• 记$\mathrm{err}^*(\boldsymbol{x}) = 1 - \max_{c\in\mathcal{Y}} P(c|\boldsymbol{x})$， $\mathrm{err}(\boldsymbol{x}) = 1 - \sum_c P(c|\boldsymbol{x})P(c|\boldsymbol{z})$，其中$\boldsymbol{z}$为$\boldsymbol{x}$的最近邻，试证明在样本无穷多时

$$\mathrm{err}^*(\boldsymbol{x}) \le \mathrm{err}(\boldsymbol{x}) \le \mathrm{err}^*(\boldsymbol{x})\left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \times \mathrm{err}^*(\boldsymbol{x})\right)$$

提示：柯西-施瓦兹不等式$(\sum_i a_i)^2 \le n(\sum_i a_i^2)$

设 $\max_{c\in y} P(c|x) = P(c^*|x)$

$\sum_c P(c|x)P(c|z) \le \sum_c P(c^*|x)P(c|z) = P(c^*|x)\sum_c P(c|z) = P(c^*|x)$

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

$\mathrm{err}(x) = 1 - \sum_c P(c|x)P(c|z) \geqslant 1 - P(c^*|x)$

故 $\mathrm{err}^*(x) \le \mathrm{err}(x)$，不等式左边得证.

在样本无穷多时，

$$\begin{aligned}
\mathrm{err}(x) &= 1 - \sum_c P(c|x)P(c|z) \approx 1 - \sum_c P^2(c|x) = 1 - P^2(c^*|x) - \sum_{c\ne c^*} P^2(c|x) \\
&\leqslant 1 - P^2(c^*|x) - \frac{1}{|y|-1}\left(\sum_{c\ne c^*} P(c|x)\right)^2 = 1 - P^2(c^*|x) - \frac{1}{|y|-1}(1 - P(c^*|x))^2 \\
&= (1 - P(c^*|x))\left[1 + P(c^*|x) - \frac{1}{|y|-1}(1 - P(c^*|x))\right] \\
&= \mathrm{err}^*(x)\left[-(1 - P(c^*|x)) + 2 - \frac{1}{|y|-1}(1 - P(c^*|x))\right] \\
&= \mathrm{err}^*(x)\left(2 - \frac{1}{|y|-1}\mathrm{err}^*(x)\right)
\end{aligned}$$

故 $\mathrm{err}(x) \le \mathrm{err}^*(x)\left(2 - \frac{1}{|y|-1}\mathrm{err}^*(x)\right)$，不等式右边得证.

10.4 在实践中，协方差矩阵 $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 的特征值分解常由中心化后的样本矩阵 $\mathbf{X}$ 的奇异值分解代替，试述其原因.

对任意大小为 $m\times n$ 的矩阵 X，$X^TXv = \lambda v$

令 $Xv = \sigma u$，那么 $X^T\sigma u = \lambda v$，

分别左乘A得 $XX^Tu = \frac{\lambda}{\sigma}Xv = \lambda u$，$u$对应$XX^T$的特征值为$\lambda$的特征向量.

在 $Xv = \sigma u$ 两边分别乘以u^T，得 $u^TXv = \sigma$

在 $X^T\sigma u = \lambda v$ 两边分别乘以 v^T，得 $v^TX^Tu = \frac{\lambda}{\sigma}$

$\Rightarrow \sigma^2 = \lambda$

将 $Xv = \sigma u$ 写成矩阵形式为 $XV = U\Sigma$，其中$\Sigma = \mathrm{diag}[\sigma_1, \cdots, \sigma_r, 0, \cdots, 0]$，$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \sigma_r \geqslant 0$，$r = \mathrm{rank}(A)$

$\because$ V为正交矩阵 $\therefore$ X可分解为 $X = XVV^T = U\Sigma V^T$，$U\in R^{m\times m}$，$\Sigma\in R^{m\times n}$，$V\in R^{n\times n}$

其中 U 的列向量是 XX^T 的特征向量，V 的列向量是 X^TX 的特征向量，

Σ 的非零对角元的平方即为 XX^T 和 X^TX 的共同非零特征值.

故两种分解具有等价性，而奇异值分解所需的计算和存储的成本更小. 中心化消除量纲对数据的影响、解决模型不稳定等，故在实践中使用中心化后的样本矩阵的奇异值分解来代替.

