HW9+10

HW9

1. $err^*(x) = 1 - \max_{c\in y} P(c|x)$

$err(x) = 1 - \sum_c P(c|x)\, P(c|z)$

$|y| \times \max_{c\in y} P(c|x) \geq \sum_{c\in y} P(c|x)$

$\because \sum_c P(c|z) \leq 1 \quad \therefore \ err^*(x) \leq err(x)$

下面证明: $err(x) \leq err^*(x) \left(2 - \frac{|y|}{|y|-1} \times err^*(x)\right)$

$err(x) = 1 - \sum_c P(c|x)\, P(c|z) = 1 - \sum_c P^2(c|x)$

· 由于 $\left(\sum_{c\in y} P(c|x)\right)^2 \leq |y| \left(\sum_{c\in y} P^2(c|x)\right)$

即有 $err(x) = 1 - \sum_{c\in y} P^2(c|x) \leq 1 - \frac{1}{|y|} \left(\sum_{c\in y} P(c|x)\right)^2.$

$\because \max_{c\in y} P(c|x) \geq \frac{1}{|y|} \sum_{c\in y} P(c|x)$

$\therefore \ err^*(x) \leq 1 - \frac{1}{|y|} \sum_{c\in y} P(c|x)$

$err^*(x) \left[2 - \frac{|y|}{|y|-1} err^*(x)\right] = 2 - 2\max_{c\in y} P(c|x) - \frac{|y|}{|y|-1} [err^*(x)]^2$

$\geq 2 - 2\max_{c\in y} P(c|x) - \frac{|y|}{|y|-1} + \frac{2|y|}{|y|-1} \max_{c\in y} P(c|x) - \frac{1}{|y|(|y|-1)} \left(\sum_{c\in y} P(c|x)\right)^2.$

$= \frac{|y|-2}{|y|-1} - \frac{1}{|y|(|y|-1)} \left(\sum_{c\in y} P(c|x)\right)^2 + \frac{2}{|y|-1} \max_{c\in y} P(c|x)$

即有 $err^*(x)\left[2-\frac{|y|}{|y|-1}err^*(x)\right]$

$$\geq \frac{|y|-2}{|y|-1}-\frac{1}{|y|(|y|-1)}\left(\sum_{c\in y}p(c|x)\right)^2+\frac{2}{|y|-1}\max_{c\in y}p(c|x)$$

$$\geq \frac{|y|-2}{|y|-1}-\frac{1}{|y|(|y|-1)}+\frac{2}{|y|-1}\times\frac{1}{|y|}$$

$$=\frac{|y|-2}{|y|-1}+\frac{1}{|y|(|y|-1)}$$

$$\geq 1-\frac{1}{|y|}\left(\sum_{c\in y}p(c|x)\right)^2 \geq err(x)$$

综上，证得 $err^*(x)\leq err(x)\leq err^*(x)\left(2-\frac{|y|}{|y|-1}err^*(x)\right)$

10.4. ~~因为奇异值分解后直接就得到了最大的特征值，不需要先求所有特征值再筛选。~~

对X作奇异值分解得 $X=U\Sigma V^T$，$XX^T=U\Sigma V^T(U\Sigma V^T)^T$

$$=U\Sigma V^TV\Sigma^TU^T$$

$$XX^T=U\Sigma\Sigma^TU^T=U\Lambda U^T$$

X的奇异值分解与 XX^T 特征分解的效果等价.

但对于存储高维数据的X，如 $X\in R^{10000\times 100}$

对 $XX^T\in R^{10000\times 10000}$ 特征分解的计算成本显然更高.

且X的奇异值分解的计算精度更高.

3. 求解: $\max\limits_{W} tr(W^TXX^TW)$ s.t. $W^TW=I_{d'}$.

$\Rightarrow \min \ -tr(W^TXX^TW)$ s.t. $W^TW=I_{d'}$

其中 $X=(x_1,\cdots,x_m)\in R^{d\times m}$ $W=(w_1,w_2,\cdots,w_{d'})\in R^{d\times d'}$ $I_{d'}\in R^{d'\times d'}$

令 $\theta\in R^{d'\times d'}$ 为拉格朗日乘子矩阵.

