HW10 & HW11 Reference

注意：方法不唯一，言之成理即可！

1 HW10

1.1 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 $(\sum_i a_i)^2 \le n(\sum_i a_i^2)$。

证明. 先证明左边不等式：

$$\begin{aligned}
\mathrm{err}^*(\boldsymbol{x}) &= 1 - \max_{c\in\mathcal{Y}} P(c|\boldsymbol{x}) = 1 - \max_{c\in\mathcal{Y}} P(c|\boldsymbol{x}) \cdot \sum_c P(c|\boldsymbol{z}) \\
&= 1 - \sum_c \max_{c\in\mathcal{Y}} P(c|\boldsymbol{x}) \cdot P(c|\boldsymbol{z}) \\
&\le 1 - \sum_c P(c|\boldsymbol{x}) \cdot P(c|\boldsymbol{z}) = \mathrm{err}^*(\boldsymbol{x})
\end{aligned}$$

令 $c^* = \arg\max_c P(c|\boldsymbol{x})$，再证明右边不等式：

$$\begin{aligned}
\mathrm{err}^*(\boldsymbol{x}) &= 1 - \sum_c P(c|\boldsymbol{x}) \cdot P(c|\boldsymbol{z}) \backsimeq 1 - \sum_c P(c|\boldsymbol{x})^2 \\
&\le 1 - P(c^*|\boldsymbol{x})^2 - \sum_{c\neq c^*} P(c|\boldsymbol{x})^2 \\
&\le 1 - P(c^*|\boldsymbol{x})^2 - \frac{1}{|\mathcal{Y}|-1}\left(\sum_{c\neq c^*} P(c|\boldsymbol{x})\right)^2 \\
&= 1 - P(c^*|\boldsymbol{x})^2 - \frac{1}{|\mathcal{Y}|-1}\left(1 - P(c^*|\boldsymbol{x})\right)^2 \\
&= \left(1 - P(c^*|\boldsymbol{x})\right) \cdot \left(1 + P(c^*|\boldsymbol{x}) - \frac{1}{|\mathcal{Y}|-1}\left(1 - P(c^*|\boldsymbol{x})\right)\right) \\
&= \left(1 - P(c^*|\boldsymbol{x})\right) \cdot \left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1}\left(1 - P(c^*|\boldsymbol{x})\right)\right) \\
&= \mathrm{err}^*(\boldsymbol{x}) \cdot \left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \cdot \mathrm{err}^*(\boldsymbol{x})\right)
\end{aligned}$$

综上所述，证毕！ □

1.2 在实践中，协方差矩阵 $\mathbf{X}\mathbf{X}^\top$ 的特征值分解常由中心化后的样本矩阵 $\mathbf{X}$ 的奇异值分解替代，试述其原因。

解. • 仅供参考，言之成理即可。

令 $\mathbf{X}\mathbf{X}^\top$ 的特征值分解为

$$\mathbf{X}\mathbf{X}^\top = \mathbf{Y}\mathbf{\Lambda}\mathbf{Y}^\top \tag{1}$$

令 $\mathbf{X}$ 的奇异值分解为 $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$，可得

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top$$

因为 $\mathbf{X}$ 是经过中心化的样本矩阵，因此 $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$，$\mathbf{U}^\top\mathbf{U} = \mathbf{I}$，所以

$$\mathbf{X}\mathbf{X}^\top = \mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\top)\mathbf{U}^\top \tag{2}$$

如果令 $\mathbf{Y} = \mathbf{U}$、$\mathbf{\Lambda} = \mathbf{\Sigma}\mathbf{\Sigma}^\top$，不难发现式(1)和(2)是等价的，也是就是协方差矩阵的特征值分解与中心化后的样本矩阵的奇异值分解其实是等价的。

除此外，相较于特征值分解，奇异值分解的运算要更加高效，节省存储空间。 □

1.3 求解优化问题

$$\begin{aligned} \max_{\mathbf{W}} \quad & \operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) \\ \text{s.t.} \quad & \mathbf{W}^\top\mathbf{W} = \mathbf{I}_{d'} \end{aligned}$$

