L1 Linear model

① $y = X^T\beta + \varepsilon$，n个带标签的p维样本

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X^T = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon \\ \vdots \\ \varepsilon_n \end{bmatrix} \sim N(0, \sigma^2 I) \text{ 噪声项}$$

预测时 $\hat{y}_i = x_i^T\beta$ 即可

② $\mathrm{logit}(p_i) = \log \frac{P(y_i=1|x_i)}{P(y_i=0|x_i)} = \log \frac{p_i}{1-p_i}$

? 假设 $\mathrm{logit}(p_i) = x_i^T\beta$ $\quad \frac{p_i}{1-p_i} = e^{x_i^T\beta} = -1 + \frac{1}{1-p_i}$ $\quad 1-p_i = \frac{1}{e^{x_i^T\beta}+1}$

$p_i = \mathrm{sigmoid}(x_i^T\beta) = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} = \frac{1}{1+e^{-x_i^T\beta}}$ $\quad \mathrm{sigmoid}(x) = \frac{1}{1+e^{-x}}$

③ 设 y_i^* 是GT标签，预测正确的概率

$$Pr(y_i = y_i^* | x_i) = \underbrace{p_i^{y_i^*}(1-p_i)^{1-y_i^*}}_{0-1\text{分布}} = \left(\frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)^{y_i^*}\left(\frac{1}{1+e^{x_i^T\beta}}\right)^{1-y_i^*} = \frac{e^{x_i^T\beta y_i^*}}{1+e^{x_i^T\beta}}$$

MLE：最大 全部正确的概率 $Pr(\beta) = \prod_{i=1}^n p_i^{y_i^*}(1-p_i)^{1-y_i^*} = \prod_{i=1}^n \frac{e^{x_i^T\beta y_i^*}}{1+e^{x_i^T\beta}}$ iid

$$\arg\max_\beta Pr(\beta) = \arg\max_\beta \log Pr(\beta) = \arg\max_\beta \sum_{i=1}^n (x_i^T\beta y_i^* - \log(1+e^{x_i^T\beta}))$$

$$\frac{\partial \log Pr(\beta)}{\partial \beta} = \sum_{i=1}^n \left(x_i y_i^* - \frac{x_i e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right) = \sum_{i=1}^n x_i\left(y_i^* - \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right) = \sum_{i=1}^n x_i(y_i^* - p_i)$$

④ $Pr(y_i = +1 | x_i) = \frac{1}{1+e^{-x_i^T\beta}}$ $\quad Pr(y_i = -1 | x_i) = 1 - Pr(y_i = +1 | x_i) = \frac{1}{1+e^{x_i^T\beta}}$

$P(y_i | x_i) = \frac{1}{1+e^{-y_i x_i^T\beta}}$

⑤ Generative model：联合概率分布 $P(X, Y)$

可以求条件概率分布 $P_\theta(Y|X) = \frac{P_\theta(Y, X)}{P_\theta(X)} = \frac{P_\theta(X, Y)}{\sum_{Y'} P_\theta(X, Y')}$

设网络 g_θ，输入 $h \sim N(0, I)$，输出 $X = g_\theta(h) + \varepsilon$，$\varepsilon \sim N(0, \sigma^2 I)$.

$X|h \sim N(g_\theta(h), \sigma^2 I)$，$\log p_\theta(X|h) = -\frac{1}{2}(X - g_\theta(h))^T \Sigma^{-1}(X - g_\theta(h)) + \text{常数}$

$$= -\frac{\|X - g_\theta(h)\|^2}{2\sigma^2} + \text{常数}$$

又 $\log p(h) = -\frac{\|h\|^2}{2} + \text{常数}$，$P_\theta(X) = \int p(h) P_\theta(X|h)\, dh$.

Created by Notein

Discriminative model：条件分布 $P(Y|X)$.

$y=g_\theta(x)$ 或 $\arg\max_y w^T\phi(x,y)$

尤其是用K维向量的显著维度代表分类结果的情况（后者），常让softmax为归一化输出层

$$P_\theta(y=k|X)=\frac{e^{f_\theta^{(k)}(X)}}{\sum_{i=1}^{K}e^{f_\theta^{(i)}(X)}}=\frac{1}{Z(\theta)}e^{f_\theta^{(k)}(X)}.$$ （$f_\theta(X)$为softmax前一层输出向量）

常用于描述分布

$$\mathrm{softmax}_i(x)=\frac{e^{x_i}}{\sum_{i=1}^{K}e^{x_i}}$$

Descriptive model：数据分布 $P(X)$.

$$P_\theta(X)=\frac{1}{Z(\theta)}e^{f_\theta(X)},\quad Z(\theta)=\int e^{f_\theta(x)}dx$$ $f_\theta(X)$相当于数据的显著性.

连续的softmax

⑥ Covariance matrix：$\Sigma=(C_{ij})_{p\times p}$

设 $X\in R^{n\times p}$ $C_{ij}=\mathrm{cov}(X_i,X_j)=E[(X_i-E[X_i])(X_j-E[X_j])]$

Σ_{xx} 半正定，对称.

$\mathrm{Cov}(X,Y)$对称，$\mathrm{Cov}(AX+Y+b,Z)=A\mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z)$.

⑦ Loss functions

Least square Regression：$L(y_i,X_i^T\beta)=(y_i-X_i^T\beta)^2$，优化它实际是计算 $y_i-X_i^T\beta=\varepsilon_i\sim N(0,1)$ 的MLE，即最大化所有预测的正确概率.

$$\log P(y_i|X_i)=-\frac{(y_i-X_i^T\beta)^2}{2}+\text{常数},\quad L(\beta)=\sum_{i=1}^{n}(y_i-X_i^T\beta)^2.$$

Robust Linear Regression：$L(y_i,X_i^T\beta)=|y_i-X_i^T\beta|$.（Absolute value loss）.

$$L(y_i,X_i^T\beta)=\begin{cases}\frac{1}{2}(y_i-X_i^T\beta)^2, & |y_i-X_i^T\beta|\le\delta \quad \text{(Huber loss)}\\ \delta|y_i-X_i^T\beta|-\frac{\delta^2}{2}, & \text{o.w}\end{cases}$$

Logistic Regression with 0/1 Responses：

$$\log\frac{e^{y_iX_i^T\beta}}{1+e^{X_i^T\beta}}=\log\left(\frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}\right)^{y_i}\left(\frac{1}{1+e^{X_i^T\beta}}\right)^{1-y_i}$$

$$L(y_i,X_i^T\beta)=-\log P_r(y_i|X_i,\beta)=-[y_iX_i^T\beta-\log(1+e^{X_i^T\beta})]$$

with ± Responses：

$$\log\frac{1}{1+e^{-y_iX_i^T\beta}}$$

$$L(y_i,X_i^T\beta)=-\log(1+e^{-y_iX_i^T\beta})\quad \text{(Logistic Loss)}$$

For classification $\{\pm1\}$：Logistic Loss

margin越正越大越好

$$L(y_i,X_i^T\beta)=e^{-y_iX_i^T\beta}\quad \text{(Exponential Loss)}$$

Created by Notein

$L(y_i, x_i^T\beta) = \max(0, 1 - y_i x_i^T\beta)$ (Hinge Loss)

$L(y_i, x_i^T\beta) = 1(y_i x_i^T\beta < 0)$ (0-1 Loss).

