之间没有连接

规模大效率特别差

利用Boltzmann Machine的定义，P14的$P(v,h)$与$E(h,v)$

推导出：(假使)证明

$$P(v_i=1|h,v_{-i}) = \sigma(\sum_{j=1}^{p} W_{ij}h_j + \sum_{k=1|i}^{D} L_{ik}v_k)$$

$$P(h_j=1|v,h_{-j}) = \sigma(\sum_{i=1}^{D} W_{ij}v_i + \sum_{m=1|j}^{p} J_{jm}h_m)$$

σ → Sigmoid → $\frac{1}{1+e^{-x}}$

为MCMC服务

P75

这个是P76推导出的

而sigmoid Belief network中的sigmoid function是人为定义的。

27.4 Conditional Probability

推导: $P(v_i=1 \mid h, v_{-i}) = \sigma\left(\sum_{j=1}^{p} W_{ij} h_j + \sum_{k=1\backslash i}^{D} L_{ik} v_k\right)$ $\quad \sigma = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$ sigmoid

$P(h_j=1 \mid v, h_{-j}) = \sigma\left(\sum_{i=1}^{D} W_{ij} v_i + \sum_{m=1\backslash j}^{p} J_{jm} h_m\right)$

$$P(v_i \mid h, v_{-i}) = \frac{P(v,h)}{P(h,v_{-i})} = \frac{\frac{1}{Z}\exp\{-E(h,v)\}}{\sum_{v_i}\frac{1}{Z}\exp\{-E(h,v)\}} = \frac{\exp\{v^T W h + \frac{1}{2}v^T L v + \frac{1}{2}h^T J h\}}{\sum_{v_i}\exp\{v^T W h + \frac{1}{2}v^T L v + \frac{1}{2}h^T J h\}}$$

$$= \frac{\exp\{\frac{1}{2}h^T J h\}\cdot\exp\{v^T W h + \frac{1}{2}v^T L v\}}{\exp\{\frac{1}{2}h^T J h\}\cdot\sum_{v_i}\exp\{v^T W h + \frac{1}{2}v^T L v\}} = \frac{\exp\{v^T W h + \frac{1}{2}v^T L v\}}{\exp\{v^T W h + \frac{1}{2}v^T L v\}\big|_{v_i=0} + \exp\{v^T W h + \frac{1}{2}v^T L v\}\big|_{v_i=1}}$$

对v进行实例化令 $v_{\neq i}$ 与 v_i 一样, 与 v_i 一开始就无关, 被积掉, 所以只有 v_i 改变

$$= \frac{\exp\{v^T W h + \frac{1}{2}v^T L v\}\big|_{v_i=1}}{\exp\{v^T W h + \frac{1}{2}v^T L v\}\big|_{v_i=0} + \exp\{v^T W h + \frac{1}{2}v^T L v\}\big|_{v_i=1}} = \frac{\Delta v_{i=1}}{\Delta v_{i=0} + \Delta v_{i=1}}$$

$P(v_i=1 \mid h, v_{-i})$

$$\Delta v_i = \exp\{v^T W h + \frac{1}{2}v^T L v\} = \exp\left\{\sum_{\hat{i}=1}^{D}\sum_{j=1}^{p} v_{\hat{i}} W_{\hat{i}j} h_j + \frac{1}{2}\sum_{\hat{i}=1}^{D}\sum_{k=1}^{D} v_{\hat{i}} L_{\hat{i}k} v_k\right\}$$

把这个拆为4个: $\hat{i}\neq i, k\neq i$; $\hat{i}=i, k=i$; $\hat{i}\neq i, k=i$; $\hat{i}=i, k\neq i$

$$= \exp\left\{\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{p} v_{\hat{i}} W_{\hat{i}j} h_j + \sum_{j=1}^{p} v_i W_{ij} h_j + \frac{1}{2}\Big(\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1\backslash i}^{D} v_{\hat{i}} L_{\hat{i}k} v_k + \underbrace{v_i L_{ii} v_i}_{L_{ii}=0,\ i=0,\ \hat{i},k=i} + \underbrace{\sum_{\hat{i}=1\backslash i}^{D} v_{\hat{i}} L_{\hat{i}i} v_i}_{k=i,\ \hat{i}\neq i} + \underbrace{\sum_{k=1\backslash i}^{D} v_i L_{ik} v_k}_{k\neq i,\ \hat{i}=i}\Big)\right\}$$

$\hat{i}$与k不$=i$

二者相同, 以对称

$$= \exp\left\{\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{p} v_{\hat{i}} W_{\hat{i}j} h_j + \sum_{j=1}^{p} v_i W_{ij} h_j + \frac{1}{2}\Big(\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1\backslash i}^{D} v_{\hat{i}} L_{\hat{i}k} v_k + 2\sum_{k=1\backslash i}^{D} v_i L_{ik} v_k\Big)\right\}$$

$$\Delta_{v_i=0} = \exp\left\{\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{p} v_{\hat{i}} W_{\hat{i}j} h_j + \sum_{j=1}^{p} v_i W_{ij} h_j + \frac{1}{2}\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1\backslash i}^{D} v_{\hat{i}} L_{\hat{i}k} v_k + \sum_{k=1\backslash i}^{D} v_i L_{ik} v_k\right\}$$

$$\Delta_{v_i=0} = \exp\left\{\sum_{\hat{i}=1\backslash i}^{D}\sum_{j=1}^{p} v_{\hat{i}} W_{\hat{i}j} h_j + \frac{1}{2}\sum_{\hat{i}=1\backslash i}^{D}\sum_{k=1\backslash i}^{D} v_{\hat{i}} L_{\hat{i}k} v_k\right\} = \exp\{A+B\}$$

