第 9 章 线性回归

9.1 线性回归问题

定义 9.1.1 (线性回归假设（Linear Regression Hypothesis）)

以顾客特征为例：

$$\mathbf{x} = (x_0, x_1, x_2, x_3, x_4) = (1,\ 23,\ 1{,}000{,}000,\ 0.5,\ 200{,}000),$$

其中

- $x_0 = 1$：偏置项；
- x_1：年龄（岁）；
- x_2：年薪（新台币）；
- x_3：在职年数；
- x_4：当前债务。

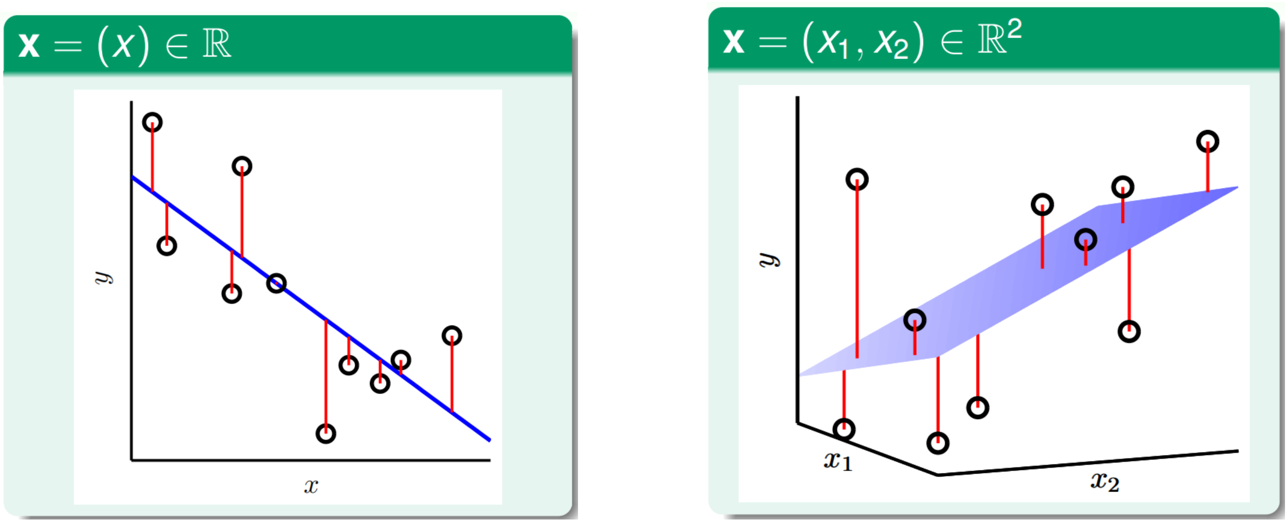
用加权线性组合近似理想信用额度

$$y \approx \sum_{i=0}^{d} w_i x_i = \mathbf{w}^\top \mathbf{x}.$$

线性回归假设

$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x},$$

与感知机形式相同，但去掉了符号函数。♣



(a) 一维特征线性回归拟合示意图

(b) 二维特征线性回归拟合示意图

图 9.1.1: 线性回归拟合示意图

命题 9.1.1 (常用误差度量：平方误差)

在回归问题中，最流行（且历史最悠久）的误差度量为平方误差

$$\mathrm{err}(\hat{y}, y) = (\hat{y} - y)^2.$$

对应的样本内与样本外误差分别为

$$E_{\mathrm{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^\top \mathbf{x}_n - y_n)^2,$$

$$E_{\mathrm{out}}(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim P}\big[(\mathbf{w}^\top \mathbf{x} - y)^2\big].$$

9.2 线性回归算法

命题 9.2.1 (平方误差的矩阵形式与最优权)

将设计矩阵与目标向量记为

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^\top - \\ \vdots \\ -x_N^\top - \end{bmatrix}_{N\times(d+1)}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1}, \quad \mathbf{w} \in \mathbb{R}^{d+1},$$

