

HW9 & HW10 Reference

注意：方法不唯一，言之成理即可！

1 HW9

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

$$\text{err}^*(\mathbf{x}) \leq \text{err}(\mathbf{x}) \leq \text{err}^*(\mathbf{x}) \left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} \times \text{err}^*(\mathbf{x}) \right)$$

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

证明。先证明左边不等式：

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

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

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

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

$$\mathbf{X}\mathbf{X}^\top = \mathbf{Y}\mathbf{\Lambda}\mathbf{Y}^\top \quad (1)$$

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

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

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，也就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。 \square

1.3 求解优化问题

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

解. 先将问题转化为等价问题，

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

然后使用拉格朗日乘子法，构造拉格朗日函数

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

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

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

解得

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

这意味着

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

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

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

\square

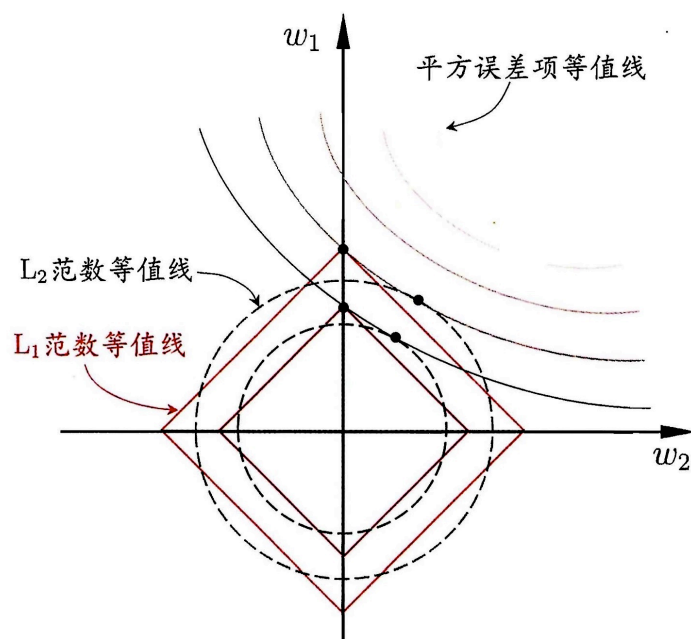


图 11.2 L_1 正则化比 L_2 正则化更易于得到稀疏解

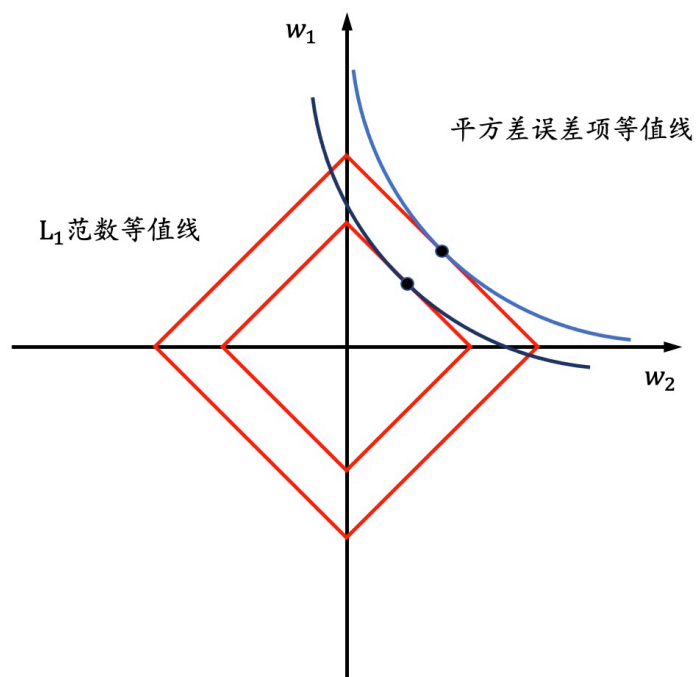


图 1: 情况演示图

2 HW10

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

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

不难发现该函数为凸函数, 其微分算子 ∇E 表示为

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

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

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

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

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

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

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

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

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

注意到 Sigmoid 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

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

因此可得

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

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