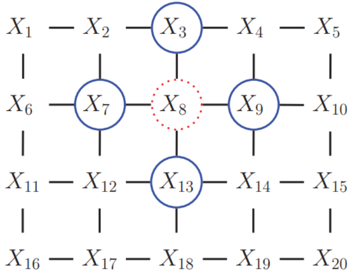
机器学习概论习题课

Yuanhao Pu & Jin Zhang

2024 年 1 月 11 日

目录

Ising Model



统计物理建模磁性材料中原子的行为。$X_i \in \{-1, +1\} \to$ 表示原子自旋方向；令 $s \sim t$ 表示 s 与 t 互为邻居，则（未归一化的）先验分布为

$$\log \tilde{p}(\boldsymbol{x}) = -\sum_{s\sim t} w_{st} x_s x_t - \sum_s b_s x_s$$

$$= -\frac{1}{2}\boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x} (\text{有时忽略 bias})$$

Image Denoising

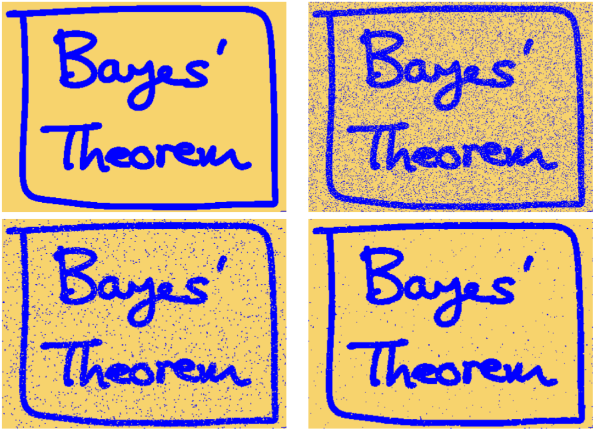
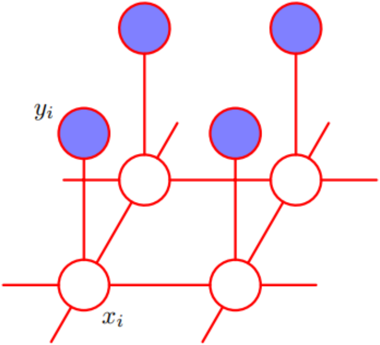


Figure 8.30 Illustration of image de-noising using a Markov random field. The top row shows the original binary image on the left and the corrupted image after randomly changing 10% of the pixels on the right. The bottom row shows the restored images obtained using iterated conditional models (ICM) on the left and using the graph-cut algorithm on the right. ICM produces an image where 96% of the pixels agree with the original image, whereas the corresponding number for graph-cut is 99%.

Ising Model for Image Denoising



$y_i \in \{-1, +1\}$ 含噪声图中的像素点
$x_i \in \{-1, +1\}$ 无噪声图中的像素点
建模联合分布

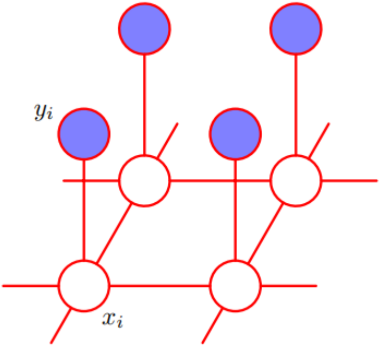
$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$$

其中先验 $p(\mathbf{x})$ 有

$$p(\mathbf{x}) = \frac{1}{Z_0}\exp(-E_0(\mathbf{x}))$$

$$E_0(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^\top \mathbf{W}\mathbf{x} = -\sum_{i=1}^{D}\sum_{j\in \mathrm{nbr}_i} w_{ij}x_i x_j$$

Ising Model for Image Denoising



设 L_i 为 $x \sim y$ 间的能量函数，

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_i p(y_i|x_i) \propto \exp(-\sum_i L_i(x_i))$$

则联合分布

$$p(x, y) = \frac{1}{Z}\exp(-E(\boldsymbol{x}))$$

$$E(\boldsymbol{x}) = E_0(x) - \sum_i L_i(x_i)$$

Mean field

由平均场假设，考虑

$$q(\boldsymbol{x}) = \prod_i q(x_i, \mu_i)$$

为了得到 μ_i 的更新策略，展开 $\log \tilde{p}(\boldsymbol{x})$ 并提取出与 x_i 相关的项：

$$\log \tilde{p}(\boldsymbol{x}) = x_i \sum_{j \in \text{nbr}_i} w_{ij} x_j + L_i(x_i) + const$$