则样本内误差可写成简洁矩阵形式

$$E_{\mathrm{in}}(\mathbf{w}) = \frac{1}{N}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2.$$

性质

- $E_{\mathrm{in}}(\mathbf{w})$ 连续、可微且凸。
- 最优权 $\mathbf{w}_{\mathrm{LIN}}$ 的必要条件为梯度为零：

$$\nabla_{\mathbf{w}} E_{\mathrm{in}}(\mathbf{w}) = \frac{2}{N}\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{0}.$$

命题 9.2.2 (线性回归的梯度)

给定设计矩阵 $X \in \mathbb{R}^{N\times(d+1)}$ 与目标向量 $y \in \mathbb{R}^N$，样本内平方误差

$$E_{\mathrm{in}}(\mathbf{w}) = \frac{1}{N}\|X\mathbf{w} - y\|^2 = \frac{1}{N}\Big(\mathbf{w}^\top X^\top X\mathbf{w} - 2\mathbf{w}^\top X^\top y + y^\top y\Big).$$

标量情形（单权 w）

$$E_{\mathrm{in}}(w) = \frac{1}{N}(aw^2 - 2bw + c) \quad \Longrightarrow \quad \nabla E_{\mathrm{in}}(w) = \frac{2}{N}(aw - b).$$

向量情形（多权 $\mathbf{w}$）

$$E_{\mathrm{in}}(\mathbf{w}) = \frac{1}{N}(\mathbf{w}^\top A\mathbf{w} - 2\mathbf{w}^\top \mathbf{b} + c) \quad \text{其中 } A = X^\top X,\ \mathbf{b} = X^\top y,$$

$$\nabla E_{\mathrm{in}}(\mathbf{w}) = \frac{2}{N}(X^\top X\mathbf{w} - X^\top y).$$

定理 9.2.1 (最优线性回归权重)

任务：寻找 $\mathbf{w}_{\text{LIN}}$，使得梯度为零

$$\nabla E_{\text{in}}(\mathbf{w}) = X^\top X\mathbf{w} - X^\top y = 0.$$

- 若 $X^\top X$ 可逆（$\det(X^\top X) \neq 0$），则存在唯一解

$$\boxed{\mathbf{w}_{\text{LIN}} = (X^\top X)^{-1} X^\top y}.$$

- 若 $X^\top X$ 奇异，则最优解不唯一；可选取

$$\boxed{\mathbf{w}_{\text{LIN}} = X^\dagger y},$$

其中 $X^\dagger$ 为伪逆（pseudo-inverse），通过不同定义方式给出，通常取

$$X^\dagger = (X^\top X)^\dagger X^\top.$$

这种情况常见于 $N \geq d+1$ 且特征近似线性相关。

实践建议　为实现数值稳定，推荐使用成熟库中的线性求解器（如 SVD 或 QR 分解）直接求 $\mathbf{w}_{\text{LIN}}$，而避免显式计算 $(X^\top X)^{-1}$，尤其在 $X^\top X$ 接近奇异时。♡

算法 9.2.1: 通过伪逆求解线性回归

输入: 数据集 $\mathcal{D} = \left\{(\mathbf{x}_n, y_n)\right\}_{n=1}^N$，其中 $\mathbf{x}_n \in \mathbb{R}^{d+1}$，$y_n \in \mathbb{R}$

输出: 最优权重向量 $\mathbf{w}_{\text{LIN}} \in \mathbb{R}^{d+1}$

构造设计矩阵 $\mathbf{X} \leftarrow \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix}_{N\times(d+1)}$ 和目标向量 $\mathbf{y} \leftarrow \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1}$；

计算伪逆 $\mathbf{X}^\dagger \leftarrow (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top$;

$\mathbf{w}_{\text{LIN}} \leftarrow \mathbf{X}^\dagger\mathbf{y}$;

return $\mathbf{w}_{\text{LIN}}$.