$$\Delta_{v_i=1} = \exp\left\{A + B + \sum_{j=1}^{p} W_{ij} h_j + \sum_{k=1\backslash i}^{D} L_{ik} v_k\right\}$$

$$\therefore P(v_i=1 \mid h, v_{-i}) = \frac{\Delta v_{i=1}}{\Delta v_{i=0} + \Delta v_{i=1}} = \frac{\exp\left\{A + B + \sum_{j=1}^{p} W_{ij} h_j + \sum_{k=1\backslash i}^{D} L_{ik} v_k\right\}}{\exp\{A+B\} + \exp\left\{A + B + \sum_{j=1}^{p} W_{ij} h_j + \sum_{k=1\backslash i}^{D} L_{ik} v_k\right\}}$$

sigmoid: $\frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$

$$= \frac{\exp\left\{\sum_{j=1}^{p} W_{ij} h_j + \sum_{k=1\backslash i}^{D} L_{ik} v_k\right\}}{1 + \exp\left\{\sum_{j=1}^{p} W_{ij} h_j + \sum_{k=1\backslash i}^{D} L_{ik} v_k\right\}} = \sigma\left(\sum_{j=1}^{p} W_{ij} h_j + \sum_{k=1\backslash i}^{D} L_{ik} v_k\right)$$

27.5 Variational Inference (Hinton)

变量代换说明在P14

根据P10类似: $\mathcal{L} = \text{上LBO} = \log P_\theta(v) - KL(q_\phi \| P_\theta) = \sum_h q_\phi(h|v) \cdot \log P_\theta(v,h) + H[q]$

基于平均场 $q_\phi(h|v) = \prod_{j=1}^{p} q_\phi(h_j|v)$, $q_\phi(h_j=1|v) = \phi_j$, $q_\phi(h_j=0|v) = 1-\phi_j$, $\phi = \{\phi_j\}_{j=1}^{p}$

$$\hat{\phi}_j = \arg\max_{\phi_j} \mathcal{L} = \arg\max_{\phi_j} \sum_h q_\phi(h|v)\left[-\log Z + v^T \cdot W \cdot h + \frac{1}{2} v^T \cdot L \cdot v + \frac{1}{2} h^T J h\right] + H[q]$$

$$= \arg\max_{\phi_j} \underbrace{\sum_h q_\phi(h|v)}_{1} \cdot \underbrace{\left[-\log Z + \frac{1}{2} v^T \cdot L \cdot v\right]}_{\text{与 } \phi_j, h \text{ 都无关}} + \sum_h q_\phi(h|v)\left[v^T \cdot W \cdot h + \frac{1}{2} h^T J h\right] + H[q]$$

$$= \arg\max_{\phi_j} \sum_h q_\phi(h|v)\left[v^T W h + \frac{1}{2} h^T J h\right] + H[q]$$

$$= \arg\max_{\phi_j} \sum_h q_\phi(h|v) \cdot v^T W h + \frac{1}{2} \sum_h q_\phi(h|v) \cdot h^T J h + H[q] = \arg\max_{\phi_j} ① + ② + ③$$

$$① = \sum_h q_\phi(h|v) \cdot \sum_{i=1}^{D} \sum_{j=1}^{P} v_i W_{ij} h_j = \sum_h \prod_{j=1}^{P} q_\phi(h_j|v) \cdot \sum_{i=1}^{D} \sum_{j=1}^{P} v_i W_{ij} h_j$$

其中一项: $\sum_h \prod_{j=1}^{P} q_\phi(h_j|v) \cdot v_1 W_{12} h_2 = \sum_{h_2} q_\phi(h_2|v) \cdot v_1 W_{12} h_2 \cdot \underbrace{\sum_{h \setminus h_2} \prod_{j=1 \setminus 2}^{P} q_\phi(h_j|v)}_{\sum_{h_1} q_\phi(h_1|v) \cdot \sum_{h_3} q_\phi(h_3|v) \cdot \cdots \cdot \sum_{h_P} q_\phi(h_P|v) = 1}$

$$= \sum_{h_2} q_\phi(h_2|v) \cdot v_1 W_{12} h_2$$

$\because h_2 = \{0,1\}$

$\therefore = q_\phi(h_2=1|v) \cdot v_1 W_{12} = \phi_2 \cdot v_1 W_{12}$

$\therefore ① = \sum_{i=1}^{D} \sum_{j=1}^{P} v_i W_{ij} \phi_j$

$i=j, j=j \Rightarrow J_{ij}=0$

$i=j, j\neq j$ } 2者等值，故与 $\frac{1}{2}$ 相消

$i\neq j, j=j$

$i\neq j, j\neq j \rightarrow$ 常数 C

在上述前提下，忽略主后的其中一项。

对称性

$$② = \frac{1}{2} \sum_h q_\phi(h|v)\, h^T J h = \frac{1}{2} \sum_h \prod_{j=1}^{P} q_\phi(h_j|v) \cdot \sum_{i=1}^{P} \sum_{j=1}^{P} h_i J_{ij} h_j$$

其中一项: $\sum_h \prod_{j=1}^{P} q_\phi(h_j|v) \cdot h_1 J_{12} h_2 = \sum_{h_1} \sum_{h_2} q_\phi(h_1|v) q_\phi(h_2|v) \cdot h_1 J_{12} h_2 \cdot \underbrace{\sum_{h \setminus h_1 h_2} \prod_{j=3}^{P} q_\phi(h_j|v)}_{1}$

$$\therefore = \sum_{h_1} \sum_{h_2} q_\phi(h_1|v) q_\phi(h_2|v) \cdot h_1 J_{12} h_2$$