基于课上的结论，对所有的 $j \neq i$ 求期望，

$$q_i(x_i) \propto \exp(x_i \sum_{j \in \text{nbr}_i} w_{ij} \mu_j + L_i(x_i))$$

进一步的，令

$$m_i = \sum_{j \in \mathrm{nbr}_i} w_{ij}\mu_j, \quad L_i^+ = L_i(+1), \quad L_i^- = L_i(-1)$$

则边缘后验可以直接求得

$$\begin{aligned} q_i(x_i = 1) &= \frac{\exp(1 * m_i + L_i^+)}{\exp(1 * m_i + L_i^+) + \exp(-1 * m_i + L_i^-)} \\ &= \frac{1}{1 + \exp(-2m_i + L_i^- - L_i^+)} = \sigma(2a_i) \end{aligned}$$

其中 $a_i = m_i + 0.5(L_i^+ - L_i^-)$，则参数 μ_i 有

$$\mu_i = \mathbb{E}_{q_i}[x_i] = q_i(x_i = +1) - q_i(x_i = -1) = tanh(a_i)$$

$$\text{Update:} \mu_i^t = tanh\Big(\sum_{j \in \mathrm{nbr}_i} w_{ij}\mu_j^{t-1} + 0.5(L_i^+ - L_i^-)\Big)$$

Bayesian Linear Regression

对线性模型 $y = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + \varepsilon, \varepsilon \sim \mathcal{N}(0, \lambda^{-1})$
写作分布形式 $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \lambda) = \prod_{n=1}^{N} \mathcal{N}(y_n|\boldsymbol{w}^\top \phi(\boldsymbol{x}_n), \lambda^{-1})$
Bayesian： 引入关于 $\boldsymbol{w}$ 的先验分布 $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$
则由 $p(\boldsymbol{w}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})$， 得到后验

$$p(\boldsymbol{w}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N, \boldsymbol{S}_N)$$
$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \lambda\boldsymbol{\Phi}^\top \boldsymbol{y})$$
$$\boldsymbol{S}_N^{-1} = \boldsymbol{S}_0^{-1} + \lambda\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$$

(过程详见 PRML)

Bayesian Linear Regression

实际操作时考虑简化问题，$p(\boldsymbol{w}|\alpha)=\mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\alpha^{-1}\boldsymbol{I})$，则

$$\boldsymbol{m}_N=\lambda S_N\boldsymbol{\Phi}^\top\boldsymbol{y}$$
$$\boldsymbol{S}_N^{-1}=\alpha\boldsymbol{I}+\lambda\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$$

代入对数似然，并提取出相关项

$$\log p(\boldsymbol{w}|\boldsymbol{y})=-\frac{\lambda}{2}\sum_{n=1}^{N}(y_n-\boldsymbol{w}^\top\phi(\boldsymbol{x}_n))^2-\frac{\alpha}{2}\boldsymbol{w}^\top\boldsymbol{w}+\text{const}$$

能证明，最大化梯度似然与最小化岭回归损失等价 (正则项系数 $\tau=\frac{\alpha}{\lambda}$)

Variational Bayesian for Linear Regression

若假定 $\boldsymbol{\eta} = (\lambda, \alpha)$ 均未知，选取超参数使边缘似然达到最大的方式称为 evidence procedure；
利用变分 Bayes 估计参数，取先验

$$p(\boldsymbol{W}, \lambda, \alpha) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, (\lambda\alpha)^{-1}\boldsymbol{I})\, Ga(\lambda|a_N^{\lambda}, b_N^{\lambda})\, Ga(\alpha|a_N^{\alpha}, b_N^{\alpha})$$