例题 9.1 选择题：线性回归预测向量的矩阵公式

得到 $\boldsymbol{w}_{\text{LIN}}$ 后，我们可以计算预测值 $\hat{y}_n = \boldsymbol{w}_{\text{LIN}}^T\boldsymbol{x}_n$。如果将所有 $\hat{y}_n$ 收集到一个类似于我们构造 $\boldsymbol{y}$ 的向量 $\hat{\boldsymbol{y}}$ 中，那么 $\hat{\boldsymbol{y}}$ 的矩阵公式是什么？

1) $\boldsymbol{y}$
2) $\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y}$
3) $\boldsymbol{X}\boldsymbol{X}^\dagger\boldsymbol{y}$
4) $\boldsymbol{X}\boldsymbol{X}^\dagger\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{y}$

解答　正确选项为 $\boxed{3}$。由于 $\hat{\boldsymbol{y}} = \mathbf{X}\boldsymbol{w}_{\text{LIN}}$，利用 $\mathbf{w}_{\text{LIN}} = \mathbf{X}^\dagger\mathbf{y}$，则 $\hat{\boldsymbol{y}} = \mathbf{X}\mathbf{X}^\dagger\mathbf{y}$。■

9.3 泛化问题

命题 9.3.1 (解析解的优势：比 VC 界更简单的保证)

给定噪声水平 σ^2，记 $\hat{\mathbf{y}} = X\mathbf{w}_{\text{LIN}}$，则

$$E_{\text{in}}(\mathbf{w}_{\text{LIN}}) = \frac{1}{N}\|\mathbf{y}-\hat{\mathbf{y}}\|^2 = \frac{1}{N}\|(I-H)\mathbf{y}\|^2, \quad \text{其中} H = XX^\dagger.$$

几何解释（$\mathbf{H}$ 矩阵）

- H 把任意 $\mathbf{y} \in \mathbb{R}^N$ 投影到 $\text{span}(X)$，故 $\hat{\mathbf{y}} = H\mathbf{y}$。
- $I-H$ 把 $\mathbf{y}$ 映射到其残差 $\mathbf{y}-\hat{\mathbf{y}} \perp \text{span}(X)$。
- 自由度：$\text{Tr}(I-H) = N-(d+1)$。（由于 $\text{Tr}(H) = d+1$，利用迹的可加性即得）

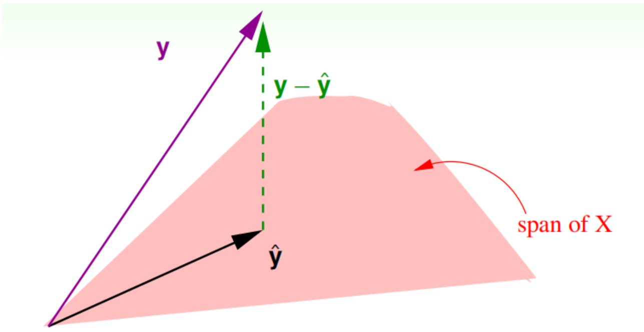


图 9.3.1: 线性回归中观测向量在列空间的投影与残差示意图

命题 9.3.2 (线性回归学习曲线)

设输入数据矩阵 $\mathbf{X} \in \mathbb{R}^{N\times(d+1)}$，其列空间为 $\mathcal{S} = \text{span}(\mathbf{X}) \subseteq \mathbb{R}^N$，观测目标为

$$\mathbf{y} = \mathbf{f} + \varepsilon, \quad \text{其中} \mathbf{f} \in \mathcal{S},\ \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_N).$$