$\because h_1, h_2 = \{0,1\}$

$\therefore (0,0) \times ; (0,1) \times ; (1,0) \times ; (1,1) \checkmark$

$\therefore = q_\phi(h_1=1|v)\, q_\phi(h_2=1|v)\, J_{12} = \phi_1 \phi_2 J_{12}$

$\therefore ② = \sum_{j=1}^{P} \sum_{m=1 \setminus j}^{P} \phi_j \phi_m J_{jm}$ (要扣去其中一个 h 的值)

because 原因

注意，这里 ϕ_j 与其它 ϕ_m 有关，$\sum_{m=1 \setminus j}^{P}$

故一个一个的求，一轮一轮目标：坐标上升

σ: Sigmoid: $\frac{1}{1+e^{-x}}$ $(x_1, x_2, \cdots, x_D)$ 一个一个输出

$$③ = -\sum_j^P \left[\phi_j \log \phi_j + (1-\phi_j) \log(1-\phi_j)\right]$$

$$\frac{\partial ②}{\partial \phi_j} = \sum_{i=1}^{P} v_i W_{ij} \quad ; \quad \frac{\partial ③}{\partial \phi_j} = \sum_{m=1 \setminus j}^{P} \phi_m J_{jm} \quad ; \quad \frac{\partial ③}{\partial \phi_j} = \log \frac{\phi_j}{1-\phi_j}$$

$$\frac{\partial [① + ② + ③]}{\partial \phi_j} \triangleq 0$$

$$\phi_j = \sigma\left(\sum_{i=1}^{D} v_i W_{ij} + \sum_{m=1 \setminus j}^{P} \phi_m J_{jm}\right)$$

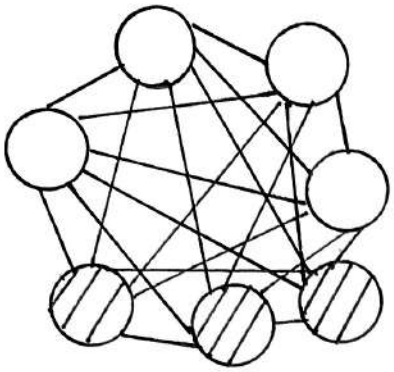
$\hat{\phi} = \{\hat{\phi}_j\}_{j=1}^{P}$ 不动点方程 解法⇒坐标上升

$q_\phi(h|v) \approx P_{model}(h|v)$ 求后验

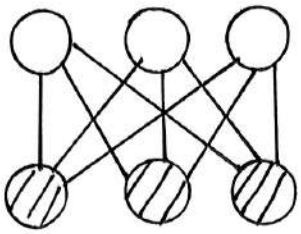
P77

28. Deep Boltzmann Machine

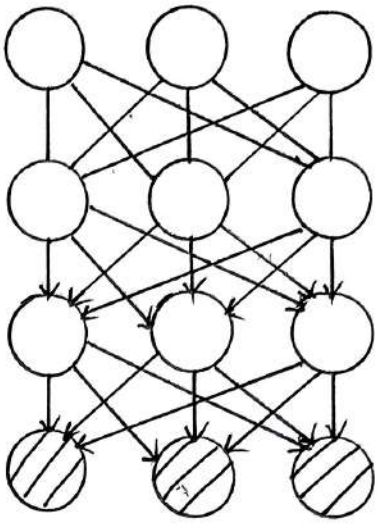
28.1 Introduction



General Boltzmann Machine 1983



RBM 1986, 2007



DBN 2006

SGA = stochastic Gradient Ascend

$\Delta W = \alpha(E_{P_{data}}[vh^T] - E_{P_{model}}[vh^T])$

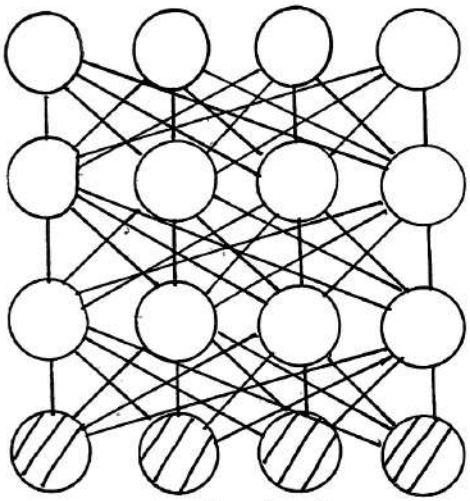
positive phase　negative phase

$P_{data} = P_{data}(v,h) = P_{data}(v) \cdot P_{model}(h|v)$

variational Inference P77

$P_{model} = P_{model}(v,h)$ → PCD (持续对比散度)

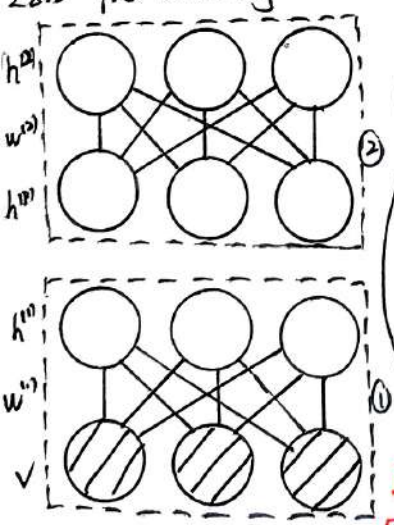
- Pre-training (stacking RBM)
- Fine-tuning
  - wake-sleep
  - BP



DBM 2008

- Pre-training (stacking RBM)
- SGA

28.2 pre-training



P74

在DBM中，直接试 $W^{(1)}$ 给 $W^{(1)}$，在 $W^{(1)}$ 基础上训练，用 $P(h^{(1)};W^{(2)})$ 去近似表达 $P(h^{(1)};W^{(1)},W^{(2)})$ → 用 $P(h^{(1)};W^{(1)})$ or $P(h^{(1)};W^{(2)})$ 表示 $P(h^{(1)};W^{(1)},W^{(2)})$ 都好；但不准，下节介绍如何准确