ReLU求导的反面

---

⑧ 最小二乘法.

$Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$, $\hat{\beta} = \arg\min_{\beta} L(\beta)$, $L(\beta) = |Y - X\beta|^2$

$\frac{\partial}{\partial\beta} L(\beta) = -2X^T(Y - X\beta) = 0$, $\hat{\beta} = (X^TX)^{-1}X^TY$. $\hat{Y} = X(X^TX)^{-1}X^TY$.

$\hat{\beta}$可以看作Y投影至X列空间的归一化权重，$\hat{Y}$即投影. $\varepsilon = Y - \hat{Y}$为无法用$X$表示的正交向量.

$\hat{\beta} = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\beta_{gt} + \varepsilon) = \beta_{gt} + (X^TX)^{-1}X^T\varepsilon$

$E[\hat{\beta}] = \beta_{gt}$, $Var[\hat{\beta}] = Var[(X^TX)^{-1}X^T\varepsilon] = E[((X^TX)^{-1}X^T\varepsilon)((X^TX)^{-1}X^T\varepsilon)^T]$

$= (X^TX)^{-1}X^T E[\varepsilon\varepsilon^T] X (X^TX)^{-1,T}$

$= (X^TX)^{-1}X^T(\sigma^2 I)X(X^TX)^{-1}$

$= \sigma^2(X^TX)^{-1}$

$\hat{\beta} \sim N(\beta_{gt}, \sigma^2(X^TX)^{-1})$

⑨ 假设分布$p(x)$, $q(x)$.

Entropy: $H(P) = E_P[-\log P(X)] = \sum_x P(x)[-\log P(x)]$. 期望编码长度.

Cross Entropy: $CE(P|Q) = E_P[-\log Q(X)] = \sum_x P(x)[-\log Q(x)]$ 用预测分布$q(x)$的期望码长

KL Divergence: $KL(P|Q) = CE(P|Q) - H(P) = \sum_x P(x)[-\log\frac{Q(x)}{P(x)}]$ 期望码长差，度量分布差

⑩ MLE的正确性: $\max_{\theta} L(\theta)$, $L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log P_\theta(y_i|x_i)$.

设$(x_i, y_i) \sim P_d(x, y)$. 根据大数定理

$\max L(\theta) = \max E_{P_d}[\log P_\theta(y|x)]$

$= \min\{-E_{P_d}[\log P_\theta(y|x)]\}$

$= \min\{CE(P_d|P_\theta)\}$

$= \min\{CE(P_d|P_\theta) - H(P_d)\}$

$= \min KL(P_d|P_\theta)$. 故MLE是在最小化预测分布与数据分布的KL散度.

⑫ 各模型的梯度公式:

Discriminative model: $P_\theta(y=k|x) = \frac{e^{f_\theta^{(k)}(x)}}{\sum_{i=1}^{K} e^{f_\theta^{(i)}(x)}} = \frac{1}{Z(\theta)} e^{f_\theta^{(k)}(x)}$.

$\frac{\partial}{\partial\theta}\log P_\theta(y=k|x) = \frac{\partial}{\partial\theta}(f_\theta^{(k)}(x) - \log Z(\theta))$

Created by Notein

$$= \frac{\partial}{\partial\theta} f_\theta^{(k)}(X) - \frac{1}{Z(\theta)} \frac{\partial}{\partial\theta} Z(\theta)$$

$$= \frac{\partial}{\partial\theta} f_\theta^{(k)}(X) - \frac{1}{Z(\theta)} \sum_i \left(e^{f_\theta^{(i)}(X)} \frac{\partial}{\partial\theta} f_\theta^{(i)}(X) \right)$$

$$= \sum_i (1(y=k) - P_i) \frac{\partial}{\partial\theta} f_\theta^{(i)}(X)$$

负梯度

$-P \to 0$

$$= \frac{\partial}{\partial\theta} f_\theta(X)^T (\underbrace{Y}_{\text{期望分布}} - \underbrace{P}_{\text{预测分布}})$$ 假设Y是独热码.

$1-P_k \to 0$

$$= \frac{\partial}{\partial\theta} f_\theta(X)^T (Y - E_\theta(Y|X)).$$

Descriptive model: $P_\theta(X) = \frac{1}{Z(\theta)} e^{f_\theta(X)}$, $Z(\theta) = \int e^{f_\theta(X)} dx$

$$\frac{\partial}{\partial\theta} \log P_\theta(X) = \frac{\partial}{\partial\theta} f_\theta(X) - \frac{\partial}{\partial\theta} \log Z(\theta)$$

$$= \frac{\partial}{\partial\theta} f_\theta(X) - \frac{1}{Z(\theta)} \int e^{f_\theta(X)} \frac{\partial}{\partial\theta} f_\theta(X) dX$$

$$= \frac{\partial}{\partial\theta} f_\theta(X) - \sum_X P_\theta(X) \frac{\partial}{\partial\theta} f_\theta(X) dX$$

$$= \frac{\partial}{\partial\theta} f_\theta(X) - \underbrace{E_\theta\left[\frac{\partial}{\partial\theta} f_\theta(X)\right]}_{\text{所有输入的梯度期望}}$$

减小梯度偏离期望的程度

如果$f_\theta(X)$都等于E_θ,说明到了局部最小值

Generative model:

$$\frac{\partial}{\partial\theta} \log P_\theta(X) = \frac{1}{P_\theta(X)} \frac{\partial}{\partial\theta} \int P_\theta(h, X) dh$$

$$= \frac{1}{P_\theta(X)} \int \left[\frac{\partial}{\partial\theta} \log P_\theta(h, X)\right] P_\theta(h, X) dh$$

$$= \int \left[\frac{\partial}{\partial\theta} \log P_\theta(h, X)\right] P_\theta(h|X) dh.$$

分布X的梯度=所有可行h的(h,X)联合分布梯度的期望

$$= E_{P_\theta(h|X)}\left[\frac{\partial}{\partial\theta} \log P_\theta(h, X)\right]$$

$$\log P_\theta(h_i, X_i) = \log P_\theta(h_i) + \log P_\theta(X_i|h_i) = -\frac{\|X - g_\theta(h)\|^2}{2\sigma^2} - \frac{\|h\|^2}{2} + \text{常数}$$

(11) Gradient Descent: $\theta_{t+1} = \theta_t - \eta_t L_i'(\theta_t)$ RNN

SGD: $\theta_{t+1} = \theta_t - \eta_t \sum_{i \in batch} L_i'(\theta_t)$

Momentum: $V_t = \gamma V_{t-1} + \eta_t g_t$ EMA梯度 LST

$\theta_t = \theta_{t-1} - V_t = \theta_{t-1} - \eta_t g_t - \underline{\gamma V_{t-1}}$ 多了所谓动量 $= \theta_{t-1} - \sum_{n=0}^{t} \gamma^n \eta_{t-n} g_{t-n}$