解. 先将问题转化为等价问题，

$$\begin{aligned} \min_{\mathbf{W}} \quad & -\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) \\ \text{s.t.} \quad & \mathbf{W}^\top\mathbf{W} = \mathbf{I}_{d'} \end{aligned}$$

然后使用拉格朗日乘子法，构造拉格朗日函数

$$L(\mathbf{W}, \mathbf{\Lambda}) = -\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) + \operatorname{tr}\left(\mathbf{\Lambda}(\mathbf{W}^\top\mathbf{W} - \mathbf{I}_{\mathbf{d'}})\right)$$

其中，$\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \cdots, \lambda_{d'})$，于是令

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{W}} &= -\frac{\partial}{\partial \mathbf{W}}\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) + \frac{\partial}{\partial \mathbf{W}}\operatorname{tr}\left(\mathbf{\Lambda}(\mathbf{W}^\top\mathbf{W} - \mathbf{I}_{\mathbf{d'}})\right) \\ &= -2\mathbf{X}\mathbf{X}^\top\mathbf{W} + 2\mathbf{W}\mathbf{\Lambda} \\ &= 0 \end{aligned}$$

解得

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

这意味着

$$\mathbf{X}\mathbf{X}^\top\boldsymbol{w}_i = \lambda_i\boldsymbol{w}_i$$

也就是说，取 $\mathbf{X}\mathbf{X}^\top$ 的最大的前 d' 个特征值所对应的特征向量即可得 $\mathbf{W}$。将之代入到目标函数即可得

$$\operatorname{tr}\left(\mathbf{W}^\top\mathbf{X}\mathbf{X}^\top\mathbf{W}\right) = \sum_{i=1}^{d'}\lambda_{d'}.$$

□

1.4　令 $\mathbf{M}=\mathbf{P}\mathbf{P}^\top$，那么下列问题还是凸优化问题吗？试证明之。

$$\begin{aligned}\min_{\mathbf{P}} \quad & \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \| \boldsymbol{x}_i-\boldsymbol{x}_j \|_{\mathbf{M}}^2 \\ \text{s.t.} \quad & \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{C}} \| \boldsymbol{x}_i-\boldsymbol{x}_j \|_{\mathbf{M}}^2 \geq 1\end{aligned}$$

凸优化问题一般具有如下形式

$$\begin{aligned}\min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x)\leq b_i,\ i=1,\dots,m \\ & h_i(x)=0,\ i=1,\dots,p\end{aligned}$$

其中函数 $f_0,f_1,\dots,f_m$ 是凸函数，$h_0,h_1,\dots,h_p$ 是仿射函数。

证明. 令

$$\begin{aligned}&\boldsymbol{\Delta}_{i,j}=\boldsymbol{x}_i-\boldsymbol{x}_j \\ &f(\mathbf{P})=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \| \boldsymbol{x}_i-\boldsymbol{x}_j \|_{\mathbf{M}}^2=\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \boldsymbol{\Delta}_{i,j}^\top\mathbf{P}\mathbf{P}^\top\boldsymbol{\Delta}_{i,j} \\ &g(\mathbf{P})=1-\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{C}} \boldsymbol{\Delta}_{i,j}^\top\mathbf{P}\mathbf{P}^\top\boldsymbol{\Delta}_{i,j}\end{aligned}$$

此时，可将上述优化问题转化为

$$\begin{aligned}\min \quad & f(\mathbf{P}) \\ \text{s.t.} \quad & g(\mathbf{P})\leq 0\end{aligned}$$

根据凸优化问题的一般形式可知，只要证明 $f(\mathbf{P})$ 和 $g(\mathbf{P})$ 同时都为凸函数，即可证该问题是一个凸优化问题。接下来证明 $f(\mathbf{P})$ 和 $g(\mathbf{P})$ 是否为凸函数：

$$\begin{aligned}\frac{\partial f}{\partial \mathbf{P}}&=2\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \boldsymbol{\Delta}_{i,j}\boldsymbol{\Delta}_{i,j}^\top\mathbf{P} \\ \frac{\partial^2 f}{\partial \mathbf{P}\partial \mathbf{P}^\top}&=2\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{M}} \boldsymbol{\Delta}_{i,j}^\top\boldsymbol{\Delta}_{i,j} \\ \frac{\partial g}{\partial \mathbf{P}}&=-2\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{C}} \boldsymbol{\Delta}_{i,j}\boldsymbol{\Delta}_{i,j}^\top\mathbf{P} \\ \frac{\partial^2 g}{\partial \mathbf{P}\partial \mathbf{P}^\top}&=-2\sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{C}} \boldsymbol{\Delta}_{i,j}^\top\boldsymbol{\Delta}_{i,j}\end{aligned}$$