$P(v) = \sum_{h^{(1)}} P(v,h^{(1)}) = \sum_{h^{(1)}} P(h^{(1)}) \cdot P(v|h^{(1)})$

右侧中：$P(v) = \sum_{h^{(1)}} P(v,h^{(1)})$　$P(h^{(1)}) = \sum_{v} P(v,h^{(1)})$

$P(v,h^{(1)})$ 与 $W^{(1)}$ 有关，进而表达结束，在①中在②中与 $W^{(1)}, W^{(2)}$ 有关

①②加，连接起来

∵ $h^{(1)}$ 与 W 有关

∴ $P(v) = \sum_{h^{(1)}} P(h^{(1)};W^{(1)}) \cdot P(v|h^{(1)};W^{(1)})$，P70. CD-k 求出 W

对 $h^{(1)}$ sampling：

$P(h^{(1)}|v,W^{(1)}) = \prod_{i=1}^{3} P(h_i^{(1)}|v;W^{(1)})$ 在RBM中同时解 P62

对比时①中 $h^{(1)}$ 采样，采样后的数据在②中当作 v 重复上述①的步骤，比如得到 $v^1, h^{(1)}$

$P(h^{(1)};W^{(2)}) = \sum_{h^{(2)}} P(h^{(1)},h^{(2)};W^{(2)})$

The $h^{(1)}$:

$P(h^{(1)}) = \sum_{h^{(2)},v} P(v,h^{(1)};h^{(2)})$　由 $W^{(1)}, W^{(2)}$ 表示

↓

$P(h^{(1)};W^{(1)},W^{(2)})$ → $h^{(1)}$ 与 $W^{(1)}, W^{(2)}$ 有关

Intuition：

用 $P(h^{(1)};W^{(1)})$ 和 $P(h^{(1)};W^{(2)})$ 几何平均近似 $P(h^{(1)};W^{(1)},W^{(2)})$

P78

28.3 double counting problem

V：样本集合 $v \in V$

H：采样样本集合 $h^{(1)} \in H$，$h^{(1)}$ 依赖于 v

真正的：$P(h^{(1)}; w^{(1)}, w^{(2)})$

直觉：同时利用 $P(h^{(1)}; w^{(2)})$ 和 $P(h^{(1)}; w^{(1)})$ 去近似 $P(h^{(1)}; w^{(1)}, w^{(2)})$，对于 $v \in V$

$$P(h^{(1)}; w^{(1)}) = \sum_{v} P(v, h^{(1)}; w^{(1)}) = \sum_{v} P(v) \cdot P(h^{(1)} \mid v; w^{(1)}) \approx \frac{1}{N} \sum_{v \in V} P(h^{(1)} \mid v; w^{(1)})$$

Aggregated posterior

$$P(h^{(1)}; w^{(2)}) = \sum_{h^{(2)}} P(h^{(1)}, h^{(2)}; w^{(2)}) \overset{?}{=} \sum_{h^{(2)}} P(h^{(2)}) \cdot P(h^{(1)} \mid h^{(2)}; w^{(2)}) \approx \frac{1}{N} \sum_{h^{(2)} \in H} P(h^{(1)} \mid h^{(2)}; w^{(2)}) \quad ?$$

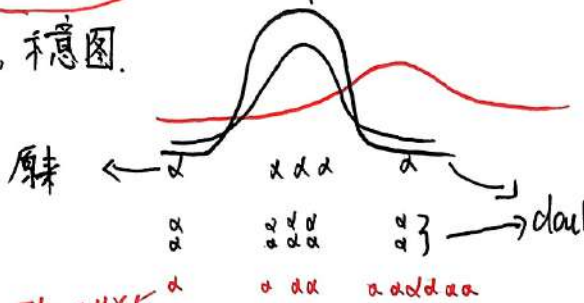
$$E_{P(v)}[P(h^{(1)} \mid v; w^{(1)})] \overset{\text{MC sampling}}{\approx} \frac{1}{N} \sum_{v \in V} P(h^{(1)} \mid v; w^{(1)})$$

用的 V　　用的 H

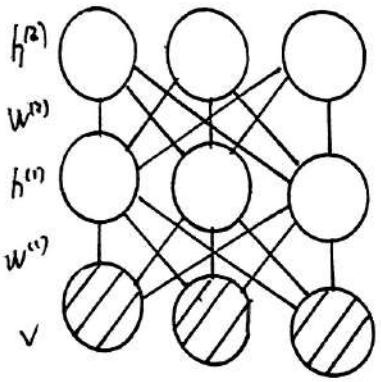
$\therefore P(h^{(1)}; w^{(1)}) \approx \frac{1}{N} \sum_{v \in V} P(h^{(1)} \mid v; w^{(1)})$；$P(h^{(1)}; w^{(2)}) \approx \frac{1}{N} \sum_{h^{(2)} \in H} P(h^{(1)} \mid h^{(2)}; w^{(2)})$，若简单相加

且 v 依赖于 V，$h^{(1)}$ 也依赖于 V，由此可知，若层数增加，$h^{(1)}$ 仍依赖于 V，会重复计算

v 与 H 都从 V 来的，double counting，示意图



$v, w^{(1)}, h^{(1)}, h^{(2)}, w^{(2)}$ 合并，P78 中282图熟



? $P(h^{(1)}; w^{(2)})$ 的表示存在疑问 $\sum_{h^{(2)}} P(h^{(2)}) P(h^{(1)} \mid h^{(2)}; w^{(2)})$ 不合理 ✗