Adagrad: $C_t = C_{t-1} + g_t^2 = \sum_{n=0}^{t} g_t^2$ R度缩放梯度

$\theta_{t+1} = \theta_t - \eta_t \boxed{\frac{g_t}{\sqrt{C_t + \varepsilon}}}$ 会越来越小

RMSprop: $C_t = \beta C_{t-1} + (1-\beta) g_t^2$ EMA+R度缩放 $\frac{g_t}{\sqrt{C_t + \varepsilon}}$ 变小有所好转

Adam: $V_t = \gamma V_{t-1} + (1-\gamma) g_t$ 动量+EMA+R度缩放

Created by Notein

$$C_t = \beta C_{t-1} + (1-\beta) g_t^2$$

$$V_t \leftarrow V_t/(1-\gamma^t),\ C_t \leftarrow C_t/(1-\beta^t)$$

$$E\left[\frac{V_t}{1-\gamma^t}\right] = \frac{(1-\gamma)}{1-\gamma^t}\sum_{i=1}^{t}\gamma^{t-i}E[g_i] = E[g_i]$$

$$\theta_{t+1} = \theta_t - \eta_t \frac{V_t}{\sqrt{C_t+\varepsilon}}$$

---

⑬ Langevin：在 SGD 中加入随机噪声扰动，有利于跳出 Local Minima

Brownian Motion： $X_{t+\Delta t} = X_t + \sigma\varepsilon_t\sqrt{\Delta t}$ 方差随 Δt 增大

假设 $X_0 = x$, $\Delta t = \frac{t}{n}$ $\quad X_t = x + \sum_{i=1}^{n}\sigma\varepsilon_i\sqrt{\frac{t}{n}} = x + \sigma\sqrt{t}\sum_{i=1}^{n}\frac{\varepsilon_i}{\sqrt{n}} \sim N(x, \sigma^2 t I)$

$$D[X_{t+\Delta t}] = E[(X_t + \sigma\varepsilon_t\sqrt{\Delta t})^2]$$

$$E[X_{t+\Delta t}] = E[X_t] \qquad = E[X_t^2] + 2\sigma\sqrt{\Delta t}E[\varepsilon_t X_t] + \sigma^2\Delta t E[\varepsilon_t^2]$$

$$= D[X_t] + \sigma^2\Delta t I$$

$$D[X_t] = D[x] + \sigma^2 t I$$

Descriptive Model： $X_{t+\Delta t} = \underbrace{X_t - \frac{1}{2}U'(X_t)\Delta t}_{GD} + \underbrace{\sqrt{\Delta t}\,\varepsilon_t}_{Langevin}$

Generative Model： $h_{t+\Delta t} = h_t + \frac{1}{2}\frac{\partial}{\partial h}\log p_\theta(h, x_i)\Delta t + \sqrt{\Delta t}\,\varepsilon_t$

⑭ Linear Discriminant Analysis (LDA)：学习一个线性分类器（一个方向），将样本集合投影到 $z = x_i^T\beta$ 使得类间方差最大化、类内方差最小化。

$\forall x_i \in \Omega^+,\ p(x_i|y=+1) \sim N(\mu^+, \Sigma^+)$

$\forall x_i \in \Omega^-,\ p(x_i|y=-1) \sim N(\mu^-, \Sigma^-)$

$$\sigma^2_{between} = (E_{i\in\Omega^+}[x_i^T\beta] - E_{i\in\Omega^-}[x_i^T\beta])^2$$

$$= (E_{i\in\Omega^+}[x_i^T]\beta - E_{i\in\Omega^-}[x_i^T]\beta)^2$$

$$= [(\mu^+ - \mu^-)^T\beta]^2$$

$$\sigma^2_{within} = n_{pos}\sigma^2_{pos} + n_{neg}\sigma^2_{neg}$$

$$n_{pos} = |\Omega^+|,\ n_{neg} = |\Omega^-|$$

$$\sigma^2_{pos} = E_{i\in\Omega^+}[(x_i^T\beta - E_{i'\in\Omega^+}[x_{i'}^T\beta])^2]$$

$$= E_{i\in\Omega^+}[\beta^T(x_i - E_{i'\in\Omega^+}[x_{i'}])(x_i^T - E_{i'\in\Omega^+}[x_{i'}^T])\beta]$$

$$= \beta^T E_{i\in\Omega^+}[(x_i - E_{i'\in\Omega^+}[x_{i'}])^2]\beta$$

$$= \beta^T\Sigma^+\beta$$

$$E_{i\in\Omega^+, j\in\Omega^-}[(x_i - x_j)^2] = E[x_i^T x_i] - 2E[x_i^T x_j] + E[x_j^T x_j]$$

$$= \Sigma^+ + \mu^{+2} - 2\mu^+\mu^- + \Sigma^- + \mu^{-2}$$

$$= (\Sigma^+ + \Sigma^-) + (\mu^+ - \mu^-)^2$$

$$E_{i\in\Omega^+, j\in\Omega^-}[x_i^T - x_j^T]\beta = (E_{i\in\Omega^+}[x_i^T] - E_{i\in\Omega^-}[x_i^T])$$

$$\sigma^2_{neg} = \beta^T\Sigma^-\beta$$

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}} \qquad \max_\beta S$$

$$= \frac{\beta^T(\mu^+ - \mu^-)^2\beta}{\beta^T(n_{pos}\Sigma^+ + n_{neg}\Sigma^-)\beta} = \frac{\beta^T S_B\beta}{\beta S_W\beta}$$

$$S_B = (\mu^+ - \mu^-)(\mu^+ - \mu^-)^T$$

$$S_W = n_{pos}\Sigma^+ + n_{neg}\Sigma^-$$

只关心 β 的方向，大小可任意缩放，因此可令 $\beta^T S_W\beta = 1$

$$\max_\beta \beta^T S_B\beta \quad s.t.\ \beta^T S_W\beta = 1$$

Lagrange 乘子法： $L = \beta^T S_B\beta - \lambda(\beta^T S_W\beta - 1)$ 对称

$$\frac{\partial L}{\partial\beta} = 2S_B\beta - 2\lambda S_W\beta = 0$$

Created by Notein

假设S_W满秩 $S_W^{-1}S_B\beta=\lambda\beta$，说明$\beta$是$S_W^{-1}S_B$的特征向量

$S_W^{-1}(\mu^+-\mu^-)(\mu^+-\mu^-)^T\beta=\lambda\beta$，$\beta\propto S_W^{-1}(\mu^+-\mu^-)$，可以取单位向量

L2 SVM

$\hat{y}=\text{sign}(w^Tx+b)=\begin{cases}+1, & w^Tx+b\geq 0\\ -1, & w^Tx+b<0\end{cases}$

⑮ Margin: $\gamma_i=y_i(w^Tx_i+b)$

$\gamma_i>0$时分类正确，越大质量越好 $\frac{w^T}{\|w\|}x+\frac{b}{\|w\|}=0$ 分类面 $x=\frac{-bw}{w^Tw}$ margin=0

