Learning Theory

bias / variance tradeoff

设x表示预测样本，y_D为x在数据集中的标记，y是x的真实标记，$f(x;D)$为由数据集D学得的模型f对x的预测输出，$\bar{f}(x)$为模型对不同数据集预测值的期望，则有：

1° generalization error.

$$Err(x) = E[(y - f(x))^2]$$

2° ~~bias~~. variance

$$\begin{cases} \bar{f}(x) = E_D[f(x;D)] \\ Var(x) = E_D[(f(x;D) - \bar{f}(x))^2] \end{cases}$$

对不同数据集预测值的期望

3° noise

$$\varepsilon^2 = E_D[(y_D - y)^2]$$

4° bias.

$$bias^2(x) = (\bar{f}(x) - y)^2$$

$$E[f;D] = E_D[(f(x;D) - y_D)^2]$$

$$= E_D[(f(x;D) - \bar{f}(x) + \bar{f}(x) - y_D)^2]$$

$$= E_D[(f(x;D) - \bar{f}(x))^2] + E_D[(\bar{f}(x) - y_D)^2] + \underbrace{2E_D[(f(x;D) - \bar{f}(x))(\bar{f}(x) - y_D)]}_{=0}$$

$$= E_D[(f(x;D) - \bar{f}(x))^2] + E_D[(\bar{f}(x) - y)^2] + E_D[(y - y_D)^2] + \underbrace{2E_D[(\bar{f}(x) - y)(y - y_D)]}_{=0}$$

$$= \underbrace{E_D[(f(x;D) - \bar{f}(x))^2]}_{\text{variance}} + \underbrace{(\bar{f}(x) - y)^2}_{\text{bias}^2} + \underbrace{E_D[(y - y_D)^2]}_{\text{noise}}$$

Variance → 算法受数据扰动的影响.

bias^2 → 算法的拟合能力

noise → 学习问题本身的难度.

learning theory

union bound $\quad p(A_1 \cup A_2 \cdots \cup A_k) \le p(A_1) + \cdots + p(A_k)$

hoeffding inequality $\quad$ 假设 $z_1, \ldots, z_n$ iid. $\sim$ Bernoulli (ϕ), $\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} z_i$

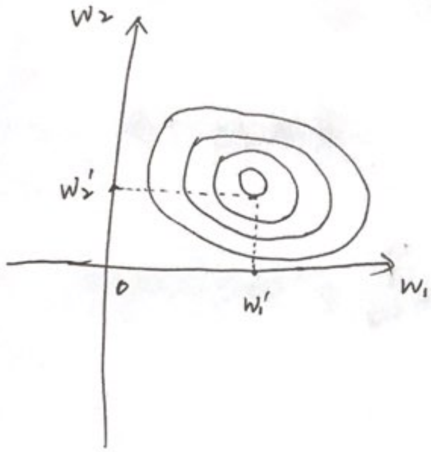
$$p(|\phi - \hat{\phi}| > \varepsilon) \le 2e^{-2\varepsilon^2 n}.$$

Regularization

目的：缓解过拟合问题，限制参数过大或过多。

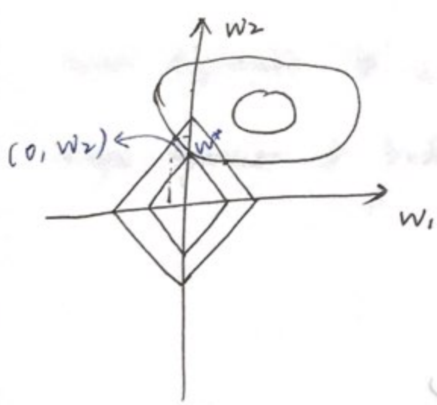
L1正则化： $E = E_{in} + \lambda \sum_j |w_j|$， $\sum_j |w_j| \le c$.

L2正则化： $E = E_{in} + \lambda \sum_j \|w_j\|^2$， $\sum_j \|w_j\|^2 \le c$.



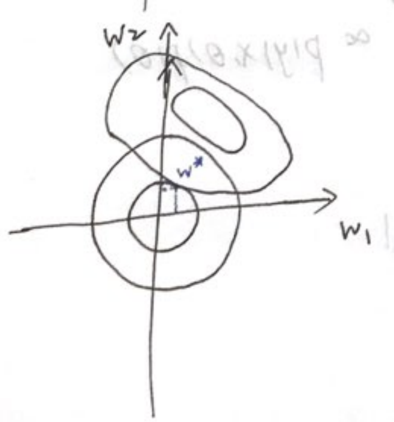
1) 误差函数等高线

此时我们希望找到 (w_1, w_2) 使得误差最小，即最里面的等高线。但此时得到的 (w_1, w_2) 比较大。



2) L1正则化

此时我们希望最小化误差的同时，C也越小。对于同一条等高线来说，当 $\sum_j |w_j| \le c$ 与其只有一个交点，即相切时 (w_1, w_2) 最小。且此时很容易落在坐标轴上，从而使得某些维度坐标为0，从而得到稀疏解。



3) L2正则化

同理，对于同一条等高线来说，当 $\sum_j w_j^2 \le c$ 与其相切时，C最小。且此时 w^* 的坐标靠近原点，不会落在坐标上，所以得到的 w^* 较小，且不为0。

从贝叶斯角度（先验/概率）的角度看 L_1 和 L_2 正则化

1) L_1 正则化 ~ Laplace 先验

Laplace 分布:

$$f(x|\mu,b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

b 越小，数据越集中在 μ 附近

假设参数 $\theta \sim \text{Laplace}(0, \frac{1}{\lambda})$，则:

$$p(\theta_j) = \frac{\lambda}{2} e^{-\lambda|\theta_j|}$$

$$\therefore p(\theta) = \prod_{j=1}^{m} p(\theta_j) = \prod_{j=1}^{m} \frac{\lambda}{2} e^{-\lambda|\theta_j|}$$

$$\therefore \log p(\theta) = \sum_{j=1}^{m} \log\frac{\lambda}{2} - \lambda \sum_{j=1}^{m} \cancel{\log} |\theta_j|$$

最大后验概率:

$$p(\theta|x,y) = \frac{p(x,y,\theta)}{p(x,y)} = \frac{p(y|x,\theta)\,p(\theta)}{p(x,y)} \propto p(y|x,\theta)\,p(\theta)$$

$$\therefore -\log p(\theta|x,y) \approx \sum_{j=1}^{N} \|y_i - h_\theta(x_i)\|^2 + \lambda \sum_{j=1}^{m} |\theta_j|$$

2) l_2正则化 ~ 高斯先验

假设 θ ~ 高斯分布，则：

$$p(\theta_j) = \frac{\lambda}{\sqrt{\pi}} e^{-\lambda\|\theta_j\|^2}$$

$$\therefore \log p(\theta) = \log \prod_{j=1}^{m} p(\theta_j) = \sum_{j=1}^{m} \log \frac{\lambda}{\sqrt{\pi}} - \lambda \sum_{j=1}^{m} \|\theta_j\|^2.$$

$\therefore$ 最大后验概率：

$$-\log p(\theta | x, y) \approx \sum_{i=1}^{n} \|y_i - h_\theta(x_i)\|^2 + \lambda \sum_{j=1}^{m} \|\theta_j\|^2.$$

$$\begin{cases} \text{Lasso Regression} \Rightarrow l_1\text{正则化的线性回归.} \\ \text{Ridge Regression} \Rightarrow l_2\text{正则化的线性回归.} \end{cases}$$