应为 $\sum_{h^{(2)}} P(h^{(2)}) \cdot P(h^{(1)} \mid h^{(2)}; w^{(2)})$

但 $P(h^{(1)} \mid h^{(2)}; w^{(2)})$ 并未明确该如何计算

28.4 Pre-training continue

True: $P(h^{(1)}; w^{(1)}, w^{(2)})$

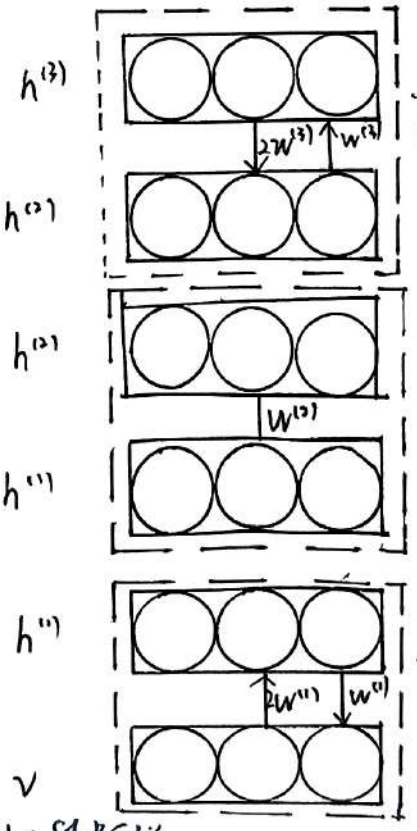
$P(h^{(1)}; w^{(1)}) = \sum_{v} P(v) \cdot P(h^{(1)} \mid v; w^{(1)})$ → DBM

$P(h^{(1)}; w^{(2)}) = \sum_{h^{(2)}} P(h^{(2)}) \cdot P(h^{(1)} \mid h^{(2)}; w^{(2)})$ → DBN

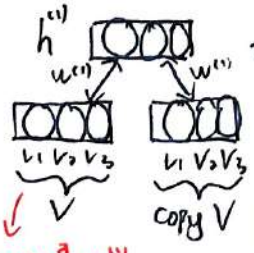
很少用 Pre-training

↓

现如今很少有人用；联合学习可以解决学习问题。



相当于 4 层 DBM



看作 ⇐

用块 Gibbs 联合采样

$h^{(1)} \to v$ 是 $w^{(1)}$

v, copy(v) $\to h^{(1)}$ 是 $2w^{(1)}$

29. Generative Model

29.1 Supervised learning vs unsupervised learning

target (二者对比) (分类, 回归, 标记, 降维, 聚类, 特征学习, 密度估计, 生成数据)

监督
- 概率模型
  - 判别模型 ($P(Y|X)$): LR, MEMM, CRF
  - 生成模型
- 非概率模型: PLA, SVM, KNN, NN, Tree Model

非监督
- 概率模型: 生成模型
- 非概率模型: PCA, LSA, k-means, Auto-encoder

深度生成模型 (加入/涉入 Deep Learning)

Energy-based Model: Boltzmann Machine

VAE

GAN / GSN

Autoregressive Model

Flow-based Model

传统生成模型:

Naive Bayes

Mixture Model: GMM

Time-Series Model:

HMM | Kalman Filter | Particle Filter

Non-parameteric:

Bayesian Model: Gauss Process, Dirichlet process

Mixed membership model: LDA

Factorial Model: Factorial ~~Analysis~~ Analysis, P-PCA, ICA,

生成模型: 关注样本分布本身

PCA → P-PCA → FA

k-means → GMM

Auto-encoder → VAE

LSA → PLSA → LDA

29.2 Presentation & Inference & Learning

Presentation: 形神兼备:

- 形
  - Discrete vs Continuous
  - Directed vs UnDirected Model
  - Latent Variable Model vs Fully-observed Model
  - Shallow vs Deep (层)
  - ~~稀疏~~ Sparse vs Dense (连接) 稠密 HMM vs BM
- 神
  - Parameteric Model vs Non-parameteric
  - Implicit Density vs Explicit Density GAN

Implicit Density: 不等于 $P(A)$

Explicit Density: 等于 $P(A)$ 分布计论

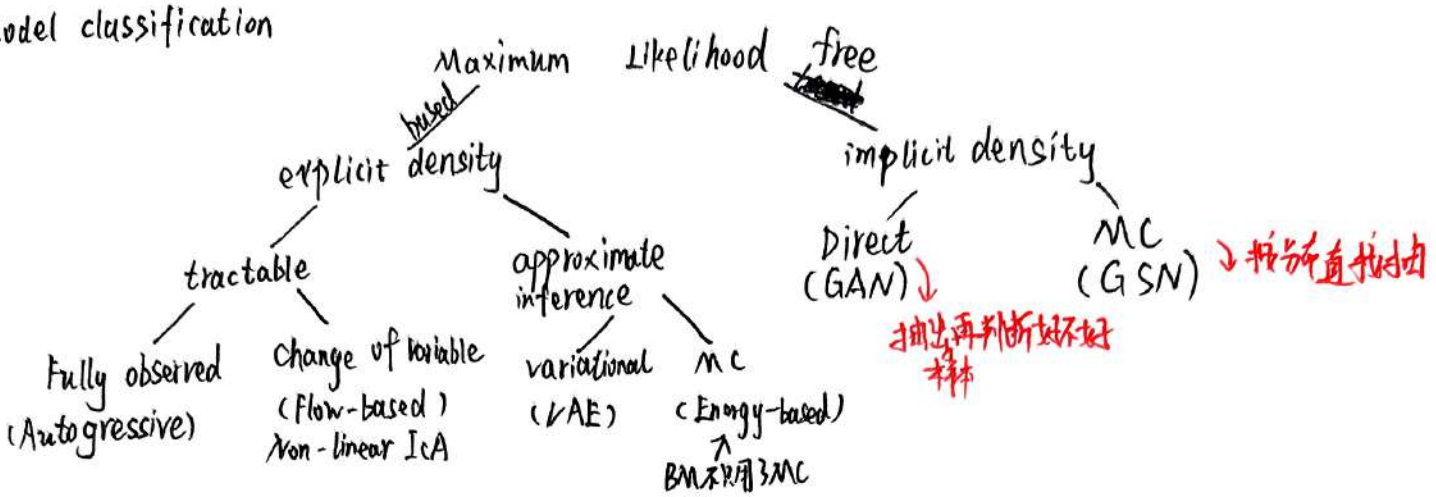
Inference:

tractable v.s intractable

Learning:

likelihood-based Model vs. likelihood-free Model GAN

29.3 Model classification



29.4 probabilistic graph vs Neural network (Bayesian Network vs Neural Network)