归一化后，$\gamma_i=\frac{y_i(w^Tx_i+b)}{\|w\|}$，而我们只关注分类最难的样本：$\hat{\gamma}=\min_i \gamma_i$

归一化的γ_i等于x_i到分类面的距离，因为 $x_i-\gamma_i\frac{w}{\|w\|}y_i$ 在分类面上 有向

⑯ SVM任务原本为 $\max_{\tau,w,b}\ \tau$ s.t. $\forall i, y_i(w^Tx_i+b)\geq\tau$，$\|w\|=1$

$w^T(x_i-\gamma_i\frac{w}{\|w\|}y_i)+b$
$=w^Tx_i+b-w^T\frac{(w^Tx_i)w+b\|w\|}{\|w\|^2}$
$=w^Tx_i+b-(w^Tx_i+b)=0$

其中，$\hat{\gamma}=\frac{\tau}{\|w\|}$，因此 $\max_{\tau,w,b}\frac{\tau}{\|w\|}$ s.t. $\forall i, y_i(w^Tx_i+b)\geq\tau$.

由于$\|w\|$变化无影响，让$\tau=1$可以等价为 $\min_{w,b}\frac{1}{2}\|w\|^2$ s.t. $\forall\ 1-y_i(w^Tx_i+b)\leq 0$

满足$1-y_i(w^Tx_i+b)=0$的样本x_i称为支持向量，代表那些最难分辨、margin最小的

$\arg\max\frac{1}{\|w\|}=\arg\max\frac{1}{\|w\|^2}=\arg\min\frac{1}{2}\|w\|^2$

$L(\alpha,w,b)=\frac{1}{2}\|w\|^2-\sum_{i=1}^n\alpha_i[y_i(w^Tx_i+b)-1]$，$\min_{w,b}\max_{\alpha:\alpha_i\geq 0}L(w,b,\alpha)$

$\frac{\partial L}{\partial w}=w-\sum_{i=1}^n\alpha_iy_ix_i=0$，$w=\sum_{i=1}^n\alpha_iy_ix_i=\sum_{i:\alpha_i>0}^n\alpha_iy_ix_i$

因此w相当于$\alpha_i>0$的样本构造出的方向，这些样本称为支持向量。那些$\alpha_i=0$代表$1-y_i(w^Tx_i+b)<0$

⑰ With outliers时SVM即转为

$$\min_{\xi,w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^n\xi_i$$

s.t. $y_i(w^Tx_i+b)\geq 1-\xi_i$，$i=1,2,\cdots,n$

$\xi_i\geq 0$

等价于 $\min_{\xi,w,b}\frac{1}{2}\|w\|^2+C\sum_{i=1}^n\max(0,1-y_i(w^Tx_i+b))$，即惩罚outlier距离

⑱ $\frac{1}{2}\|w\|^2$好在(1)使用Lagrangian时方便能得到$w=\sum_{i=1}^n\alpha_iy_ix_i$从而转换为核函数形式

(2) 凸函数且处处可导，且让margin能最大化.

⑲ With outliers时，$L(\alpha,\beta,w,b)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^n\xi_i+\sum_{i=1}^n\alpha_i(1-y_i(w^Tx_i+b)-\xi_i)-\sum_{i=1}^n\beta_i\xi_i$.

KKT条件 $\frac{\partial L}{\partial w}=0$，$\frac{\partial L}{\partial b}=0$，$1-y_i(w^Tx_i+b)-\xi_i\leq 0$，$\xi_i\geq 0$.

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

$\alpha_i=0\Rightarrow 1-y_i(w^Tx_i+b)<0$，样本非支持向量且在安全区内，$\xi_i=0$ $\beta_i>0$

$0<\alpha_i<C\Rightarrow 1-y_i(w^Tx_i+b)=0$，样本为支持向量 $\xi_i=0$ $\beta_i>0$

$\alpha_i=C\Rightarrow 1-y_i(w^Tx_i+b)>0$，样本为outlier $\xi_i>0$ $\beta_i=0$

都是支持向量

相当于关注权重拉满

$\alpha_i+\beta_i=C$

Created by Notein

⑳ SVM Inference：$y_i = W^T X_i + b$

$$= \sum_{j=1}^{n} \alpha_j y_j X_j X_i + b$$

$$= \sum_{j=1}^{n} \alpha_j y_j \langle X_j, X_i \rangle + b$$ （核函数形式）

对于非线性的分类任务，可以用 $\phi(x)$ 先把 X_i 投影到另一个空间中，从而得到通用SVM

$$y_i = \sum_{j=1}^{n} \alpha_j y_j \langle \phi(X_j), \phi(X_i) \rangle + b$$

$$= \sum_{j=1}^{n} \alpha_j y_j K(X_j, X_i) + b$$

$K(X_i, X_j)$ 即核函数，如果 $K(X_j, X_i)$ 已知，通常它比 $\langle \phi(X_j), \phi(X_i) \rangle$ 运算更 Cheap。

㉑ 核函数的充要条件（Mercer定理）：核矩阵对称（$K(X_i, X_j) = K(X_j, X_i)$）

半正定（$\forall z$，$z^T K z \geq 0$）。

㉒ 常见核函数：

可拆 ☆ RBF Kernel：$K(X_i, X_j) = \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right)$

简单多项式核：$K(X_i, X_j) = (X_i^T X_j)^d$

余弦相似度核：$K(X_i, X_j) = \frac{X_i^T X_j}{\|X_i\| \|X_j\|}$

一般不可拆 Sigmoid 核：$K(X_i, X_j) = \tanh(\alpha X_i^T X_j + c) = \frac{1 - \exp(-2\alpha X_i^T X_j - 2c)}{1 + \exp(-2\alpha X_i^T X_j - 2c)}$

比较接近感知机的一层

L3 Regularized Learning

㉓

	train	test	reason
overfitting	error small	error too large	too much parameters, learned from noises.
underfitting	error too large	error too large/ irrelevant	not enough parameters, missed meaningful info

㉔ Ridge Regression：$L(\beta) = \|Y - X\beta\|^2 + \underbrace{\lambda \|\beta\|^2}_{\text{正则项}}$，$\hat{\beta}_\lambda = \arg\min_\beta L(\beta)$

$$\frac{\partial L}{\partial \beta} = \frac{\partial [(Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta]}{\partial \beta} = -2X^T(Y - X\beta) + 2\lambda\beta = 0$$

$$\hat{\beta}_\lambda = \frac{X^T Y}{X^T X + \lambda I_p}$$

正则项的意义：防止β中少数较大值主导拟合

Created by Notein

㉕ Kernel Regression：用 $f(x)=\sum_{i=1}^{p} c_i K(x, x_i)$ 拟合，x_i 来自训练样本.

$$l(c)=\sum_{i=1}^{n}\|y_i-\sum_{j=1}^{n} c_j K_{ij}\|^2+\lambda\sum_{i,j} c_i c_j K_{ij}=\|Y-Kc\|^2+\underbrace{\lambda c^T K c}_{\text{正则项}}$$

$$\hat{c}_\lambda=\arg\min_c l(c)$$

$$\frac{\partial l}{\partial c}=-2K^T(Y-Kc)+2\lambda Kc=0,\ \hat{c}_\lambda=\frac{Y}{\lambda I_p+K}$$

正则项的意义：防止 $c^T\phi(x)$ 中少数较大值主导拟合，使各样本影响更平均

与 Ridge Regression 关系：$\beta=c^T\phi(x)$ 即岭回归.

