$L1$ 和 $L2$ 正则化的概率解释

杨航锋

正则化在机器学习中主要用于控制模型的复杂度、解决过拟合和追求更优预测效果的重要手段，而常见的正则化有 $L1$ 正则化和 $L2$ 正则化。 $L1$ 正则化可以产生稀疏权值矩阵，即产生一个稀疏模型，用于特征筛选； $L2$ 正则化可以防止过拟合，提升模型的泛化能力。

$L1$ 正则化和 $L2$ 正则化的符号化描述

假设待优化函数为 $f(\theta)$ ，其中 $\theta \in \mathbb{R}^n$ ，那么优化问题可以转化为求

$$\arg\min_{\theta} \ f(\theta)$$

- $L1$ 正则化，即对参数 θ 加上 $L1$ 范数约束

$$\arg\min_{\theta} \ J_1(\theta) = f(\theta) + \lambda\|\theta\|_1$$

- $L2$ 正则化，即对参数 θ 加上 $L2$ 范数的平方约束

$$\arg\min_{\theta} \ J_2(\theta) = f(\theta) + \lambda\|\theta\|_2^2$$

从贝叶斯先验概率看正则化

假设输入空间是 $X \in \mathbb{R}^n$,输出空间是 Y ，不妨假设含有m个样本数据($x^{(1)}$,$y^{(1)}$)、($x^{(2)}$,$y^{(2)}$)、$\cdots$、($x^{(m)}$,$y^{(m)}$)，其中$x^{(i)} \in X$、$y^{(i)} \in Y$ 。

贝叶斯学派认为参数 θ 也是服从某种概率分布的，即先给定 θ 的先验分布为 $p(\theta)$ ，然后根据贝叶斯定理 $P(\theta|(X,Y)) = \dfrac{P((Y,X);\theta) \times P(\theta)}{P(X,Y)} \sim P(Y|X;\theta) \times P(\theta)$ （这里的 $Y|X$ 仅仅是一种记号，代表给定的 X 对应相关的 Y ），因此通过极大似然估计可求参数 θ 。

$$\arg\max_{\theta} \ L(\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)p(\theta)$$

等价于求解对数化极大似然函数 $l(\theta)$

$$
\begin{aligned}
\arg\max_{\theta}\ l(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} \log p(y^{(i)} | x^{(i)}; \theta) + \sum_{i=1}^{m} \log p(\theta) \\
\Leftrightarrow \arg\min_{\theta}\ -l(\theta) &= -\log L(\theta) \\
&= -\sum_{i=1}^{m} \log p(y^{(i)} | x^{(i)}; \theta) - \sum_{i=1}^{m} \log p(\theta) \\
&= f(\theta) - \sum_{i=1}^{m} \log p(\theta)
\end{aligned}
$$

- $L1$ 正则化的概率解释

  假设 θ 服从的先验分布为均值为 0 参数为 λ 的拉普拉斯分布，即 $\theta \sim La(0, \lambda)$ 其中，$p(\theta) = \frac{1}{2\lambda} e^{-\frac{|\theta|}{\lambda}}$。因此，上述优化函数可转换为：

$$
\begin{aligned}
&\arg\min_{\theta}\ f(\theta) - \sum_{i=1}^{m} \log p(\theta) \\
&= f(\theta) - \sum_{i=1}^{m} \log \frac{1}{2\lambda} e^{-\frac{|\theta_i|}{\lambda}} \\
&= f(\theta) - \sum_{i=1}^{m} \log \frac{1}{2\lambda} + \frac{1}{\lambda} \sum_{i=1}^{m} |\theta_i| \\
&\Leftrightarrow \arg\min_{\theta}\ f(\theta) + \lambda \|\theta\|_1
\end{aligned}
$$

  从上面的数学推导可以看出， $L1$ 正则化可以看成是：通过假设权重参数 θ 的先验分布为拉普拉斯分布，由最大后验概率估计导出。

- $L2$ 正则化的概率解释

  假设 θ 服从的先验分布为均值为 0 方差为 σ^2 的正态分布，即 $\theta \sim \mathcal{N}(0, \sigma^2)$ 其中，$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\theta^2}{2\sigma^2}}$ 。因此，上述优化函数可转换为：

$$
\begin{aligned}
&\arg\min_{\theta}\ f(\theta) - \sum_{i=1}^{m} \log p(\theta) \\
&= f(\theta) - \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-\theta_i^2}{2\sigma^2}} \\
&= f(\theta) - \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{2\sigma^2} \sum_{i=1}^{m} \theta_i^2 \\
&\Leftrightarrow \arg\min_{\theta} f(\theta) + \lambda \|\theta\|_2^2
\end{aligned}
$$

从上面的数学推导可以看出，$L2$ 正则化可以看成是：通过假设权重参数 θ 的先验分布为正态分布，由最大后验概率估计导出。