Probabilistic Graph => $P(x)$ 的表示

Neural Network => 函数逼近器

PG	NN
数：结构化的	深层、稠密
稀疏，浅层	计算图
条件独立假设	可解释性未知
具备可解释性	

Inference：精确 PG 近似(MC，变分) NN 推断很容易，但没意义(跟W后)

Learning：Likelihood maximum estimate (MLE)

适合：high level reasoning

梯度下降(BP) → 高效的求导方法 { 链式求导法则 / 动态规划 / 递归+缓存 }

表示学习，low level reasoning

(以连结主义) Boltzmann Machine { 无向图模型 / 神经网络 }

神经网络 { 右角定性 / 随机性 }

29.5 Reparametrization Trick

假设：① 概率分布 $P(y)=N(\mu,\sigma^2)$ 实际上可能是一个很复杂的分布

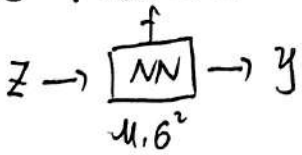
$z\sim N(0,1)$，$\theta=\{\mu,\sigma^2\}$

$y=\mu+\sigma\cdot z$

$z^{(i)}\sim N(0,1)$ z具有随机性

$y^{(i)}=\mu+\sigma\cdot z^{(i)}$ 随机性转移到 $y^{(i)}$ 上

$y=f(\mu,\sigma,z)$



$J(y)$：目标函数

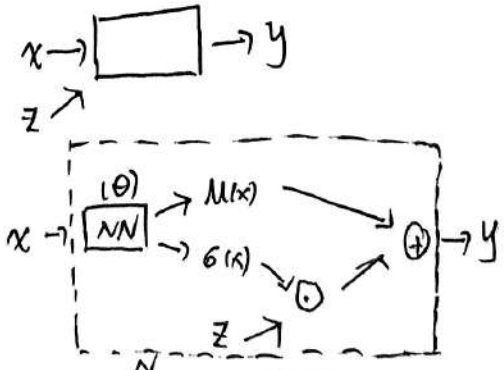
$$\frac{\nabla J(y)}{\nabla \theta}=\frac{\nabla J(y)}{\nabla y}\cdot\frac{\nabla y}{\nabla \theta}$$

② 条件概率分布

$P(y|x)=N(x;\mu,\sigma^2)$

$z\sim N(0,1)$

$y=\mu_\theta(x)+\sigma_\theta(x)\cdot z$

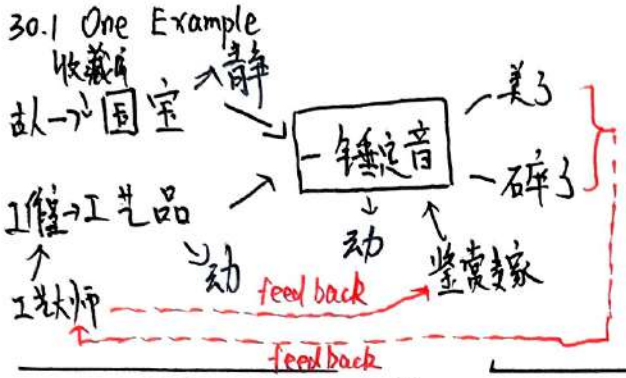


$$J_\theta(y)=\sum_{i=1}^{N}\|y-y^{(i)}\|^2$$

$$\frac{\nabla J_\theta(y)}{\nabla \theta}=\frac{\nabla J_\theta(y)}{\nabla y}\frac{\nabla y}{\nabla \mu}\frac{\nabla \mu}{\nabla \theta}+\frac{\nabla J_\theta(y)}{\nabla y}\frac{\nabla y}{\nabla \sigma}\frac{\nabla \sigma}{\nabla \theta}$$

30. Generative Adversarial Network → Implicit Density Model

30.1 One Example



目标：成为高水平，可以以假乱真的大师.

双高 { 高水平的鉴赏专家 → 手段
高水平的工艺品大师 → 目标

↳ 有高专家是成为高大师的先决条件

30.2 Model representation

国宝：$\{x_i\}_{i=1}^N$: P_{data}

工艺品：$P_g(x;\theta_g)$ ：假设 $z \sim P_z(z)$ 找一个容易的分布

generator

$z \to$ NN $\to x$ NN

$x = G(z;\theta_g)$

$G(z;\theta_g)$

通过 $P_g(x;\theta_g)$，用可微生成网络 $G(z;\theta_g)$ 去逼近 P_g，从采样角度 → Implicit Density

鉴赏专家：$D(x;\theta_d)$：代表 x 是真品(国宝)的概率

P_{data} → NN → Y/N

$z \to$ NN → Y/N

高专家：{ if x is from P_{data}, then $D(x)\uparrow$ $\log D(x)\uparrow$
{ if x is from P_g, then $D(x)\downarrow$ / $1-D(x)\uparrow$ $\log(1-D(G(z;\theta_g)))\uparrow$
(z is from $P_z(z)$)

$$\max_D E_{x\sim P_{data}}[\log D(x)] + E_{z\sim P_z}[\log(1-D(G(z;\theta_g)))]$$