• 求解20页的优化问题

主成分分析—求解

$$\max_{W} \operatorname{tr}(\boldsymbol{W}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W}) \quad \text{s.t. } \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}_{d'}$$

• 使用拉格朗日乘子法可得

$$\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{W} = \boldsymbol{\Lambda}\boldsymbol{W}$$

只需对协方差矩阵$\boldsymbol{X}\boldsymbol{X}^\top$进行特征值分解，并将求得的特征值排序：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$，再取前$d'$个特征值对应的特征向量构成$\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d'})$

这就是主成分分析的解。

$$\max \ \operatorname{tr}(W^T X X^T W) \quad s.t. \ W^T W = I_{d'}$$

$$\Leftrightarrow \min \ -\operatorname{tr}(W^T X X^T W) \quad s.t. \ W^T W = I_{d'}$$

其拉格朗日函数为 $L(W, \Lambda) = -\operatorname{tr}(W^T X X^T W) + (W^T W - I)\Lambda$

其中 $\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_{d'} \end{bmatrix} \in R^{d' \times d'}$, $I = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \in R^{d' \times d'}$, $W \in R^{d \times d'}$

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

$$= -XX^T W - (XX^T)^T W + 2W\Lambda = -2XX^T W + 2W\Lambda$$

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

$XX^T \in R^{d \times d}$ 有d个特征向量，我们只需要其中d'个.

对 $XX^T W = W\Lambda$ 两边同时左乘 W^T，得 $W^T X X^T W = W^T W \Lambda = \Lambda$

我们的优化目标 $\operatorname{tr}(W^T X X^T W) = \operatorname{tr}(\Lambda) = \sum_{i=1}^{d} \lambda_i$

∴ 最大化迹即选择最大的d'个λ_i，和它们对应的特征向量w_i组成矩阵 $W = (w_1, w_2, \cdots w_{d'})$

这就是主成分分析的解.

附加题：令$\mathbf{M}=PP^T$，那么下列问题还是凸优化问题吗？试证明之。

$$\min_P \sum_{(x_i,x_j)\in\mathcal{M}} \|x_i-x_j\|_{\mathbf{M}}^2 \quad \text{s.t.} \sum_{(x_i,x_j)\in\mathcal{C}} \|x_i-x_j\|_{\mathbf{M}}^2 \geq 1$$

目标函数：$f_0 = \sum \|x_i - x_j\|_M^2 = \sum (x_i-x_j)^T PP^T (x_i-x_j)$

不等式约束：$f_1 = 1 - \sum \|x_i - x_j\|_M^2 = 1-\sum (x_i-x_j)^T PP^T (x_i-x_j)$

$$\frac{\partial f_0}{\partial P} = \sum\left[(x_i-x_j)(x_i-x_j)^T P + (x_i-x_j)(x_i-x_j)^T P\right] = \sum\left[2(x_i-x_j)(x_i-x_j)^T P\right]$$

$$\frac{\partial f_0}{\partial P\partial P^T} = \sum 2(x_i-x_j)(x_i-x_j)^T$$ ，元素不会为 0

$$\frac{\partial f_1}{\partial P} = -\sum\left[2(x_i-x_j)(x_i-x_j)^T P\right]$$

$$\frac{\partial f_1}{\partial P\partial P^T} = -\sum 2(x_i-x_j)(x_i-x_j)^T$$

$\frac{\partial f_0}{\partial P\partial P^T}$ 与 $\frac{\partial f_1}{\partial P\partial P^T}$ 异号，故 f_0 与 f_1 不可能同时为凸函数，故该问题不是凸优化问题.

11.5 结合图 11.2, 试举例说明 L_1 正则化在何种情形下不能产生稀疏解.

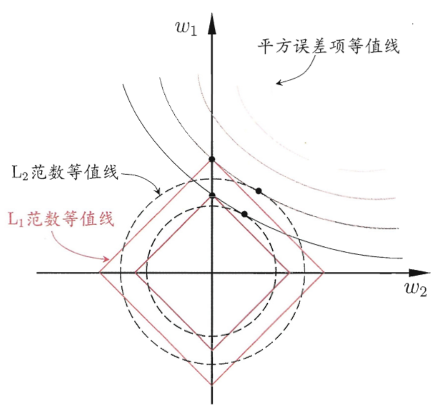

图 11.2 L_1 正则化比 L_2 正则化更易于得到稀疏解

对于 L1 范数而言，约束为红色菱形框及其内，

我们要在这之中找到一点尽可能使得平方误差项更小.

