

## IML 第五次作业

### 作业 10.1

设  $\max_{c \in \mathcal{Y}} P(c|\mathbf{x}) = P(c_0|\mathbf{x})$ , 则  $P(c_0|\mathbf{x}) \geq P(c|\mathbf{x})$ , 又因为  $\text{err}^*(\mathbf{x}) = 1 - \max_{c \in \mathcal{Y}} P(c|\mathbf{x})$ , 则

$$\sum_c P(c|\mathbf{x})P(c|\mathbf{z}) \leq \sum_c P(c_0|\mathbf{x})P(c|\mathbf{z}) = P(c_0|\mathbf{x}) \sum_c P(c|\mathbf{z}) = P(c_0|\mathbf{x})$$

所以

$$\text{err}(\mathbf{x}) = 1 - \sum_c P(c|\mathbf{x})P(c|\mathbf{z}) \geq 1 - P(c_0|\mathbf{x})$$

所以

$$\text{err}^*(\mathbf{x}) \leq \text{err}(\mathbf{x})$$

又因为  $\text{err}(\mathbf{x}) = 1 - \sum_c P(c|\mathbf{x})P(c|\mathbf{z})$ , 由样本无穷多可得

$$\text{err}(\mathbf{x}) \sim 1 - \sum_c P^2(c|\mathbf{x})$$

则

$$1 - \sum_c P^2(c|\mathbf{x}) = 1 - P^2(c_0|\mathbf{x}) - \sum_{c \neq c_0} P^2(c|\mathbf{x}) \quad (1)$$

$$\leq 1 - P^2(c_0|\mathbf{x}) - \frac{1}{|\mathcal{Y}| - 1} \left[ \sum_{c \neq c_0} P(c|\mathbf{x}) \right]^2 \quad (2)$$

$$= 1 - P^2(c_0|\mathbf{x}) - \frac{1}{|\mathcal{Y}| - 1} [1 - P(c_0|\mathbf{x})]^2 \quad (3)$$

$$= [1 - P(c_0|\mathbf{x})] \left[ 1 + P(c_0|\mathbf{x}) - \frac{1 - P(c_0|\mathbf{x})}{|\mathcal{Y}| - 1} \right] \quad (4)$$

$$= \text{err}^*(\mathbf{x}) \left[ 2 - \frac{|y|}{|y| - 1} \times \text{err}^*(\mathbf{x}) \right] \quad (5)$$

### 习题 10.4

任意实矩阵  $A \in \mathbb{R}^{m \times n}$  均可分解成

$$A = U\Sigma V^\top$$

其中  $U$  是满足  $UU^\top = I$  的  $m$  阶酉矩阵;  $V$  是满足  $VV^\top = I$  的  $n$  阶酉矩阵;  $\Sigma$  为除了对角元以外, 其余元素皆为 0 的  $m \times n$  的矩阵, 则

$$AA^\top = U\Sigma V^\top V \Sigma^\top U^\top$$

$$A^T A = V \Sigma U^T U \Sigma^T V^T$$

可见  $U$  的列向量就是  $AA^T$  的特征向量,  $V$  的列向量就是  $A^T A$  的特征向量,  $\Sigma$  非零对角元平方就是  $AA^T$  和  $A^T A$  的共同非零特征值。

所以两种分解是等价的, 但是奇异值分解需要的计算和存储成本更低, 因此实践中更常用。

### 作业 10.3

设  $X = (x_1, x_2, \dots, x_m) \in \mathbb{R}^{d \times m}$ ,  $W = (w_1, w_2, \dots, w_{d'}) \in \mathbb{R}^{d \times d'}$ 。其优化目标的拉格朗日函数为

$$\begin{aligned} L(W, \Lambda) &= -\text{tr}(W^T X X^T W) + \langle \Lambda, W^T W - I \rangle \\ &= -\text{tr}(W^T X X^T W) + \text{tr}(\Lambda^T (W^T W - I)) \end{aligned}$$

其中  $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{d'}) \in \mathbb{R}^{d' \times d'}$  为拉格朗日乘子矩阵, 对拉格朗日函数关于  $W$  求导可得

$$\begin{aligned} \frac{\partial L(W, \Lambda)}{\partial W} &= \frac{\partial}{\partial W} [-\text{tr}(W^T X X^T W) + \text{tr}(\Lambda^T (W^T W - I))] \\ &= -\frac{\partial}{\partial W} \text{tr}(W^T X X^T W) + \frac{\partial}{\partial W} \text{tr}(\Lambda^T (W^T W - I)) \\ &= -2X X^T W + W \Lambda + W \Lambda^T \\ &= -2X X^T W + W (\Lambda + \Lambda^T) \\ &= -2X X^T W + 2W \Lambda \end{aligned}$$

$$\text{令 } \frac{\partial L(W, \Lambda)}{\partial W} = 0 \text{ 可得}$$

$$\begin{aligned} -2X X^T W + 2W \Lambda &= 0 \\ X X^T W &= W \Lambda \end{aligned}$$