记投影矩阵为 $\mathbf{H} = \mathbf{X}\mathbf{X}^\dagger$，则模型预测 $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ 是 $\mathbf{y}$ 在 $\mathcal{S}$ 上的正交投影，残差为

$$\mathbf{y}-\hat{\mathbf{y}} = (\mathbf{I}-\mathbf{H})\varepsilon.$$

于是有：

- 样本内误差的期望为
$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(\mathbf{w}_{\text{LIN}})] = \frac{1}{N}\mathbb{E}\left[\|(\mathbf{I}-\mathbf{H})\varepsilon\|^2\right] = \sigma^2\left(1-\frac{d+1}{N}\right);$$
- 样本外误差的期望为（推导更复杂，但形式对称）
$$\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(\mathbf{w}_{\text{LIN}})] = \sigma^2\left(1+\frac{d+1}{N}\right).$$

证明 由于 $\mathbf{f} \in \text{span}(\mathbf{X})$，我们有 $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \mathbf{H}(\mathbf{f}+\varepsilon) = \mathbf{f}+\mathbf{H}\varepsilon$。因此残差为

$$\mathbf{y}-\hat{\mathbf{y}} = (\mathbf{f}+\varepsilon)-(\mathbf{f}+\mathbf{H}\varepsilon) = (\mathbf{I}-\mathbf{H})\varepsilon.$$

因此，样本内误差为

$$E_{\text{in}} = \frac{1}{N}\|\mathbf{y}-\hat{\mathbf{y}}\|^2 = \frac{1}{N}\|(\mathbf{I}-\mathbf{H})\varepsilon\|^2.$$

取期望得：

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}] = \frac{1}{N}\mathbb{E}\left[\varepsilon^\top(\mathbf{I}-\mathbf{H})^2\varepsilon\right].$$

由于 $\mathbf{H}$ 是正交投影矩阵，满足 $\mathbf{H}=\mathbf{H}^\top=\mathbf{H}^2$，因此 $(\mathbf{I}-\mathbf{H})^2=\mathbf{I}-\mathbf{H}$。于是

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}]=\frac{1}{N}\mathbb{E}\left[\boldsymbol{\varepsilon}^\top(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}\right]=\frac{1}{N}\operatorname{Tr}\left((\mathbf{I}-\mathbf{H})\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]\right).$$

由于 $\boldsymbol{\varepsilon}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$，协方差为 $\sigma^2\mathbf{I}$，因此

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}]=\frac{1}{N}\operatorname{Tr}\left((\mathbf{I}-\mathbf{H})\sigma^2\mathbf{I}\right)=\frac{\sigma^2}{N}\operatorname{Tr}(\mathbf{I}-\mathbf{H})=\frac{\sigma^2}{N}(N-\operatorname{Tr}(\mathbf{H})).$$

注意 $\operatorname{Tr}(\mathbf{H})=\operatorname{rank}(\mathbf{H})=d+1$，因此

$$\mathbb{E}_{\mathcal{D}}[E_{\text{in}}]=\sigma^2\left(1-\frac{d+1}{N}\right).$$

对于样本外误差，设新样本点为 $(\mathbf{x}_0,y_0)$，其中

$$y_0=\mathbf{x}_0^\top\mathbf{w}_f+\varepsilon_0,\quad \varepsilon_0\sim\mathcal{N}(0,\sigma^2),\quad \text{且}\,\mathbf{f}=\mathbf{X}\mathbf{w}_f.$$

最小二乘解为

$$\mathbf{w}_{\text{LIN}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{f}+\boldsymbol{\varepsilon}),$$

预测为

$$\hat{y}_0=\mathbf{x}_0^\top\mathbf{w}_{\text{LIN}}=\mathbf{x}_0^\top\mathbf{w}_f+\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}.$$

因此误差为

$$y_0-\hat{y}_0=\varepsilon_0-\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}.$$

于是有

$$\mathbb{E}[(y_0-\hat{y}_0)^2]=\mathbb{E}[\varepsilon_0^2]+\mathbb{E}\left[\left(\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\right)^2\right].$$

