l1相对于l2更容易获得稀疏解的个人看法

杨航锋

从优化(最小化)损失函数的角度来看，稀疏解 $\theta(\theta \in \mathbb{R}^n)$ 产生的条件是:

1、如果损失函数 $J(\theta)$ 在 θ 处可导，且 θ 满足 $\left.\frac{\partial J(\theta)}{\partial \theta}\right|_{\theta=0} = 0$；

2、如果损失函数 $J(\theta)$ 在 θ 处不可导，且 $\left.\frac{\partial J(\theta)}{\partial \theta}\right|_{\theta=0^+} > 0$、$\left.\frac{\partial J(\theta)}{\partial \theta}\right|_{\theta=0^-} < 0$。

假设未加入正则化之前的损失函数为 $l(\theta)$，且 $\left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=0} = \delta \neq 0$，则有

$$
\begin{aligned}
J_{l1}(\theta) &= l(\theta) + \lambda \|\theta\|_1 \\
J_{l2}(\theta) &= l(\theta) + \lambda \|\theta\|_2^2
\end{aligned}
$$

分别计算 $J_{l1}(\theta)$、$J_{l2}(\theta)$ 在 $\theta = 0$ 处的导函数情况

$$
\begin{aligned}
&\left.\frac{\partial J_{l1}(\theta)}{\partial \theta}\right|_{\theta=0} = l'(\theta) + \lambda sign(\theta) \\
&\begin{cases}
\left.\frac{\partial J_{l1}(\theta)}{\partial \theta}\right|_{\theta=0^+} = \delta + \lambda \\
\left.\frac{\partial J_{l1}(\theta)}{\partial \theta}\right|_{\theta=0^-} = \delta - \lambda
\end{cases} \\
&\left.\frac{\partial J_{l2}(\theta)}{\partial \theta}\right|_{\theta=0} = l'(\theta) + 2\lambda\theta = \delta
\end{aligned}
$$

因此在 θ 的各个分量中，当 δ 为一个不为 0 的常量时，$\delta + \lambda$、$\delta - \lambda$ 产生异号的可能性更大（导数值异号），$J_{l1}(\theta)$ 在该点取得极小值；而 $J_{l2}(\theta)$ 该点的导数值为常量故取不到极小值，所以 $l1$ 正则化相对要比 $l2$ 正则化更容易产生稀疏解。