因为 $\boldsymbol{\Delta}_{i,j}^\top\boldsymbol{\Delta}_{i,j}=(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)\geq 0$，那么

$$\begin{aligned}\frac{\partial^2 f}{\partial \mathbf{P}\partial \mathbf{P}^\top}&\geq 0 \\ \frac{\partial^2 g}{\partial \mathbf{P}\partial \mathbf{P}^\top}&\leq 0\end{aligned}$$

因此 $f(\mathbf{P})$ 是凸函数，$g(\mathbf{P})$ 不是凸函数，所以该问题不是凸优化问题。 □

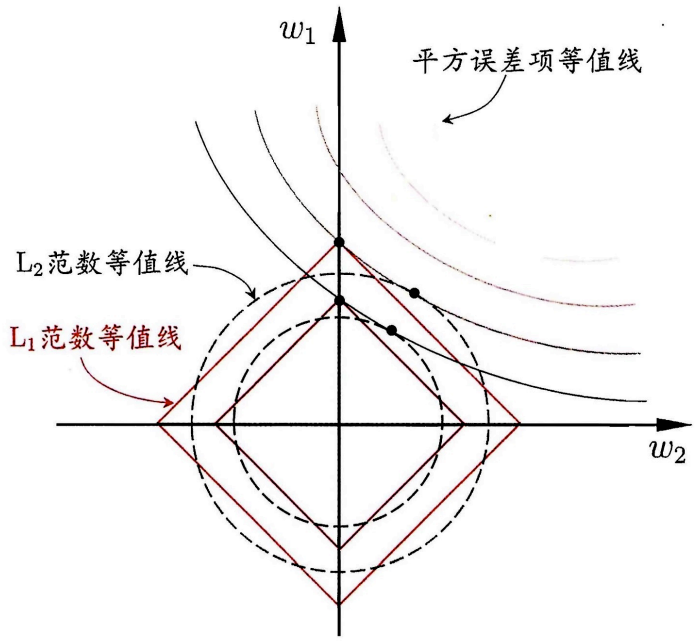

图 11.2 L_1 正则化比 L_2 正则化更易于得到稀疏解

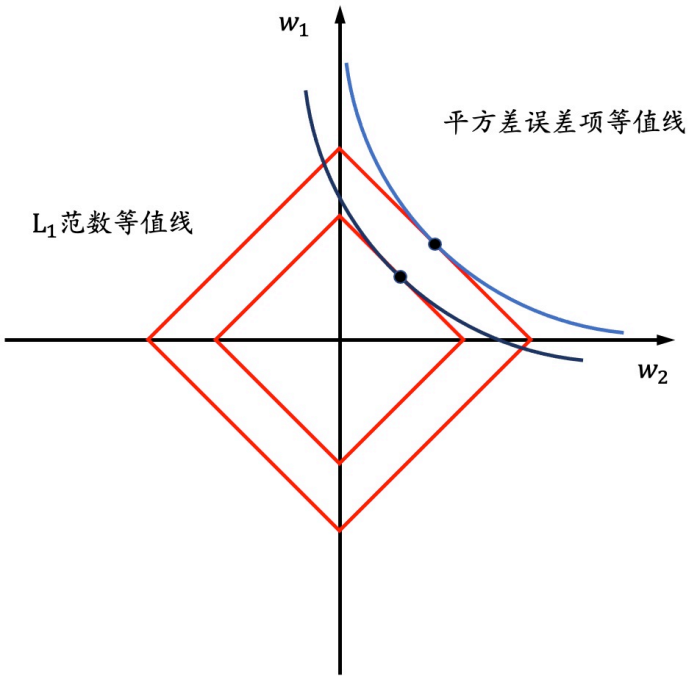

图 1: 情况演示图

2 HW11

2.1 [课本习题 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。

解. 如图1所示，当平方差误差项等值线的斜率较大的时候，其与 L_1 范数等值线的交点就不再位于坐标轴上，因此将无法产生稀疏解。

□

2.2 [课本习题 11.7] 试述直接求解 L_0 范数正则化会遇到的困难。

解. L_0 范数是统计向量非零元素的个数，不连续、不可微、非凸，无法通过凸优化的方式求解，需要采用遍历方式才能找到最优解，因此难度是 NP-难的。 □

2.3 [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。

证明. 先证明线性回归函数，其损失函数为

$$E(\boldsymbol{w}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{w})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{w})$$