第一项为 σ^2。记 $A=\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$，则第二项为

$$\mathbb{E}[(A\boldsymbol{\varepsilon})^2]=\mathbb{E}[\boldsymbol{\varepsilon}^\top A^\top A\boldsymbol{\varepsilon}]=\sigma^2\operatorname{Tr}(A^\top A).$$

在均匀采样假设下，有 $\mathbb{E}[\operatorname{Tr}(A^\top A)]=\frac{d+1}{N}$，因此

$$\mathbb{E}_{\mathcal{D}}[E_{\text{out}}]=\sigma^2+\sigma^2\cdot\frac{d+1}{N}=\sigma^2\left(1+\frac{d+1}{N}\right).$$

■

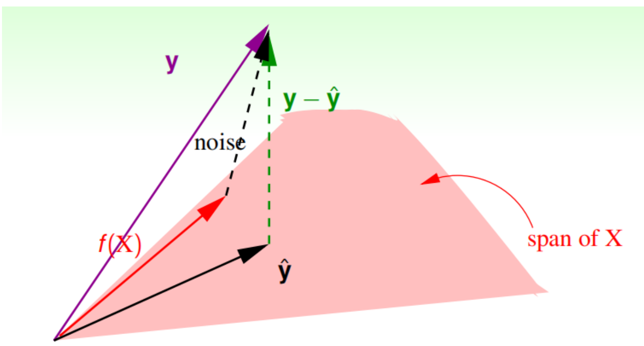


图 9.3.2: 线性回归中观测向量的信号与噪声分解示意图

定理 9.3.1 (线性回归的学习曲线)

设噪声水平为 σ^2，特征维度为 d，训练样本数为 N，则线性回归的样本内与样本外误差的期望

为：

$$\overline{E}_{\text{out}}(\mathbf{w}_{\text{LIN}}) = \sigma^2\left(1+\frac{d+1}{N}\right),$$

$$\overline{E}_{\text{in}}(\mathbf{w}_{\text{LIN}}) = \sigma^2\left(1-\frac{d+1}{N}\right).$$

推论：

- 当 $N\to\infty$ 时，$\overline{E}_{\text{in}},\overline{E}_{\text{out}}\to\sigma^2$，误差收敛于噪声水平；
- 期望泛化误差为：

$$\overline{E}_{\text{out}}-\overline{E}_{\text{in}} = \frac{2(d+1)}{N}\cdot\sigma^2;$$

- 泛化误差与 VC 理论中的最坏情形上界同阶（d+1），说明“学习确实发生了”。

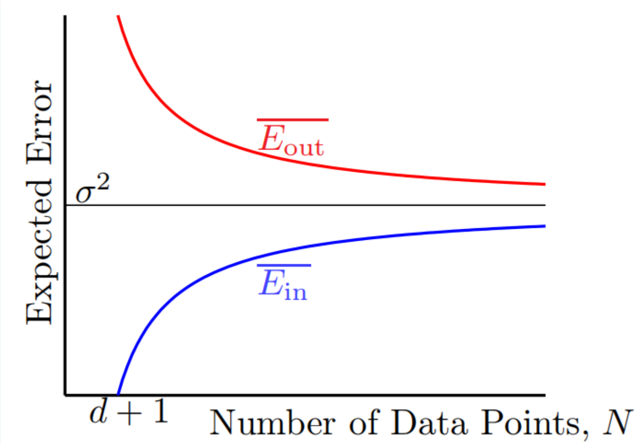


图 9.3.3: 数据点数量与期望误差关系示意图

9.4 线性回归用于二分类

命题 9.4.1 (线性分类 vs. 线性回归)