高大师：if x is from P_g, then $D(G(z;\theta_g))\uparrow$ / $\log[1-D(G(z;\theta_g))]\downarrow$

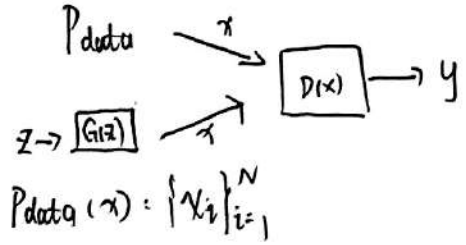
$$\min_G E_{z\sim P(z)}[\log(1-D(G(z;\theta_g)))]$$

总目标：$\min_G \max_D E_{x\sim P_{data}}[\log D(x)] + E_{z\sim P_z}[\log(1-D(G(z;\theta_g)))]$ $P_g \to P_{data}$

先让骗成为高专家，再让大师成为高大师

⇕ 恒等？为何

30.3 Global optimality



$P_{data}(x) = \{x_i\}_{i=1}^N$

$P_g(x;\theta_g)$: generator $(P_z(z) + G(z;\theta_g))$

$y|x$: discriminator 判别器

$y\|x$	正真 1	假 0
P	D(x)	1-D(x)

记 $V(D,G) = E_{x\sim P_{data}}[\log D(x)] + E_{x\sim P_g}[\log(1-D(x))]$

For fixed G, 求 $\max_D V(D,G)$

$$\max_D V(D,G) = \max_D \int P_{data}\cdot\log D\,dx + \int P_g\cdot\log(1-D)\,dx$$

$$= \max_D \int [P_{data}\cdot\log D + P_g\cdot\log(1-D)]\,dx$$

$$\frac{\partial}{\partial D}(\max_D V(D,G)) = \int \frac{\partial}{\partial D}[P_{data}\cdot\log D + P_g\cdot\log(1-D)]\,dx$$

$$= \int [P_{data}\cdot\frac{1}{D} + P_g\cdot\frac{-1}{1-D}]\,dx \triangleq 0$$

$$\Rightarrow D_G^* = \frac{P_d}{P_d+P_g}$$

将 D_G^* 代入 $V(D,G)$，则有：

$$\min_G \max_D V(D,G) = \min_G V(D_G^*, G) = \min_G E_{x\sim P_{data}}\left[\log \frac{P_d}{P_d+P_g}\right] + E_{x\sim P_g}\left[\log\left(1-\frac{P_d}{P_d+P_g}\right)\right]$$

$$= \min_G E_{x\sim P_d}\left[\log \frac{P_d}{P_d+P_g}\right] + E_{x\sim P_g}\left[\log \frac{P_g}{P_d+P_g}\right] \longrightarrow$$

与 KL 散度很像

$KL(q\|p) = q(x)\log\frac{q(x)}{p(x)} = E_{q(x)}\left[\log\frac{q(x)}{p(x)}\right]$

但 P_d+P_g 两个概率密度函数相加属于 $[0,2]$ 3. 不是概率密度函数，稍加处理

$$= \min_G E_{x\sim P_d}\left[\log \frac{P_d}{(P_d+P_g)/2}\cdot\frac{1}{2}\right] + E_{x\sim P_g}\left[\log\frac{P_g}{(P_d+P_g)/2}\cdot\frac{1}{2}\right]$$

$E(\log\frac{1}{2}) = -\log 2$

$$= \min_G KL\left(P_d \,\|\, \frac{P_d+P_g}{2}\right) + KL\left(P_g \,\|\, \frac{P_d+P_g}{2}\right) - \log 4$$

KL 散度非负

$$\geq -\log 4$$

当 $P_d = \frac{P_d+P_g}{2} = P_g$ 时，"=" 成立

故 $P_g^* = P_d$，$D_G^* = \frac{1}{2}$

$D_G^* = \frac{1}{2}$，专家已经无法一锤定音了

---

31. Variational Autoencoder

31.1 representation

Latent Variable Model

$z \rightarrow x$

GMM：k 个 Gaussian Dist 混合：$z \sim$ Categorical Dist（离散）

z	1	2	…	k
p	p_1	p_2	…	p_k

$\sum_{i=1}^{k} p_i = 1$　　$x|z=i \sim N(x|\mu_i, \Sigma_i)$

VAE：infinite Gaussian Dist：$\begin{cases} z \sim N(0, I) \\ x|z \sim N(\mu_\theta(z), \Sigma_\theta(z)) \end{cases}$

$P_\theta(x) = \int_z P_\theta(x,z)\,dz = \int_z P_\theta(z)\cdot P_\theta(x|z)\,dz$

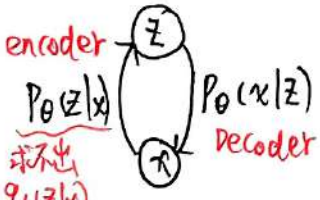
intractable

$P_\theta(z|x) = \frac{P_\theta(z)P_\theta(x|z)}{P_\theta(x)}$

intractable　无法直接求出 $P_\theta(z|x)$

31.2 Learning

Latent Variable Model



求不出 $q_\phi(z|x)$

$P(z) = N(0, I)$

$P_\theta(x|z) = N(\mu_\theta(z), \Sigma_\theta(z))$

$P_\theta(z|x)$ is intractable 用 $q_\phi(z|x)$ 逼近

$\log P(x) = ELBO + KL(q_\phi(z|x) \,\|\, P_\theta(z|x))$

EM：E-step：当 $q = P_\theta(z|x)$ 时，$KL = 0$，$\arg\max_\theta P_\theta(x) = \arg\max_\theta ELBO$

expectation is ELBO

M-step：$\hat{\theta} = \arg\max_\theta ELBO = \arg\max_\theta E_{P_\theta(z|x)}[\log P_\theta(x,z)]$

$$\langle \hat{\theta}, \hat{\phi} \rangle = \arg\min KL(q_\phi(z|x) \,\|\, P_\theta(z|x))$$

$$= \arg\max ELBO$$

$$= \arg\max E_{q_\phi(z|x)}[\log P_\theta(x,z)] + H[q_\phi]$$

$$\langle\hat{\theta},\hat{\phi}\rangle = \arg\max \mathop{E}_{q_\phi(z|x)}\left[\log P_\theta(x|z)\right] - KL(q_\phi(z|x)\,\|\,P_\theta(z))$$

