CS229_4

CS229

最小二乘法的概率解释

为什么在线性回归问题中我们选择最小二乘法定义代价函数$J(\theta)$？本小节将就这一问题进行讨论。

首先，我们假设对于每一个样本实例$(x^{(i)}, y^{(i)})$，特征变量x和目标值y的关系如下：

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

其中，$\epsilon^{(i)}$表示误差。

让我们进一步假设误差$\epsilon^{(i)}$服从正态分布（也称为高斯分布），即$\epsilon^{(i)} \sim N(0, \sigma^2)$。因此，误差$\epsilon$为独立同分布（Independent and Identical Distribution，IID）。

$$P(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(\epsilon^{(i)})^2}{2\sigma^2})$$

当给定参数θ和x时，目标值y也服从正态分布，即$y^{(i)}|x^{(i)}; \theta \sim N(\theta^T x^{(i)}, \sigma^2)$。

$$P(y^{(i)}|x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2})$$

注：$x^{(i)}$与θ之间为分号，表示θ为已知变量。

又因为似然函数（Likelihood Function）如下：

$$L(\theta) = L(\theta; X, Y) = P(Y|X; \theta)$$

其中，Y表示一个长度为训练集大小的向量，X表示维度为训练集数*特征变量数的矩阵。

将上述结论带入似然函数可得：

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}) \end{aligned}$$

为了计算出参数θ，我们采用极大似然估计。为了便于计算，我们可将上式转变为最大化对数似然。

$$\begin{aligned} \ell(\theta) &= logL(\theta) \\ &= log\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^{(i)} - \theta^T x)^2}{2\sigma^2}) \\ &= \sum_{i=1}^{m} log\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y^{(i)} - \theta^T x)^2}{2\sigma^2}) \\ &= m\ log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2 \end{aligned}$$

因此，我们不难发现最大化对数似然，实际上在最小化$\frac{1}{2}\sum_{i=1}^{m}(y^{(i)} - \theta^T x^{(i)})^2$。