# HW1

PB23000209 赵文凯

## 1 模型偏差（Model Bias）

### (a)

设数据满足：

$$y = f_{\text{true}}(x) + \varepsilon$$

其中

$$\mathbb{E}[\varepsilon] = 0,\ \mathrm{Var}(\varepsilon) = \sigma^2$$

我们从训练集随机抽子集并各自训练，得到一族模型

$$\hat{f}^{(1)}, \ldots, \hat{f}^{(n)}$$

其均值为

$$\bar{f}(x) = \mathbb{E}_{\text{train}}[\hat{f}(x)]$$

泛化 MSE：

$$\mathbb{E}_{\text{train}}\left[\mathbb{E}_{x,y}(y - \hat{f}(x))^2\right] = \mathbb{E}_{\text{train}}\left[\mathbb{E}_x\left(\mathbb{E}_{y|x}(y - \hat{f}(x))^2\right)\right]$$

代入

$$y = f_{\text{true}}(x) + \varepsilon, \mathbb{E}[\varepsilon] = 0,\ \mathbb{E}[\varepsilon(f_{\text{true}} - \hat{f})] = 0$$

得

$$\mathbb{E}_{y|x}(y - \hat{f}(x))^2 = \mathbb{E}(\varepsilon + f_{\text{true}}(x) - \hat{f}(x))^2 = \sigma^2 + (f_{\text{true}}(x) - \hat{f}(x))^2,$$

再对训练采样取期望

$$\mathbb{E}_{\text{train}}\left[(f_{\text{true}} - \hat{f})^2\right] = \mathbb{E}_{\text{train}}\left[(f_{\text{true}} - \bar{f} + \bar{f} - \hat{f})^2\right] = (f_{\text{true}} - \bar{f})^2 + \mathbb{E}_{\text{train}}\left[(\bar{f} - \hat{f})^2\right],$$

于是

$$\mathbb{E}_{\text{train}}[\text{GE}] = \sigma^2 + \underbrace{(f_{\text{true}} - \bar{f})^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}_{\text{train}}[(\bar{f} - \hat{f})^2]}_{\mathrm{Var}(f)}.$$

从而得到：

$$\mathbb{E}_{\text{train}}[\text{GE}] = \sigma^2 + \text{Bias}^2 + \mathrm{Var}(f)$$

### (b)

- 非负性：

$$\text{Bias}^2 = \mathbb{E}_{\text{train}}[(\bar{f} - f_{\text{true}})^2] \geq 0$$

- 何时恒大于 0：模型容量不足以表达真函数时（欠拟合）

- 例如：真实关系 (y=x^2)，而模型类仅包含所有一元线性函数

### (c)

- Var 主要与 N 相关。通常 $\mathrm{Var} = \mathcal{O}(1/n)$；模型越复杂，方差越大。
- Bias^2 主要与模型复杂度相关。复杂度越高，Bias 通常越小。

---

## 2 数据偏差（Data Bias）

### (a)

令

$$Z_i = I(h(x_i) \neq y_i)$$

则

$$\mathrm{err}_S(h) = \frac{1}{n} \sum_{i=1}^{n} Z_i, \quad \mathbb{E}[Z_i] = \mathrm{err}_D(h).$$

由 Hoeffding 不等式（a=0,b=1）：

$$\Pr\left[\left|\frac{1}{n}\sum Z_i - \mathbb{E}\right| > \varepsilon\right] \leq 2\exp(-2n\varepsilon^2),$$

即

$$\Pr\left(|\mathrm{err}_D(h) - \mathrm{err}_S(h)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

### (b)

$$\Pr\left[\exists h \in H,\ |\mathrm{err}_D(h) - \mathrm{err}_S(h)| > \varepsilon\right] \leq \sum_{h \in H} 2e^{-2n\varepsilon^2} = 2|H|e^{-2n\varepsilon^2}.$$

### (c)

- 直观：$|\mathrm{err}_D - \mathrm{err}_S|$ 随样本数 n 指数级缩小；随假设类规模 |H| 线性增大。
- 影响因素：样本数 (n)（越大越好）、模型类复杂度/容量（越大越易过拟合）、数据噪声。
- 缓解：增大数据量/数据增强、正则化/限制容量、改进采样使训练分布更接近 D。

---

## 3 贝叶斯分类器

### (a)

取对数得

$$\log L = \sum_{i=1}^{n_1} \log p(x_i^{(1)} | \mu_1, \Sigma) + \sum_{j=1}^{n_2} \log p(x_j^{(2)} | \mu_2, \Sigma)$$