先从 $z^{(i)} \sim q_\phi(z|x)$ 采样，再固定 $z^{(i)}$，求 $\log P_\theta(x|z^{(i)})$

先 encoder 再 decoder 闭环

采用 SGVI / SGVB / SVI / Amortized Inference, P36

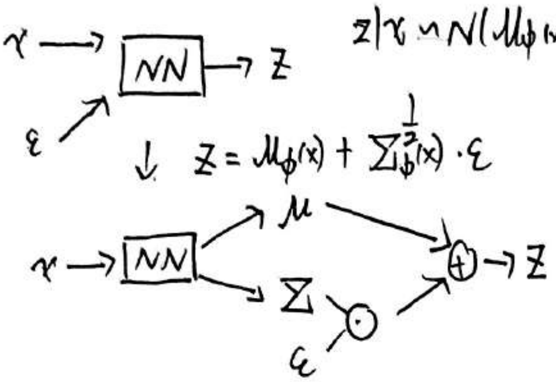
$z|x$

$\varepsilon \sim N(0,1)$

$z|x \sim N(\mu_\phi(x), \Sigma_\phi(x))$

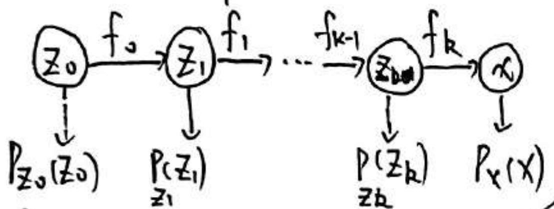
$(\mu, \sigma^2) \to (\mu, \Sigma)$

$\mu + \sigma\varepsilon \to \mu + \Sigma^{\frac{1}{2}}\varepsilon$



32. Normalizing Flow

32.1 Model Representation



Change of Variables Theorem

① Assuming $x = f(z)$, $z, x \in \mathbb{R}^p$

$z \sim P_z(z)$, $x \sim P_x(x)$

f is continuous, invertible（可逆）

$\because \int_z P(z)\,dz = 1 = \int_x P_x(x)\,dx$

$\therefore |P_z(z)\,dz| = |P_x(x)\,dx|$ $\quad P_x(x), P_z(z) > 0$

$\therefore P_x(x) = \frac{|dz|}{|dx|}\cdot P_z(z)$

$\because x = f(z)$, f is invertible

$\therefore z = f^{-1}(x)$

$\therefore P_x(x) = \left|\frac{\partial f^{-1}(x)}{\partial x}\right|\cdot P_z(z)$ 一维（一个值）

$P_x(x) = \left|\frac{\nabla f^{-1}(x)}{\nabla x}\right|\cdot P_z(z)$ p维（矩阵）

$\frac{\partial f^{-1}(x)}{\partial x}$: Jacobian Matrix

$\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)$: Jacobian Determinant

故：$P_x(x) = \left|\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)\right|\cdot P_z(z)$

$y = f(x)$

$x = f^{-1}(y)$

$\frac{dy}{dx} = \frac{\partial f(x)}{\partial x}$ ①

$\frac{dx}{dy} = \frac{\partial f^{-1}(y)}{\partial y}$ ②

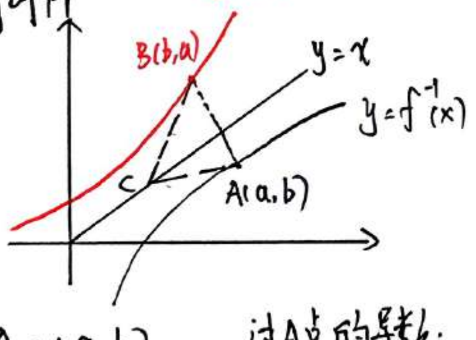
①·② $= \frac{dy}{dx}\cdot\frac{dx}{dy} = 1$

②

$$P_x(x) = \left|\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)\right| P_z(z)$$

$$= \left|\det\left(\frac{\partial f(z)}{\partial z}\right)\right|^{-1} P_z(z)$$

作图解释



A : (a, b)

B : (b, a)

C : (c, c)

过A点的导数：$f^{-1\prime}(a) = \tan\theta = \frac{b-c}{a-c}$

过B点的导数：$f'(b) = \tan\theta' = \frac{a-c}{b-c}$

$\tan\theta\cdot\tan\theta' = 1$

$f^{-1\prime}(a)\cdot f'(b) = 1$

$f'(b) = (f^{-1\prime}(a))^{-1}$

用MLE学习，$P_x(x) = \left|\det\left(\frac{\partial f(z)}{\partial z}\right)\right|^{-1} P_z(z)$

在本题Ex：$P_x(x) = \left|\det\left(\frac{\partial f(z_k)}{\partial z_k}\right)\right|^{-1} P_{z_k}(z_k)$

再用change of variables Theorem进行代换即可得一条链，即每一个节点都利用 change of Variables Theorem（代 $P_{z_k}(z_k)$ 与 $P_{z_{k-1}}(z_{k-1})$）