在DNN中引入正则项会使分母带 λI_p，它能稳定对角线元素，防止结果过于病态。 深入 ⇒ Neural Tangent Kernel.

㉖ Spline Regression：$f(x)=\alpha_0+\sum_{j=1}^{p}\alpha_j \max(0, x-k_j)$

$$l(\alpha)=\sum_{i=1}^{n}\|y_i-\alpha_0-\sum_{j=1}^{p}\alpha_j\max(0, x_i-k_j)\|^2+\lambda\sum_{j=1}^{p}\alpha_j^2$$

$$=\|Y-\alpha^T R(X-k)\|^2+\lambda\|[{}^{0}_{I_p}]\alpha\|^2$$

$$\alpha=[\alpha_0,\cdots,\alpha_p],\ R(x)=[1, \max(0,x),\cdots,\max(0,x)]$$

与Kernel Regression关系：$\alpha_0=0$ 且 $K(x_i, x_j)=\max(0, x_i-x_j)$ 时.

$$f(x)=\sum_{j=1}^{p}\alpha_j K(x, x_j)$$

㉗ Lasso Regression：$l(\beta)=\frac{1}{2}\|Y-X\beta\|_{L_2}^2+\lambda\|\beta\|_{L_1}$，$\hat{\beta}_\lambda=\arg\min_\beta l(\beta)$ 在 $\beta=0$ 不可导 分类讨论

Least Absolute Shrinkage and Selection Operator

$$\frac{\partial l}{\partial\beta}=-X^T(Y-X\beta)+\lambda(-1)(\beta<0)$$

$$=X^TX\beta-X^TY+\lambda(-1)(\beta<0)=0$$

$$\beta=\frac{\langle X,Y\rangle-\lambda(-1)(\beta<0)}{\|X\|^2}$$

$P=1$ 时 $\hat{\beta}$ 不一定有解

$$\hat{\beta}_\lambda=\begin{cases}(\langle Y,X\rangle-\lambda)/\|X\|_{L_2}^2, & \langle Y,X\rangle\geq\lambda \quad \beta<0\\ (\langle Y,X\rangle+\lambda)/\|X\|_{L_2}^2, & \langle Y,X\rangle<-\lambda \quad \beta=0\\ 0 & |\langle Y,X\rangle|\leq\lambda\end{cases}$$

$$\hat{r}=\langle Y,X\rangle/\|X\|_{L_2}^2 \qquad =\mathrm{sign}(\hat{r})\max(0, |\hat{r}|-\lambda/\|X\|_{L_2}^2)$$

λ 的作用：(1) Shrinkage：把模型权重绝对值减小

(2) Selection：某些权重会被减到零.

与Ridge Regression的关系：Ridge的平方正则化让 λ 出现在分母，使各维度特征更平均.

Lasso的绝对正则化让 λ 出现在分子，使维度特征变得稀疏

㉘ Primal form of Lasso：$\min\|Y-X\beta\|_{L_2}^2/2$ s.t. $\|\beta\|_{L_1}\leq t$

Dual form of Lasso：$\min\|Y-X\beta\|_{L_2}^2/2+\lambda\|\beta\|_{L_1}$

证明：设 $\hat{\beta}$，$\hat{\beta}_\lambda$ 分别为 Primal，Dual 形式的不同解. 令 $t=\|\hat{\beta}_\lambda\|_{L_1}$

Created by Notein

$\|Y-X\hat\beta\|_{L_2}^2/2 < \|Y-X\hat\beta\|_{L_2}^2/2$，$\|\hat\beta\|_{L_1} \le \|\hat\beta_\lambda\|_{L_1}$ 假设 $\hat\beta_\lambda$ 不是Primal form的解

$\|Y-X\hat\beta\|_{L_2}^2/2+\lambda\|\hat\beta\|_{L_1} < \|Y-X\hat\beta_\lambda\|_{L_2}^2/2+\lambda\|\hat\beta_\lambda\|_{L_1}$，$\hat\beta_\lambda$ 不是Dual的最优解，矛盾

㉙ Bayesian Regression：考虑MLE. $P(\beta|X,Y)=\dfrac{P(\beta|X)P(Y|X,\beta)}{P(Y|X)}=P(\beta)P(Y|X,\beta)\cdot C$

与β无关　　引入先验

$$l(\beta)=\log P(\beta|X,Y)$$
$$=\log P(\beta)+\log P(Y|X,\beta)+C$$

$\beta\sim N(0,\tau^2I_p)$, $Y\sim N(X\beta,\sigma^2I)$

$$=-\frac{1}{2\sigma^2}\|Y-X\beta\|_{L_2}^2-\frac{1}{2\tau^2}\|\beta\|_{L_2}^2+C$$

$$\frac{\partial l}{\partial\beta}=\frac{X^T}{\sigma^2}(Y-X\beta)-\frac{1}{\tau^2}\beta=0 \quad \hat\beta=\frac{X^TY}{X^TX+\frac{\sigma^2}{\tau^2}I_p}$$

与Ridge Regression的关系：$\lambda=\frac{\sigma^2}{\tau^2}$

㉚ 与SVM合流：$l(w)=\sum_{i=1}^n \max(0,1-y_iw^TX_i)+\frac{\lambda}{2}\|w\|^2$ （简化掉b）. SVM with outlier second form

$$\frac{\partial l}{\partial w}=-\sum_{i=1}^n 1(y_iw^TX_i<1)y_iX_i+\lambda w$$

相似.

Ridge Logistic Loss：$l(w)=\sum_{i=1}^n \log[1+\exp(-y_iw^TX_i)]+\frac{\lambda}{2}\|w\|^2$

$$\frac{\partial l}{\partial w}=-\sum_{i=1}^n\frac{\exp(-y_iw^TX_i)}{1+\exp(-y_iw^TX_i)}y_iX_i+\lambda w=-\sum_{i=1}^n \mathrm{sigmoid}(-y_iw^TX_i)y_iX_i+\lambda w.$$

㉛ 含Ridge Loss的损失函数解空间的形式：

Linear：$l(\beta)=\sum_{i=1}^n L(y_i;x_i^T\beta)+\lambda\|\beta\|^2 \Rightarrow \hat\beta=\sum_{i=1}^n\alpha_ix_i$

反证，假设 $\tilde\beta=\hat\beta+\sum_{k=1}^K K_kx_k$，$\forall i,k$，$x_k\perp x_i$，则

$$x_i^T\tilde\beta=x_i^T\left(\sum_{j=1}^n\alpha_jx_j+\sum_{k=1}^K K_kx_k\right)$$
$$=\sum_{j=1}^n\alpha_jx_i^Tx_j=x_i^T\hat\beta$$ 损失的前半相等

又 $\|\hat\beta\|^2\le\|\tilde\beta\|^2$，因此 $l(\hat\beta)\le l(\tilde\beta)$，$\hat\beta$ 是最优解. $\|\tilde\beta\|^2=\|\sum\alpha_ix_i\|^2+\|\sum K_kx_k\|^2\ge\|\hat\beta\|^2$

Feature：$l(\beta)=\sum_{i=1}^n L(y_i;\phi(x_i)^T\beta)+\lambda\|\beta\|^2 \Rightarrow \hat\beta=\sum_{i=1}^n\alpha_i\phi(x_i)$

反证，假设 $\tilde\beta=\hat\beta+\sum_{k=1}^K K_k\phi(x_k)$，$\forall i,k$，$\phi(x_k)\perp\phi(x_i)$，后同理

L4 Neural Networks.

