l_1, l_2 正则化

优化目标：

在原目标(代价)函数中添加惩罚项，

$\tilde{J}(w;X,y) = J(w;X,y) + \alpha\Omega(w)$

1. 基于约束条件的最优化

模型复杂度用VC维衡量。VC维与系数w个数成线性关系，VC维越大，模型越复杂。

用l_0范数来限制个数

$$\min_w J(w;X,y)$$

$$s.t.\ \|w\|_0 \le C$$

向量中非0的个数

但该问题NP问题，无法求解

因此要求w接近于0，尽量小，所以用l_1,l_2范数近似l_0范数。

即：$\min J(w;X,y)$ 或 $\min J(w;X,y)$

$s.t\ \|w\|_1 \le C$ $s.t\ \|w\|_2 \le C$

拉格朗日乘子法：

$L(w,\alpha) = J(w;X,y) + \alpha(\|w\|_1 - C)$ $\alpha > 0$

$L(w,\alpha) = J(w;X,y) + \alpha(\|w\|_2 - C)$

假设α最优解是α^*，则最优等价于

$\min_w J(w;X,y) + \alpha^*\|w\|_1$ 或

$\min J(w;X,y) + \alpha^*\|w\|_2$

这与 $\min \tilde{J}(w;X,y)$ 等价

所以是在原$J(w;X,y)$增加约束条件

$$\begin{cases} \|w\|_1 \le C \\ \|w\|_2 \le C \end{cases}$$

2. 基于最大后验概率估计

将w看作随机变量，也具有某种分布，从而

$$P(w|X,y) = \frac{P(y|X,w)P(w)}{P(y)} \propto P(y|X,w)P(w)$$

$$MAP = \underbrace{\log P(y|X,w)}_{\text{似然函数}} + \underbrace{\log P(w)}_{\text{增加对w的相关分布的先验假设}}$$

收集到$\{X,y\}$后，根据w在$\{X,y\}$下的后验概率对w进行修正。

① 假设w_j的先验分布是均值为0的高斯分布

$$w_j \sim N(0,\sigma^2)$$

$$\log P(w) = \log \prod_j P(w_j) = \log \prod_j \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(w_j)^2}{2\sigma^2}}\right]$$

$$= -\frac{1}{2\sigma^2}\sum_j w_j^2 + C'$$

等价于l_2范则项

② 假设均值为0，参数为a；拉普拉斯分布

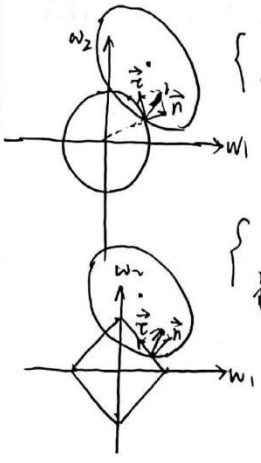
$$P(w_j) = \frac{1}{2a} e^{\frac{-|w_j|}{a}}$$

则 $\log P(w_j) = \log \prod_j \frac{1}{2a} e^{\frac{-|w_j|}{a}}$

$$= -\frac{1}{a}\sum_j |w_j| + C'$$

等价于增加l_1范则项

3. 直观上



{ 又被沿切线移动，最终和角同与圆相切。

{ 沿切线移动最终到顶点处

千姓论上

设原目标函数 $J(w)$ 的最优解为 w^*，假设 $J(w)$ 在 w^* 二阶泰勒展开有

$$\hat{J}(w) = J(w^*) + \frac{1}{2}(w-w^*)^T H (w-w^*)$$

H：Hessian 矩阵

w^* 为 $J(w)$ 最优解，$\therefore J(w^*)=0$

$$\therefore \hat{J}(w) = \frac{1}{2}(w-w^*)^T H (w-w^*)$$

$\therefore \hat{J}(w)$ 取得最小值为

$$\nabla_w \hat{J}(w) = H(w-w^*) = 0$$

添加 L2 正则项后

$$\nabla_w \tilde{J}(w) = \nabla_w \hat{J}(w) + \nabla_w \Omega(w)$$
$$= H(w-w^*) + \alpha w$$

设最优解为 $\tilde{w}$，则有

$$H(\tilde{w}-w^*) + \alpha\tilde{w} = 0$$

$$\therefore \tilde{w} = (H+\alpha I)^{-1} H w^*$$

H 是对称阵，可特征分解 $H = Q\Lambda Q^T$ Q 为特征

$$\tilde{w} = Q(\underbrace{\Lambda+\alpha I}_{\lambda_j})^{-1}\Lambda Q^T w^*$$

$\frac{\lambda_j}{\lambda_j+\alpha}$

$\therefore \tilde{w}$ 是 w^* 在 H 的每个特征向量上分量以 $\frac{\lambda_j}{\lambda_j+\alpha}$ 比例收缩得到。

① 若 $\lambda_j \gg \alpha$，则 w_j^* 受正则化影响较小

② 若 $\lambda_j \ll \alpha$，则 w_j^* 受正则化影响较大，收缩到接近于0的值，但 $w_j^* \neq 0$，$\therefore \tilde{w}_j \neq 0$，因而不会稀疏。