机器学习笔记二——为什么选择最小二乘

本笔记是在斯坦福大学的机器学习公开课上学习总结的，纯属个人意见。

在进行求解方程 $AX = Y$ 时，通常遇到该方程组是超定方程组的情况（行数大于列数），再求解时再上一个笔记中用了 $\min \frac{1}{2}(AX - Y)^2$ 这一目标进行求解，再次讨论一下为什么。

假设 $Y = AX + E$ 即 $y_i = a^T x_i + e_i$，这里的 e 为独立同分布的随机变量，且 $e \sim N(0, \sigma^2)$

则 $P(e) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{e^2}{2\sigma^2}}$

则 $P(y \mid x; e) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - a^T x)^2}{2\sigma^2}}$

所以 $y \sim N(a^T x, \sigma^2)$

我们的目标可以认为是对参数进行合理的估计，使得我们的模型是原始数据的最佳估计，为此定义似然函数为：

$$L(e) = P(y \mid x; e) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - ax)^2}{2\sigma^2}}$$

为了使 L(e)最大化，进行如下变形：

$$
\begin{aligned}
l(e) &= \log L(e) \\
&= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - ax)^2}{2\sigma^2}} \\
&= \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - ax)^2}{2\sigma^2}}\right) \\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{m} \frac{(y - ax)^2}{2\sigma^2}
\end{aligned}
$$

为使 l(e)最大化，仅需要 $\min \sum_{i=1}^{m} (y - ax)^2$，这就是前面我们讨论的最小二乘问题。

因此，进行最小二乘求解的一个很重要的假设是 e 相互独立且 $e \sim N(0, \sigma^2)$