假设变分分布满足

$$q(\boldsymbol{w}, \lambda, \alpha) = q(\boldsymbol{w}, \lambda) q(\alpha)$$

Variational Bayesian for Linear Regression

$$q(\mathbf{w}, \alpha, \lambda) = \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \lambda^{-1}\mathbf{V}_N)\mathrm{Ga}(\lambda|a_N^\lambda, b_N^\lambda)\mathrm{Ga}(\alpha|a_N^\alpha, b_N^\alpha)$$

where

$$\begin{aligned}
\mathbf{V}_N^{-1} &= \overline{\mathbf{A}} + \mathbf{X}^{\mathbf{X}} \\
\mathbf{w}_N &= \mathbf{V}_N \mathbf{X}^T \mathbf{y} \\
a_N^\lambda &= a_0^\lambda + \frac{N}{2} \\
b_N^\lambda &= b_0^\lambda + \frac{1}{2}(||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 + \mathbf{w}_N^T \overline{\mathbf{A}} \mathbf{w}_N) \\
a_N^\alpha &= a_0^\alpha + \frac{D}{2} \\
b_N^\alpha &= b_0^\alpha + \frac{1}{2}\left(\frac{a_N^\lambda}{b_N^\lambda}\mathbf{w}_N^T\mathbf{w}_N + \mathrm{tr}(\mathbf{V}_N)\right) \\
\overline{\mathbf{A}} &= \langle\alpha\rangle\mathbf{I} = \frac{a_N^\alpha}{b_N^\alpha}\mathbf{I}
\end{aligned}$$

Variational Bayesian for Linear Regression

Update: 交替更新 $q(\boldsymbol{w}, \lambda)$ 和 $q(\alpha)$

$$p(\mathcal{D}) = \int\int\int p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \lambda)p(\mathbf{w}|\alpha)p(\lambda)d\mathbf{w}d\alpha d\lambda$$

We can compute a lower bound on $\log p(\mathcal{D})$ as follows:

$$\begin{aligned} L(q) &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{a_N^\lambda}{b_N^\lambda}(y_i - \mathbf{w}_N^T\mathbf{x}_i)^2 + \mathbf{x}_i^T\mathbf{V}_N\mathbf{x}_i\right) \\ &+\frac{1}{2}\log|\mathbf{V}_N| + \frac{D}{2} \\ &-\log\Gamma(a_0^\lambda) + a_0^\lambda\log b_0^\lambda - b_0^\lambda\frac{a_N^\lambda}{b_N^\lambda} + \log\Gamma(a_N^\lambda) - a_N^\lambda\log b_N^\lambda + a_N^\lambda \\ &-\log\Gamma(a_0^\alpha) + a_0^\alpha\log b_0^\alpha + \log\Gamma(a_N^\alpha) - a_N^\alpha\log b_N^\alpha \end{aligned}$$

推导过于复杂，重点掌握使用变分推断的流程，并理解 ELBO 为边缘似然的下界（参考 github 上 HW11&HW12.pdf 往年作业题）

Boltzmann Machine

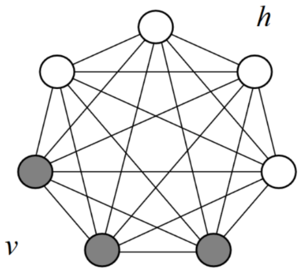


图 1: Boltzmann machine 模型的概率图

可观测变量 $v_i \in \{0,1\}$
隐变量 $h_i \in \{0,1\}$
能量函数

$$E(\boldsymbol{v},\boldsymbol{h}) = -(\sum_{i<j} w_{ij}x_ix_j + \sum_i b_ix_i)$$

$$= -(\boldsymbol{v}^\top W\boldsymbol{h} + \frac{1}{2}\boldsymbol{v}^\top L\boldsymbol{v} + \frac{1}{2}\boldsymbol{h}^\top J\boldsymbol{h} + \boldsymbol{b}_v^\top \boldsymbol{v} + \boldsymbol{b}_h^\top \boldsymbol{h})$$

$$P(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{Z}\exp(-E(\boldsymbol{v},\boldsymbol{h}))$$

VI for BM

变分推断通过 ELBO 优化变分分布 $Q_\phi(\boldsymbol{h}|\boldsymbol{v})$ 推断 $P(\boldsymbol{h}|\boldsymbol{v})$

$$ELBO = \mathcal{L} = \sum_h Q_\phi(\boldsymbol{h}|\boldsymbol{v}) \log P_\theta(\boldsymbol{v}, \boldsymbol{h}) + H(Q_\phi)$$

$$= \sum_h Q_\phi(\boldsymbol{h}|\boldsymbol{v})[-\log Z + \boldsymbol{v}^\top W\boldsymbol{h} + \frac{1}{2}\boldsymbol{v}^\top L\boldsymbol{v} + \frac{1}{2}\boldsymbol{h}^\top J\boldsymbol{h} + \boldsymbol{b}_v^\top \boldsymbol{v} + \boldsymbol{b}_h^\top \boldsymbol{h}] + H(Q$$