将  $W$  和  $\Lambda$  展开可得

$$X X^T w_i = \lambda_i w_i, \quad i = 1, 2, \dots, d'$$

此式为矩阵特征值和特征向量的定义式, 其中  $\lambda_i$  和  $w_i$  分别表示矩阵  $X X^T$  的特征值和单位特征向量。

又  $X X^T$  是一个实对称矩阵, 所以通过上式求得的  $w_i$  可以同时满足约束  $w_i^T w_i = 1$  和  $w_i^T w_j = 0 (i \neq j)$ 。

将  $XX^\top w_i = \lambda_i w_i$  代入目标函数可得

$$\begin{aligned}\max_W \operatorname{tr}(W^\top XX^\top W) &= \max_W \sum_{i=1}^{d'} w_i^\top XX^\top w_i \\ &= \max_W \sum_{i=1}^{d'} w_i^\top \lambda_i w_i \\ &= \max_W \sum_{i=1}^{d'} \lambda_i w_i^\top w_i \\ &= \max_W \sum_{i=1}^{d'} \lambda_i.\end{aligned}$$

令  $\lambda_1, \lambda_2, \dots, \lambda_{d'}$  和  $w_1, w_2, \dots, w_{d'}$  分别为矩阵  $XX^\top$  的前  $d'$  个最大的特征值和单位特征向量构成的  $W$ ，就是主成分分析的解。

## 附加题 10

欲最小化的对象

$$f(P) = \sum_{(x_i, x_j) \in \mathcal{M}} \|x_i - x_j\|_M^2 = \sum_{(x_i, x_j) \in \mathcal{M}} (x_i - x_j)^\top PP^\top (x_i - x_j)$$

因为

$$\begin{aligned}\frac{\partial f(P)}{\partial P} &= \sum_{(x_i, x_j) \in \mathcal{M}} (x_i - x_j)^\top \left( P^\top + P \frac{\partial P^\top}{\partial P} \right) (x_i - x_j) \\ \frac{\partial^2 f(P)}{\partial P \partial P^\top} &= \sum_{(x_i, x_j) \in \mathcal{M}} (x_i - x_j)^\top (I + C) (x_i - x_j)\end{aligned}$$

而约束条件为

$$g(P) = 1 - \sum_{(x_i, x_j) \in \mathcal{C}} \|x_i - x_j\|_M^2 = 1 - \sum_{(x_i, x_j) \in \mathcal{C}} (x_i - x_j)^\top PP^\top (x_i - x_j) \leq 0$$

其二阶导

$$\frac{\partial^2 g(P)}{\partial P \partial P^\top} = - \sum_{(x_i, x_j) \in \mathcal{C}} (x_i - x_j)^\top (I + C) (x_i - x_j)$$

因为  $\sum_{(x_i, x_j) \in \mathcal{F}} (x_i - x_j)^\top (x_i - x_j) \geq 0$ ，所以  $\frac{\partial^2 f(P)}{\partial P \partial P^\top}$  与  $\frac{\partial^2 g(P)}{\partial P \partial P^\top}$  不可能符号相同，不可能同时为凸函数，所以这个问题不是凸优化问题。

## 习题 11.5

$L_1$  正则化产生稀疏解  $\Leftrightarrow$  平方误差等值线和正则化的等值线的交点出现在坐标轴上。

当交点不在坐标轴上时，就不会产生稀疏解，即：平方误差等值线存在斜率绝对值为 1 的点，且与  $L_1$  正则等值线相切。

**习题 11.7**

$L_0$  范数等于非零元素的个数，这个函数不是连续的，也不是凸函数，很难通过优化直接求解。

**作业 11.3**

对于回归问题，设  $f(w) = (y - Xw)^\top (y - Xw)$ ，其梯度

$$\nabla f(w) = 2X^\top (Xw - y)$$

对任意  $w, w'$ ，有

$$\|\nabla f(w') - \nabla f(w)\| = \|2X^\top (Xw' - Xw)\| = \|2X^\top X(w' - w)\| \leq 2\|X^\top X\| \|w' - w\|$$

因为  $X^\top X$  是半正定矩阵，只要  $\|X^\top X\| \neq 0$ ，就存在  $L = 2\|X^\top X\| > 0$  使得回归问题的损失函数的梯度满足  $L$ -Lipschitz 条件。

对于对率回归问题，设  $g(w) = \sum_{i=1}^m (-y_i w^\top x_i + \ln(1 + e^{w^\top x_i}))$ ，其梯度

$$\nabla g(w) = -\sum_{i=1}^m x_i (y_i + \frac{e^{w^\top x_i}}{1 + e^{w^\top x_i}})$$

令  $\phi(x) = \frac{e^x}{1+e^x}$ ， $\phi'(x) = \frac{e^x}{(1+e^x)^2} \leq \frac{1}{4}$ ，所以对任意  $w, w'$ ，有

$$\|\nabla g(w') - \nabla g(w)\| \leq \frac{1}{4} \left\| \sum_{i=1}^m x_i (w'^\top x_i - w^\top x_i) \right\| = \frac{1}{4} \|X^\top X(w' - w)\| \leq \frac{1}{4} \|X^\top X\| \|w' - w\|$$

因为  $X^\top X$  是半正定矩阵，只要  $\|X^\top X\| \neq 0$ ，就存在  $L = \frac{1}{4}\|X^\top X\| > 0$  使得回归问题的损失函数的梯度满足  $L$ -Lipschitz 条件。