对 $\mu_k$ 求偏导并置零得

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i^{(k)}$$

对 $\Sigma$ 求导并置零，并代入 $\mu_k$ 得

$$\hat{\Sigma} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(x_i^{(1)}-\hat{\mu}_1)(x_i^{(1)}-\hat{\mu}_1)^\top + \sum_{j=1}^{n_2}(x_j^{(2)}-\hat{\mu}_2)(x_j^{(2)}-\hat{\mu}_2)^\top\right).$$

## (b)

标签 0 视为类别1，标签1视为类别2

- 期望：

$$\hat{\mu}_1 = (2.50,\ 2.72,\ 2.28),$$
$$\hat{\mu}_2 = (3.50,\ 3.76,\ 3.52).$$

- 协方差：

$$\hat{\Sigma} = \begin{bmatrix} 0.136 & -0.032 & 0.064 \\ -0.032 & 0.032 & -0.0114 \\ 0.064 & -0.0114 & 0.0476 \end{bmatrix}.$$

- 对新数据 x = (2.7,2.9,3.5)：

$$p(y_1|x) = \frac{n_1\mathcal{N}(x;\hat{\mu}_1,\hat{\Sigma})}{\sum_k n_k\mathcal{N}(x;\hat{\mu}_k,\hat{\Sigma})}$$

  计算得

$$p(y_1|x_{\text{new}}) \approx 0.0214, \quad p(y_2|x_{\text{new}}) \approx 0.9786,$$

  判为"标签 1"。

## (c)

> 原题目中z应该为log $\frac{p(x|y_2)p(y_2)}{p(x|y_1)p(y_1)}$，似乎少写了log

$$z = \log\frac{p(x|y_2)p(y_2)}{p(x|y_1)p(y_1)} = \underbrace{\Sigma^{-1}(\mu_2-\mu_1)}_{w}\cdot x + \underbrace{\left(-\frac{1}{2}(\mu_2^\top\Sigma^{-1}\mu_2 - \mu_1^\top\Sigma^{-1}\mu_1) + \ln\frac{n_2}{n_1}\right)}_{b}.$$

---

# 4 神经网络

## (a)

隐藏层：

$$z^1 = x^\top W^1 + b^1 = (2, 13, -4, -15)$$
$$a^1 = \mathrm{ReLU}(z^1) = (2, 13, 0, 0)$$

输出层：

$$z^2 = a^1W^2 + b^2 = (15, -13)$$

softmax 概率：

$$a^{(2)} = \mathrm{softmax}(15, -13) = (1,\ 6.91\times10^{-13})$$

## (b)

$$x^{(1)} = (0.5, 0.5)^\top,\ x^{(2)} = (0, 2)^\top,\ x^{(3)} = (-3, 0.5)^\top$$

计算得

$$\begin{aligned} a^{1(1)} &= (0,0,0,0) \\ a^{1(2)} &= (0,1,0,0) \\ a^{1(3)} &= (0,0,2,0) \end{aligned}$$

隐藏层输出矩阵为

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

## (c)

已知：输入 x=(3,14)，目标 y=(0,1)。

前向得到：$z^{(1)} = (2, 13, -4, -15) \Rightarrow a^{(1)} = (2, 13, 0, 0)$;

$z^{(2)} = (15, -13) \Rightarrow a^{(2)} = \text{softmax}(z^{(2)}) \approx (1,\ 6.91 \times 10^{-13})$。

输出层残差

$$\delta^{(2)} = a^{(2)} - y = (1, -1).$$

二层梯度

$$\frac{\partial L}{\partial W^{(2)}} = a^{(1)} \otimes \delta^{(2)} = \begin{bmatrix} 2 & -2 \\ 13 & -13 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \frac{\partial L}{\partial b^{(2)}} = (1, -1).$$

更新后

$$W^{(2)} \leftarrow \begin{bmatrix} 0.8 & -0.8 \\ -0.3 & 0.3 \\ 1 & -1 \\ 1 & -1 \end{bmatrix}, \quad b^{(2)} \leftarrow (-0.1, , 2.1)$$

回传到隐藏层（用更新前的 (W^{(2)}))

$$\delta^{(1)} = \left(W^{(2)} \delta^{(2)}\right) \odot \mathbf{1}_{z^{(1)}>0} = (2, 2, 0, 0).$$

一层梯度与更新

$$\frac{\partial L}{\partial W^{(1)}} = x \otimes \delta^{(1)} = \begin{bmatrix} 6 & 6 & 0 & 0 \\ 28 & 28 & 0 & 0 \end{bmatrix}, \qquad \frac{\partial L}{\partial b^{(1)}} = (2, 2, 0, 0).$$

$$W^{(1)} \leftarrow \begin{bmatrix} 0.4 & -0.6 & -1 & 0 \\ -2.8 & -1.8 & 0 & -1 \end{bmatrix}, \quad b^{(1)} \leftarrow (-1.2, , -1.2, , -1, , -1)$$