基于平均场理论，拆分 $Q_\phi(\boldsymbol{h}|\boldsymbol{v}) = \prod_{j=1}^P Q_\phi(h_j|\boldsymbol{v})$,$h_j$ 是二值的，令 $Q_\phi(h_j = 1|v) = \phi_j$, 只考虑 $\mathcal{L}$ 中与 ϕ_j(即h_j) 相关的项

$$\mathcal{L} = \sum_h Q_\phi(\boldsymbol{h}|\boldsymbol{v})[\boldsymbol{v}^\top W\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}^\top J\boldsymbol{h} + \boldsymbol{b}_h^\top \boldsymbol{h}] + H(Q_\phi)$$

$$= \sum_h Q_\phi(\boldsymbol{h}|\boldsymbol{v})\boldsymbol{v}^\top W\boldsymbol{h} + \sum_h Q_\phi(\boldsymbol{h}|\boldsymbol{v})\frac{1}{2}\boldsymbol{h}^\top J\boldsymbol{h} + \sum_h Q_\phi(\boldsymbol{h}|\boldsymbol{v})\boldsymbol{b}_h^\top \boldsymbol{h} + H(Q_\phi)$$

VI for BM

$$\mathcal{L} = ① + ② + ③ + ④$$

$$① = \sum_i \sum_j \phi_j v_i w_{ij} \quad ② = \sum_j \sum_{m \neq j} \phi_j \phi_m J_{jm} \quad ③ = \sum_j \phi_j b_{hj}$$

$$④ = -\sum_j [\phi_j \log \phi_j + (1 - \phi_j) \log(1 - \phi_j)]$$

$$\frac{\partial \mathcal{L}}{\partial \phi_j} = 0 \implies \phi_j = \sigma(\sum_i v_i w_{ij} + \sum_{m \neq j} \phi_m J_{jm} + b_{hj})$$

迭代求解所有的 ϕ_j 得到 $Q_\phi(\boldsymbol{h}|\boldsymbol{v}) \simeq P(\boldsymbol{h}|\boldsymbol{v})$

1.3

已知随机变量 $\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2]\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$，计算 $P(\boldsymbol{x}_1)$，$P(\boldsymbol{x}_1\mid\boldsymbol{x}_2)$

Solve.

$$x_1\sim\mathcal{N}\left(\mu_1,\Sigma_{11}\right).$$

待定 λ 使得 $x_1-\lambda x_2$ 与 x_2 相互独立，等价于

$$\mathrm{Cov}\left(x_1-\lambda x_2,x_2\right)=0\Rightarrow\lambda=\Sigma_{22}^{-1}\Sigma_{12}$$

$x_1\,|x_2=(x_1-\lambda x_2)|\,x_2+\lambda x_2\,|x_2=(x_1-\lambda x_2)+\lambda x_2|\,x_2$ 是一正态分布，求出均值和标准差即可

$$\mathbb{E}\left[x_1\mid x_2\right]=\mu_1-\lambda\mu_2+\lambda x_2$$
$$\mathrm{Cov}\left(x_1-\lambda x_2,x_1-\lambda x_2\right)=\mathrm{Cov}\left(x_1-\lambda x_2,x_1\right)=\Sigma_{11}-\lambda\Sigma_{21}$$

$$x_1\mid x_2\sim\mathcal{N}\left(\mu_1-\lambda\mu_2+\lambda x_2,\Sigma_{11}-\Sigma_{22}^{-1}\Sigma_{12}\Sigma_{21}\right)$$

3.1

[课本习题 3.2] 试证明，对于参数 w，对率回归的目标函数 (3.18) 是非凸的，但其对数似然函数 (3.27) 是凸的。

Sol.