㉜ 浅层NN示例：

$$h_{ik}=\mathrm{sigmoid}(X_i^T\alpha_k)$$
$$P_i=\mathrm{sigmoid}(h_i^T\beta)$$
$$y_i\sim\mathrm{Bernouli}(P_i)$$

使用MLE，要使 $P=\prod_i P_i^{y_i}(1-P_i)^{1-y_i}$ 最大

等价

$$\log P=\sum_i[y_i\log P_i+(1-y_i)\log(1-P_i)]$$
$$=\sum_i[y_ih_i^T\beta-y_i\log(1+\exp(h_i^T\beta))-(1-y_i)\log(1+\exp(h_i^T\beta))]$$
$$=\sum_i[y_ih_i^T\beta-\log(1+\exp(h_i^T\beta))]$$

Created by Notein

因此 $l(\beta,\alpha)=\log P$

$$\frac{\partial l}{\partial \beta}=\sum_i\left[y_i h_i-\frac{\exp(h_i^T\beta)}{1+\exp(h_i^T\beta)}h_i\right]=\sum_i(y_i-p_i)h_i$$

$$\frac{\partial l}{\partial \alpha_k}=\frac{\partial l}{\partial h_{ik}}\frac{\partial h_{ik}}{\partial \alpha_k}=\left[\sum_i(y_i-p_i)\beta_k\right]\frac{-x_i\exp(x_i^T\alpha_k)}{[1+\exp(x_i^T\alpha_k)]^2}=\sum_i(y_i-p_i)\beta_k\underbrace{h_{ik}(1-h_{ik})}_{\text{sigmoid求导}}x_i$$

如果用CrossEntropy $l(\beta,\alpha)=-y\log p-(1-y)\log(1-p)=-\sum_i[y_i\log p_i-(1-y_i)\log(1-p_i)]$ 最小

Binary

$$\frac{\partial l}{\partial p_i}=\underbrace{-\frac{y_i}{p_i}+\frac{1-y_i}{1-p_i}}_{\text{CrossEntropy}}$$

$$\frac{\partial l}{\partial h_i}=\frac{\partial l}{\partial p_i}\frac{\partial p_i}{\partial h_i}=\left(\frac{-y_i}{p_i}+\frac{1-y_i}{1-p_i}\right)p_i(1-p_i)\beta=(p_i-y_i)\beta$$

$$\frac{\partial l}{\partial \beta}=\frac{\partial l}{\partial p}\frac{\partial p}{\partial \beta}=\sum_i\left(\frac{-y_i}{p_i}+\frac{1-y_i}{1-p_i}\right)p_i(1-p_i)h_i=\sum_i(p_i-y_i)h_i$$

如果用ReLU $h_{ik}=\max(x_i^T\alpha_k,0)$

$$\frac{\partial h_{ik}}{\partial \alpha_k}=1(\alpha_k^Tx_i>0)\cdot x_i=1(h_{ik}>0)\cdot x_i$$

㉞ 矩阵形式梯度传播：设MLP的某层为

$$h_l=f_l(s_l),\quad s_l=W_lh_{l-1}+b_l$$

反向传播：$\frac{\partial L}{\partial s_l^T}=\frac{\partial L}{\partial h_l^T}\frac{\partial h_l}{\partial s_l^T}=\frac{\partial L}{\partial h_l^T}f_l'$

$n\times n$ $n\times 1$ $1\times n$

$\frac{\partial h_l}{\partial s_l^T}=\partial h_l\left(\frac{1}{\partial s_l}\right)^T$

$$\frac{\partial L}{\partial h_{l-1}^T}=\frac{\partial L}{\partial s_l^T}\frac{\partial s_l}{\partial h_{l-1}^T}=\frac{\partial L}{\partial h_l^T}f_l'W_l$$

转置符号暗示梯度用Jacobian布局，即以h_l的列方向与s_l^T的行方向构建Jacobian矩阵，或维度相当于列向量×行向量

$$W_l=\frac{\partial s_l}{\partial h_{l-1}^T}=\left(\frac{\partial s_{l,i}}{\partial h_{l-1,j}}\right)_{n\times n}$$

$$f_l'=\frac{\partial h_l}{\partial s_l^T}=\underbrace{\mathrm{diag}\left(\frac{\partial h_{l,i}}{\partial s_{l,i}}\right)_n}$$

因为f_l是一个逐元素操作，为公式可以写成对角矩阵形式

$$\frac{\partial L}{\partial W_{lk}}=\frac{\partial L}{\partial s_{lk}}\frac{\partial s_{lk}}{\partial W_{lk}}=\frac{\partial L}{\partial s_{lk}}h_{l-1}^T$$

$$\frac{\partial L}{\partial W_l}=\frac{\partial L}{\partial s_l}h_{l-1}^T=f_l'^T\frac{\partial L}{\partial h_l}h_{l-1}^T=f_l'\frac{\partial L}{\partial h_l}h_{l-1}^T$$

规定 $\Delta h_l=\frac{\partial L}{\partial h_l}$ $\Delta W_l=\frac{\partial L}{\partial W_l}$ $D_l=f_l'$

则 $\Delta h_{l-1}=W_l^TD_l^T\Delta h_l=W_l^TD_l\Delta h_l$

$$\Delta W_l=D_l\Delta h_l h_{l-1}^T$$

W_l,b_l可以一起写为 $(\Delta W_l,\Delta b_l)=\Delta(W_l,b_l)=D_l\Delta h_l(h_{l-1}^T,1)$

㉝ CNN：$S=W\otimes h+b$

stride padding

$P=\lfloor R/2\rfloor\times 2$

$$S_{uvw}=\sum_{i=1}^{R}\sum_{j=1}^{R}\sum_{k=1}^{C}W^w_{ijk}h_{t(u-1)-p+i,\,t(v-1)-p+j,\,k}+b^w$$

$W\in R^{R\times R\times C}$，共有$W$个卷积核，卷积层总参数量为$W(R\times R\times C+1)$

输入通道数 输出通道数 b

单核输出尺寸 $=\left\lfloor\frac{H-R+P}{t}+1\right\rfloor$，后接ReLU、max-pooling、average-pooling等.