不难发现该函数为凸函数，其微分算子 ∇E 表示为

$$\nabla E(\boldsymbol{w}) = 2\mathbf{X}^\top (\mathbf{X}\boldsymbol{w} - \boldsymbol{y})$$

对于 $\forall \boldsymbol{w}, \boldsymbol{w}'$，都有

$$\begin{aligned} \| \nabla E(\boldsymbol{w}) - \nabla E(\boldsymbol{w}') \|_2 &= \| 2\mathbf{X}^\top \mathbf{X}(\boldsymbol{w} - \boldsymbol{w}') \|_2 \\ &\leq 2 \| \mathbf{X}^\top \mathbf{X} \|_2 \cdot \| \boldsymbol{w} - \boldsymbol{w}' \|_2 \end{aligned}$$

令 $L = 2 \| \mathbf{X}^\top \mathbf{X} \|_2 > 0$，可以发现线性回归函数的损失函数满足 L-Lipschitz 条件。

接着证明对率回归函数，其损失函数为

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(-y_i \boldsymbol{\beta}^\top \boldsymbol{x}_i + \ln \left(1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i} \right) \right)$$

该函数是关于 $\boldsymbol{\beta}$ 的高阶可导连续凸函数，其微分算子 $\nabla \ell$ 表示为

$$\nabla \ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(-y_i \boldsymbol{x}_i + \frac{\boldsymbol{x}_i e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta}^\top \boldsymbol{x}_i}} \right) = \sum_{i=1}^{m} \left(\frac{1}{1 + e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i}} - y_i \right) \boldsymbol{x}_i$$

对于 $\forall \boldsymbol{\beta}, \boldsymbol{\beta}'$，都有

$$\| \nabla \ell(\boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta}') \|_2 = \| \sum_i \left(\frac{1}{1 + e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i}} - \frac{1}{1 + e^{-\boldsymbol{\beta}'^\top \boldsymbol{x}_i}} \right) \boldsymbol{x}_i \|_2$$

注意到 Sigmod 函数 $f(x) = \frac{1}{1+e^{-x}}$ 上任意两点连线的斜率小于等于 $f'(0)$，因此可知

$$\frac{1}{1 + e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i}} - \frac{1}{1 + e^{-\boldsymbol{\beta}'^\top \boldsymbol{x}_i}} \leq \left(\frac{1}{1 + e^{-\boldsymbol{\beta}^\top \boldsymbol{x}_i}} \right)' \Bigg|_{\boldsymbol{\beta}^\top \boldsymbol{x}_i = 0} \left(\boldsymbol{\beta}^\top - \boldsymbol{\beta}'^\top \right) \boldsymbol{x}_i = \frac{1}{4} \boldsymbol{x}_i^\top \left(\boldsymbol{\beta} - \boldsymbol{\beta}' \right)$$

因此可得

$$\begin{aligned} \| \nabla \ell(\boldsymbol{\beta}) - \nabla \ell(\boldsymbol{\beta}') \|_2 &\leq \| \sum_i \frac{1}{4} \boldsymbol{x}_i^\top \left(\boldsymbol{\beta} - \boldsymbol{\beta}' \right) \boldsymbol{x}_i \|_2 \\ &\leq \frac{1}{4} \| \sum_i \boldsymbol{x}_i^\top \boldsymbol{x}_i \|_2 \cdot \| \boldsymbol{\beta} - \boldsymbol{\beta}' \|_2 \end{aligned}$$

令 $L = \frac{1}{4} \| \sum_i \boldsymbol{x}_i^\top \boldsymbol{x}_i \|_2 > 0$，可以发现对率回归函数的损失函数满足 L-Lipschitz 条件。

□