$$y = \frac{1}{1 + \mathrm{e}^{-\boldsymbol{w}^\top \boldsymbol{x} + b}}$$

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

$$\frac{\partial y}{\partial \boldsymbol{w}} = \frac{\boldsymbol{x} e^{-\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)}}{\left(1 + e^{-\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)}\right)^2} = \boldsymbol{x} y(1-y)$$

$$\frac{\partial^2 y}{\partial \boldsymbol{w}^\top \partial \boldsymbol{w}} = \frac{\partial y}{\partial \boldsymbol{w}^\top} \boldsymbol{x}(1-y) + \frac{\partial (1-y)}{\partial \boldsymbol{w}^\top} \boldsymbol{x} y = \boldsymbol{x}\boldsymbol{x}^\top y(1-2y)(1-y)$$

3.1

$\boldsymbol{x}\boldsymbol{x}^\top \geq 0$ 恒成立，当 $0.5 < y < 1$ 时，$y(1-2y)(1-y) < 0$ ，此时 $\frac{\partial^2 y}{\partial \boldsymbol{w}^\top \partial \boldsymbol{w}} < 0$ ，因此函数 (3.18) 非凸。

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

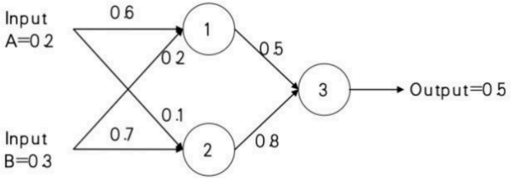
$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}^\top} \sum_{i=1}^{m} \left(-y_i \widehat{\boldsymbol{x}}_i + \frac{1}{1+\exp \boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i} \widehat{\boldsymbol{x}}_i \exp\left(\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i\right) \right)$$

$$= \sum_{i=1}^{m} \frac{\exp\left(\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i\right)}{\left(1+\exp\left(\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i\right)\right)^2} \widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top$$

由于 $\widehat{\boldsymbol{x}}_i \widehat{\boldsymbol{x}}_i^\top \geq 0$ 且 $\frac{e^{\beta^\top \hat{x}_i}}{\left(1+e^{\beta^\top \hat{x}_i}\right)^2} \geq 0$ ，因此函数 (3.27) 为凸函数。

5.4 神经网络参数更新

激活函数为 ReLU，用平方损失 $\frac{1}{2}(y-\hat{y})^2$ 计算误差，请用 BP 算法更新一次所有参数 (学习率为 1)，给出更新后的参数值给定输入值 $x=(0.2,0.3)$ 时初始时和更新后的输出值



Sol:

已知 $v_{11}=0.6, v_{12}=0.1, v_{21}=0.2, v_{22}=0.7, w_1=0.5, w_2=0.8$，令 α_1,α_2,γ 为结点 1, 2,3 的输入，$\{\beta_1,\beta_2,\hat{y}\}$ 为对应输出，

5.4

正向传播:

$$\alpha_1 = v_{11}A + v_{21}B = 0.18 = \beta_1$$
$$\alpha_2 = v_{12}A + v_{22}B = 0.23 = \beta_2$$
$$\gamma = w_1\beta_1 + w_2\beta_2 = 0.274 = \hat{y}$$
$$E = \frac{1}{2}(y - \hat{y})^2 = 0.025538$$

反向传播（链式法则）:

$$\frac{\partial E}{\partial v_{11}} = \frac{\partial E}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial \gamma}\frac{\partial \gamma}{\partial \beta_1}\frac{\partial \beta_1}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial v_{11}} = -0.0226$$

9.1

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

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

证明. 先证明左边不等式:

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

9.1

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

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

9.1

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

10.1

[课本习题 11.5] 结合图 11.2，试举例说明 $\mathbf{L}_1$ 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。

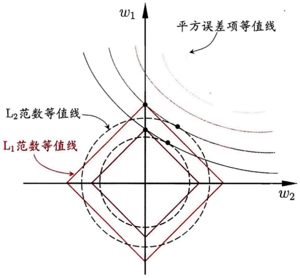


图 11.2 L_1 正则化比 L_2 正则化更易于得到稀疏解

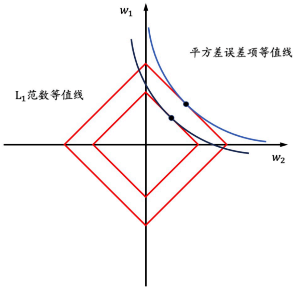


图 1: 情况演示图