当平方误差等值线在第一、三象限存在斜率为 -1 的点，

或在第二、四象限存在斜率为 1 的点时，

其与 L1 等值线的第一个交点不在坐标轴上，

此时，无法产生稀疏解.

11.7 试述直接求解 L_0 范数正则化会遇到的困难.

L_0 范数是非零元素的个数，是很好的“稀疏约束”，

但 L_0 范数不连续，非凸，难以优化求解，

只能采用遍历的方法，设特征数为 N，则要在 2^N 的情况下分别求解问题，因此比较困难.

- PPT 20页：证明回归和对率回归的损失函数的梯度是否满足L-Lipschitz条件，并求出L

L_1正则化问题的求解(1)

20

- 可使用近端梯度下降(Proximal Gradient Descend，简称PGD) 求解

[Boyd and Vandenberghe, 2004]

- 考虑更一般的问题

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1$$

- 假设$f(\boldsymbol{w})$为凸函数，且$\nabla f(\boldsymbol{w})$满足L-Lipschitz条件， 即存在常数 $L>0$, 使得

$$\forall \boldsymbol{w}, \boldsymbol{w}' \quad \|\nabla f(\boldsymbol{w}') - \nabla f(\boldsymbol{w})\| \le L\|\boldsymbol{w}' - \boldsymbol{w}\|$$

线性回归的损失函数: $f(w) = (y - Xw)^T(y - Xw)$

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

$$\|\nabla f(w') - \nabla f(w)\| = \|2X^T(Xw' - y) - 2X^T(Xw - y)\| = \|2X^TX(w' - w)\| \le 2\|X^TX\|\,\|w' - w\|$$

取 $L \ge 2\|X^TX\|$, 即满足 $\forall\, w, w'$ $\|\nabla f(w') - \nabla f(w)\| \le L\|w' - w\|$

对率回归的损失函数: $f(w) = \sum_{i=1}^{m}(-y_i w^T x_i + \ln(1 + e^{w^T x_i}))$

$$\nabla f(w) = \sum_{i=1}^{m}\left(-y_i + \frac{1}{1 + e^{-w^T x_i}}\right)x_i$$

$$\|\nabla f(w') - \nabla f(w)\| = \left\|\sum_{i=1}^{m} x_i\left(\frac{1}{1 + e^{-w'^T x_i}} - \sum_{i=1}^{m}\frac{1}{1 + e^{-w^T x_i}}\right)\right\|$$

设 $g(w) = \frac{1}{1 + e^{-w^T x}}$, $\frac{\partial g(w)}{\partial w} = -\frac{e^{-w^T x}\cdot(-x)}{(1 + e^{-w^T x})^2} = \frac{x e^{-w^T x}}{(1 + e^{-w^T x})^2}$

设 $a = e^{-w^T x}$, 则 $a > 0$ $\quad \frac{e^{-w^T x}}{(1 + e^{-w^T x})^2} = \frac{a}{(1+a)^2} = \frac{a}{a^2 + 2a + 1} = \frac{1}{a + \frac{1}{a} + 2} \le \frac{1}{4}$

故 $\frac{\partial g(w)}{\partial w} = \frac{x e^{-w^T x}}{(1 + e^{-w^T x})^2} \le \frac{1}{4}x$

由微分中值定理: $\frac{g(w') - g(w)}{w' - w} = g'(\xi)$, $\xi \in (w', w)$ 或 (w, w')

$$\le \frac{1}{4}x$$

故 $\|\nabla f(w') - \nabla f(w)\| = \left\|\sum_{i=1}^{m} x_i\left(\frac{1}{1 + e^{-w'^T x_i}} - \sum_{i=1}^{m}\frac{1}{1 + e^{-w^T x_i}}\right)\right\| \le \left\|\sum_{i=1}^{m} x_i \cdot \frac{1}{4}x_i(w' - w)\right\|$

$$\le \frac{1}{4}\left\|\sum_{i=1}^{m} x_i^2\right\|\,\|w' - w\|$$

取 $L \ge \frac{1}{4}\|X^TX\|$, 即满足 $\forall\, w, w'$ $\|\nabla f(w') - \nabla f(w)\| \le L\|w' - w\|$