间隔个数 始终放置

常见CNN结构：$X\to(\mathrm{Conv}\to\mathrm{ReLU}\to\mathrm{Max\text{-}Pooling})\times 2$

$\to(\mathrm{FC}\to\mathrm{ReLU})\times 2\to\mathrm{FC}\to\mathrm{Softmax}\to\mathrm{CrossEntropy}$

Created by Notein

→(FC→ReLU)×2 → FC ② Softmax → Cross Entropy.

坑：① FC/Conv/ReLU不可连续 ② Softmax前Pooling后不可ReLU

③ Conv后不可直接FC ④ 残差连接前不可ReLU，后可以

㉟ Batch Normalization：在输入批大小维度上各通道分别归一化

$$\mu_d = \frac{1}{n}\sum_{i=1}^{n} x_{id} \quad \sigma_d^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{id}-\mu_{id})^2$$

$$\hat{x}_{id} = \frac{x_{id}-\mu_d}{\sigma_{id}} \quad y_{id} = \underbrace{\beta_d + \gamma_d}_{\text{可学习参数，每通道一对}}\hat{x}_{id}$$

每通道内需要N×H×W共nhw个元素计算统计量.

㊱ 残差网络：$x_{l+1} = x_l + F(x_l)$

F例子：X→FC→BN→ReLU→FC→BN

㊲ RNN：$s_t = f(Ux_t + Ws_{t-1})$ 状态更新

$O_t = g(Vs_t)$ 某时观测

LSTM：在RNN基础上增加旧状态保留与输入筛选机制，即"记忆".

$\Delta C_t = f(W_c(h_{t-1}, x_t))$

$i_t = f(W_i(h_{t-1}, x_t))$ Input Gate 掌管筛选输入

$f_t = f(W_f(h_{t-1}, x_t))$ Forget Gate 掌管保留/遗忘

$C_t = C_{t-1} f_t + \Delta C_t i_t$ Cell state vector 长期记忆，有筛选与保留

$O_t = f(W_o(h_{t-1}, x_t))$ Output Gate

$h_t = O_t f(C_t)$ Hidden state vector 短期记忆，只看上状态与输入

$y_t = f(Wh_t)$

长短期记忆合一

GRU：融合Cell State与Hidden State，简化LSTM.

$z_t = f(W_z(h_{t-1}, x_t))$ Update Gate Input+Forget Gate

$r_t = f(W_r(h_{t-1}, x_t))$ Reset Gate 决定新记忆参考多少历史

$\tilde{h}_t = f(W(x_t, r_t h_{t-1}))$ New Info from Input

$h_t = (1-z_t)h_{t-1} + z_t\tilde{h}_t$ Hidden State

$y_t = f(W_h h_t)$

㊳ GAN：判别器D，生成器G，输入$h \sim N(0, \sigma^2 I)$，G_θ输出$\hat{x}$，D_ϕ判断$\hat{x}$是否是真实样本，输出是真实图像的概率。

对D来说，max V(D,G)相当于最小化一个BCE损失，增强判别能力

对G来说，min V(D,G)相当于最大化它参与的BCE损失，增强造假能力

D增强把X判别为真 D增强/G削弱把$\hat{x}$判别为假

Created by Notein

$$V(D,G)=E_{P_{data}}[\log P_D(X)]+E_{h\sim p(h)}[\log(1-P_D(G(h)))]$$

$$=E_{P_{data}}[\log D(X)]+E_{h\sim p(h)}[\log(1-D(G(h)))] \rightarrow \max_G E_{h\sim p(h)}[\log D(G(h))]$$

最小化最大化 $\min_G \max_D V(D,G)$

优化 $V(D,G)$ 本质是在求从生成分布 P_θ 与真实分布 P_{data} 间的 Jensen-Shannon Divergence：

$$\frac{\partial V}{\partial D}=\frac{P_{data}(X)}{D(X)}-\frac{P_\theta(X)}{1-D(X)}=0$$

$$D(X)=\frac{P_{data}(X)}{P_{data}(X)+P_\theta(X)}$$

$$P_{mix}=\frac{P_\theta+P_{data}}{2}=\frac{P_{data}}{2D(X)}=\frac{P_\theta}{2(1-D(X))}$$

$$JSD(P_{data}|P_\theta)=KL(P_\theta|P_{mix})+KL(P_{data}|P_{mix})$$

P_θ 与 P_{data} 分别相对 P_{mix} 的 KL 散度之和

$$=\sum_X\left[P_\theta\log\frac{P_\theta}{P_{mix}}+P_{data}\log\frac{P_{data}}{P_{mix}}\right]$$

$$=-H(P_\theta)-H(P_{data})-\sum_X[P_\theta\log P_{mix}+P_{data}\log P_{mix}]$$

$$=-H(P_\theta)-H(P_{data})-\sum_X\left[P_\theta\log\frac{P_\theta}{2(1-D(X))}+P_{data}\log\frac{P_{data}}{2D(X)}\right]$$

$$=\log 2\sum_X(P_\theta+P_{data})+\sum_X[P_\theta\log(1-D(X))+P_{data}\log D(X)]$$

$$=2\log 2+V(D,G)$$

㊴ VAE：Encoder q_ϕ 把输入 x 压缩成低维向量 h，Decoder P_θ 把 h 重建为 $\hat{x}$，目标是 $\min\|\hat{x}-x\|^2$

$$L(\phi,\theta,X)=KL(q_\phi(h|X)|P_\theta(h|X))-E_{q_\phi(h|X)}[\log P_\theta(X|h)] \qquad \min_{\phi,\theta}L(\phi,\theta,X)$$

让后验分布去接近先验　简化为先验分布　最大化重建期望

引入先验：$P_\theta(h|X)\sim N(0,I)$ 让 Encoder 学习到近似 Gaussian 的分布防止过拟合。

与 Evidence Lower Bound (ELBO) 的关系：

$$L(\phi,\theta,X)=KL(q_\phi(h|X)|P_\theta(h|X))+KL(P_{data}(X)|P_\theta(X))-C$$

$$=\underbrace{E_{P_{data}}[\log P_{data}(X)]}_{常数}-E_{P_{data}}[\log P_\theta(X)]+KL(q_\phi(h|X)|P_\theta(h|X))-C$$

$P(X)$ 无法计算

$$=-\left(E_{P_{data}}[\log P_\theta(X)]-KL(q_\phi(h|X)|P_\theta(h|X))\right)$$

$$=-ELBO(\phi,\theta,X)=-E_{q_\phi(h|X)}\left[\log\frac{P_\theta(h,X)}{q_\phi(h|X)}\right]$$

逼开与计算真实后验

虽然真实后验被省了

即 $\min L(\phi,\theta,X)$ 等效于 $\max ELBO$　$(\underbrace{\log P_\theta(X)}_{常数}=\underbrace{ELBO}_{增大\rightarrow减小}+KL\geq ELBO)$

L6 PCA, t-SNE, EM

㊵ PCA：找到数据方差最大的几个主方向，将数据投影到这些方向组成的椭圆上，舍弃那些方差最小的方向从而实现降维

(1) 去中心化，$X\in R^{n\times p}$，$X\leftarrow X-\text{Expand}(\frac{1}{n}\sum_i X_{i})$