$L(W,\theta)=-tr(W^TXX^TW)+<\theta, W^TW-I>$

$=-tr(W^TXX^TW)+tr(\theta(W^TW-I))$

若仅考虑约束 $w_i^Tw_i=1$ $(i=1,2,\cdots,d')$

此时 θ 为对角矩阵, 令新的拉格朗日乘子矩阵为 $\Lambda=diag(\lambda_1,\cdots\lambda_{d'})\in R^{d'\times d'}$

$L(W,\Lambda)=-tr(W^TXX^TW)+tr(\Lambda^T(W^TW-I))$

$\frac{\partial L(W,\Lambda)}{\partial W}=-\frac{\partial}{\partial W}tr(W^TXX^TW)+\frac{\partial}{\partial W}tr(\Lambda^T(W^TW-I))$

$=-XX^TW-XX^TW+W\Lambda+W\Lambda^T$

$=-2XX^TW+2W\Lambda$

令 $\frac{\partial L(W,\Lambda)}{\partial W}=0$ 得 $XX^T\Lambda=W\Lambda$.

~~显然 $\lambda_1\cdots\lambda$~~ $XX^Tw_i=\lambda_iw_i$ $i=1,2,\cdots,d'$

显然 $\lambda_1,\cdots,\lambda_{d'}$ 为 XX^T 的特征值,

$w_1\cdots w_{d'}$ 为 $\lambda_1,\cdots,\lambda_{d'}$ 对应单位特征向量

又∵ $w_i^Tw_j=0$ $(i\neq j)$ 为原问题的约束.

且当 ~~λ_i与λ_j~~ $\lambda_i\neq\lambda_j$ 时, w_i与w_j 正交, 即

满足原约束条件.

从 XX^T 的 d 个特征向量中找出 d' 个能使得目标函数达到最优值的特征向量作为最优解。

将 $XX^T w_i = \lambda_i w_i$ 代入 $\min\limits_{W} -tr(W^T X X^T W)$

$$
\begin{aligned}
\min_{W} -tr(W^T X X^T W) &= \max_{W} tr(W^T X X^T W) \\
&= \max_{W} \sum_{i=1}^{d'} w_i^T X X^T w_i \\
&= \max_{W} \sum_{i=1}^{d'} w_i^T \lambda_i w_i \\
&= \max_{W} \sum_{i=1}^{d'} \lambda_i w_i^T w_i \\
&= \max_{W} \sum_{i=1}^{d'} \lambda_i
\end{aligned}
$$

此时只需令 λ_1、…，$\lambda_{d'}$ 为 XX^T 的前 d' 个最大特征值，w_1、…、$w_{d'}$ 为 λ_1、…$\lambda_{d'}$ 分别对应的单位特征向量

就可使目标函数取到最优值。

HW10

11.5、11.7

11.5 L_1正则化可以产生稀疏解，是因为平方误差项等值线与L_1等值线的第一个交点位于坐标轴上。当平方误差项等值线的曲率较大时，会导致其与L_1等值线的第一个交点不再位于坐标轴上，此时无法产生稀疏解。

11.7 L_0范数是不连续的，且是非凸函数，无法通过优化直接求解，必须采用遍历的方式，从而导致这个问题是NP难问题。

3. ① 证明回归：$f(x)=wx+b$

$$L(w,b)=\frac{1}{2n}\sum_{i=1}^{n}(f(x_i)-y_i)^2.$$

$$\nabla L(w,b)=\left(\frac{\partial L}{\partial w},\frac{\partial L}{\partial b}\right)\quad \frac{\partial L}{\partial w}=\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)\cdot x_i$$

$$\frac{\partial L}{\partial b}=\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)$$

$$|f(x)-f(y)|\le L\|x-y\|$$

$$L=\frac{1}{n}\sum_{i=1}^{n}x_i^2$$

② 对数几率回归：$f(w,b)=-\frac{1}{n}\sum_{i=1}^{n}[y_i\log(f(x_i))+(1-y_i)\log(1-f(x_i))]$

$$f(x_i)=\frac{1}{1+e^{-x_i}}\qquad \frac{\partial L}{\partial w}=\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)\cdot x_i$$

~~|f(x)-f(y)| ≤ L||~~

$$\frac{\partial L}{\partial b}=\frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)$$

~~|f(x)-f(y)|~~

$$L=\frac{1}{4}\sigma_{max}(X^TX)$$

$$=\frac{1}{4}\max_i\|x_i\|^2.$$