- 线性分类

  输出空间 $\{-1,+1\}$，假设 $h(x)=\text{sign}(w^\top x)$，代价函数

$$\text{err}(\hat{y},y)=\mathbb{I}[\hat{y}\neq y],$$

  求解一般情形为 NP-hard。

- 线性回归

  输出空间 $\mathbb{R}$，假设 $h(x)=w^\top x$，代价函数

$$\text{err}(\hat{y},y)=(\hat{y}-y)^2,$$

  存在解析解

$$w_{\text{LIN}}=(X^\top X)^{-1}X^\top y.$$

- 用线性回归做二分类的启发式

对二分类数据 $\mathcal{D}$ 直接运行线性回归（高效），然后取

$$g(x) = \mathrm{sign}(w_{\mathrm{LIN}}^{\top} x).$$

该启发式在实践中表现良好，其理论解释见后文。

命题 9.4.2 (两种误差的关系与线性回归二分类近似)

设二分类标签 $y \in \{-1, +1\}$，令

- 0/1 误差：$\mathrm{err}_{0/1}(w) = \mathbb{I}[\mathrm{sign}(w^{\top} x) \neq y]$；
- 平方误差：$\mathrm{err}_{\mathrm{sqr}}(w) = (w^{\top} x - y)^2$。

逐点误差上界对任意 (x, y) 有

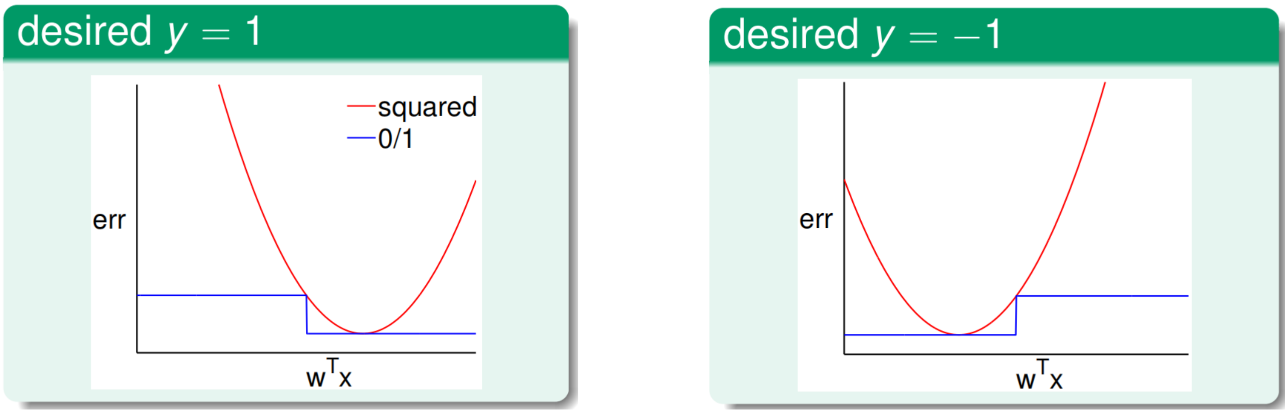
$$\mathrm{err}_{0/1}(w) \leq \mathrm{err}_{\mathrm{sqr}}(w).$$

VC 泛化界

$$E_{\mathrm{out},0/1}(w) \leq E_{\mathrm{in},0/1}(w) + \Omega\left(\frac{d+1}{N}\right) \leq E_{\mathrm{in,sqr}}(w) + \Omega\left(\frac{d+1}{N}\right),$$

其中右侧为宽松上界，但换来计算效率。

结论　w_{LIN} 既可作为高效的基线分类器，也可作为 PLA / Pocket 算法的初始化向量。



(a) 期望输出为 1 时平方误差与 0-1 误差对比图　(b) 期望输出为 - 1 时平方误差与 0-1 误差对比图

图 9.4.1: 期望输出的平方误差与 0-1 误差对比图

9.5 总结

笔记 [线性回归]

- 线性回归问题：用超平面逼近实值输出。
- 线性回归算法：利用伪逆得到解析解。
- 泛化问题：平均而言 $E_{\mathrm{out}} - E_{\mathrm{in}} \approx \dfrac{2(d+1)}{N}$。
- 线性回归用于二分类：0/1 误差上界由平方误差控制。