(2) 计算协方差矩阵，$\Sigma=\frac{1}{n}X^TX$

(3) 求特征值与特征向量，$\Sigma W=\lambda W$，特征值 $\lambda_1>\lambda_2>\cdots>\lambda_p$，对应特征向量 $W_{(1)},W_{(2)},\cdots W_{(p)}$，第 i 主成分即 $W_{(i)}$。

如何求入和 W？

① $\det[\Sigma-\lambda I]=0$　因为 $(\Sigma-\lambda I)W=0$

② 代入各 λ 求 W_i

③ 归一化各 W_i

Created by Notein

(4) 数据降维，假设要降到k维，令 $W=[W_{(1)}, W_{(2)}, \cdots, W_{(k)}]$ 为转移矩阵，

$T=XW\in R^{n\times k}$ 即每条数据在各主方向的投影

㊶ Stochastic Neighbor Embedding (SNE)：降维后数据保持相似度

原始高维数据 x_i，降维后数据 h_i

$$P_{ij}=\frac{G(x_i,x_j)}{\sum G(x_m,x_i)}$$

定义高维数据相似度 $P_{ij}=P(x_j|x_i)$

$$=\frac{\exp(-|x_i-x_j|^2/2\sigma_i^2)}{\sum_{m=1,m\neq i}^{n}\exp(-|x_m-x_i|^2/2\sigma_i^2)}\propto e^{-|x_i-x_j|^2/2\sigma_i^2}\propto N(0,\sigma_i^2)$$ 通常为1

q_{ij} 可以用 h_i, h_j 直接替换 P_{ij} 公式，如果 q_{ij} 用自由度为1的t分布代替高斯分布，

即t-SNE $q_{ij}=P(h_j|h_i)$

$$q_{ij}=\frac{T(h_j,h_i)}{\sum T(h_m,h_i)}$$

$$=\frac{(1+|h_j-h_i|^2)^{-1}}{\sum_{m=1,m\neq j}^{n}(1+|h_m-h_i|^2)^{-1}}$$

目标函数 $Loss=KL(P\|Q)=\sum_{i,j}P_{ij}\log\left(\frac{P_{ij}}{q_{ij}}\right)$，$\min_{\{h_i\}}Loss$，梯度下降即可

㊷ Local Linear Embedding (LLE)：低维上保留邻居关系，用 N_i 表示 x_i 的相邻数据

$x_i\approx\sum_{j\in N_i}w_{ij}x_j$，有可能不完全能用线性组合表示

$\min_{\{w_{ij}\}}|x_i-\sum_{j\in N_i}w_{ij}x_j|^2$ 得到 w_{ij}

$h_i\approx\sum_{j\in N_i}w_{ij}h_j$，保持 w_{ij} 不变，找低维表示，$\min_{\{h_i\}}|h_i-\sum_{j\in N_i}w_{ij}h_j|^2$ 得到 h_i

㊸ The Expectation Maximization (EM) Algorithm：建模X的潜在分布

样本X，未知统计模型参数θ，未观测到的潜在数据Z

动机是建模已知观测数据的分布但有些隐变量未知(如GMM的聚类标签)因此无法MLE

用MLE：$L(\theta;X,Z)=p(X,Z|\theta)$，$L(\theta,X)\overset{def}{=}p(X|\theta)=\prod_{i=1}^{n}p(x_i|\theta)$

$$\max_\theta L(\theta,X)$$

$$=\int_Z P(X,Z|\theta)dZ=\prod_{i=1}^{n}E_{Z_i\sim P(Z_i)}[P(x_i|Z_i,\theta)]$$

用下面的两步循环求解：

$$\log P(X|\theta)=\log\sum_Z P(X,Z|\theta)=\log\sum_Z Q(Z)\frac{P(X,Z|\theta)}{Q(Z)}$$

(1) Expectation Step (E step)：

(log满足Jensen不等式) $\geq\sum_Z Q(Z)\log\frac{P(X,Z|\theta)}{Q(Z)}$

$$Q(\theta|\theta^{(t)})=E_{Z\sim P(Z|X,\theta^{(t)})}[\log P(X,Z|\theta)]$$ 取等：$E_{Z\sim P(Z|X,\theta^{(t)})}[\log P(X,Z|\theta)]$ (含Q(Z))

给定X和Z情况下的MLE

$$=\sum_{i=1}^{n}E_{Z_i\sim P(Z_i|X_i,\theta^{(t)})}[\log P(Z_i)+\log P(x_i|Z_i,\theta)]$$

(2) Maximization Step (M step)：

Jensen不等式的取等条件：$\frac{P(X,Z|\theta)}{Q(Z)}=C$ (常数)

$$\theta^{(t+1)}=\arg\max_\theta Q(\theta|\theta^{(t)})$$

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

从MLE推导Q：$\Delta\theta=\eta\frac{\partial}{\partial\theta}\log P(X|\theta)$

$C=P(X|\theta)$

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

☆ $Q(Z)=\frac{P(X,Z|\theta)}{P(X|\theta)}=P(Z|X,\theta)$

$$=\eta\frac{1}{P(X|\theta)}\frac{\partial}{\partial\theta}\sum_Z P(X,Z|\theta)$$ 引入隐变量Z

Created by Notein

$$= \eta \frac{1}{P(X|\theta)} \sum_{Z} \frac{\partial}{\partial \theta} P(X,Z|\theta)$$

$$= \eta \frac{1}{P(X|\theta)} \sum_{Z} \underline{P(X,Z|\theta) \frac{\partial}{\partial \theta} \log P(X,Z|\theta)}$$ 只与近似概率函数有关

$$= \eta \sum_{Z} \frac{P(X,Z|\theta)}{P(X|\theta)} \frac{\partial}{\partial \theta} \log P(X,Z|\theta)$$

$$= \eta \sum_{Z} P(Z|X,\theta) \frac{\partial}{\partial \theta} \log P(X,Z|\theta)$$

$$\approx \eta \sum_{Z} P(Z|X,\theta^{(t)}) \frac{\partial}{\partial \theta} \log P(X,Z|\theta)$$ 用$\theta^{(t)}$代替θ

$$= \eta \frac{\partial}{\partial \theta} E_{Z \sim P(Z|X,\theta^{(t)})} [\log P(X,Z|\theta)] = \eta \frac{\partial}{\partial \theta} Q(\theta|\theta^{(t)})$$

用到K-Means中：$Z = [z_1, z_2, \cdots, z_n]$, $z_i \in \{1, 2, \cdots, K\}$

$$\hat{P}_k = \sum_{i=1}^{n} P(z_i = k | x_i, \theta) P(x_i|\theta), \quad \hat{\mu}_k = \frac{\sum_{i=1}^{n} P(z_i = k | x_i, \theta) x_i}{\sum_{i=1}^{n} P(z_i = k | x_i, \theta)}$$

$$z_i = \arg\min_{k} |x_i - \mu_k|$$

Created by Notein

Created by Notein