波士顿房价预测

陈久宁

2018 年 11 月 30 日

1 问题介绍

波士顿房价预测问题是一个比较经典的入门案例，其数据集来自于 [1]，当前收录于`http://lib.stat.cmu.edu/datasets/boston`

2 线性模型

若采用如下线性模型来拟合该数据

$$\hat{y} = W^T x + b \qquad \text{(线性模型：预测)}$$

其中 $W \in \mathbb{R}^{13\times 1}$，$b \in \mathbb{R}$，我们需要从训练集 $\{(X,Y)|X \in \mathbb{R}^{13\times 506}, Y \in \mathbb{R}^{1\times 506}\}$ 中求解出一组 “最优” 的 (W^*, b^*)。所谓 “最优” 是在如下意义下成立的：

$$\begin{aligned}(W^*, b^*) &= \arg\min_{W,b} \ell_2(\hat{Y}, Y) \\ &= \arg\min_{W,b} \ell_2(W^T X + b, Y) \\ &:= \frac{1}{2}\left\|W^T X + b\mathbb{1} - Y\right\|_2^2 \\ &:= \frac{1}{2\times 506}\sum_{i=1}^{506}(W^T X_i + b - Y_i)^2\end{aligned} \qquad \text{(线性模型：训练)}$$

其中 $\mathbb{1} \in \mathbb{R}^{1\times 506}$ 为各个分量值全为 1 的向量。一旦求解出“最优”的 (W^*, b^*)，就可以将其代入等式 (线性模型：预测) 中进行房价预测。

实际中 Loss 的选择并不局限于本例中的 MSE(Mean Squared Error)。给定 Loss，模型训练的过程就是在求解出最优参数的过程，训练过程中参数是未知项；模型的预测过程就是利用求解出的最优参数代入等式 (线性模型：预测) 的过程，预测过程中 y 是未知项。

2.1 解析解–最小二乘法

等式 (线性模型：预测) 实际上是存在解析解 (closed-form solution) 的。因为该式可微，最优解只在驻点 (Saddle point)–梯度全为 0 的点–处取到，梯度计算如下：

$$\begin{aligned} \nabla_W \ell = &\quad X(X^T W + b\mathbb{1}^T - Y^T) = &\quad XX^T W + X\mathbb{1}^T b - XY^T \in \mathbb{R}^{13\times 1} \\ \frac{\partial \ell}{\partial b} = &\quad \mathbb{1}(X^T W + b\mathbb{1}^T - Y^T) = &\quad \mathbb{1}X^T W + \mathbb{1}\mathbb{1}^T b - \mathbb{1}Y^T \in \mathbb{R} \end{aligned} \tag{1}$$

令梯度为 0，则有

$$\begin{bmatrix} XX^T & X\mathbb{1}^T \\ \mathbb{1}X^T & \mathbb{1}\mathbb{1}^T \end{bmatrix} \begin{bmatrix} W \\ b \end{bmatrix} - \begin{bmatrix} XY^T \\ \mathbb{1}Y^T \end{bmatrix} = 0 \tag{2}$$

从而得到解析解

$$\begin{bmatrix} W \\ b \end{bmatrix} = \begin{bmatrix} XX^T & X\mathbb{1}^T \\ \mathbb{1}X^T & \mathbb{1}\mathbb{1}^T \end{bmatrix}^{-1} \begin{bmatrix} XY^T \\ \mathbb{1}Y^T \end{bmatrix} \tag{3}$$

关于最小二乘法的类似的推导可参考 [2]

2.2 梯度下降法

Data: X,Y
Result: W,b
Init: $W^0, b^0, \epsilon^0, \epsilon_{rel}$
while $\epsilon_{rel} \geq \epsilon_{tol}$ **do**

$$\begin{aligned}
\nabla_W \ell &= XX^T W^k + X\mathbb{1}^T b^k - XY^T \\
W^{k+1} &= W^k - t_k \nabla_W \ell \\
\frac{\partial \ell}{\partial b} &= \mathbb{1}X^T W^k + \mathbb{1}\mathbb{1}^T b^k - \mathbb{1}Y^T \\
b^{k+1} &= b^k - t_k \frac{\partial \ell}{\partial b} \\
\epsilon^{k+1} &= \ell(W, b) \\
\epsilon_{rel} &= |\epsilon^{k+1} - \epsilon^k| / \epsilon^k
\end{aligned}$$

end

Algorithm 1: 梯度下降

算法1描述了梯度下降的大致过程，使用该方法时需要注意以下三点：

- 固定学习率很难得到一个满意的精度，因此一般来说，每迭代一定次数之后都会降低学习率以保证收敛的精度。
- 如果不对 X 作归一化处理，则会导致 W 的各个维度之间的权重更新是不均衡的，这对最终收敛的结果影响很大。Y 不需要作归一化处理。对 X 作归一化处理之后求到的 (W, b) 的大小会根据 X 的归一化处理“自动”地调整。
- 为了避免大梯度导致的权重更新过大引起的“梯度爆炸”现象，一个常用的手段是作梯度裁剪 (Gradient Clipping)。梯度裁剪的方式有两种，一种为 $t_k \nabla_W \ell := \max(\min(t_k \nabla_W \ell, \lambda), -\lambda)$；另一种为 $t_k \nabla_W \ell := \frac{\lambda}{\|\nabla_W \ell\|_2^2} \nabla_W \ell \quad \text{if } \|\nabla_W \ell\|_2^2 \geq \lambda$，其中 $\lambda \geq 0$ 为裁剪的阈值。前者在实际中使用的更多一些，因为：后者在计算模长上有一定开销，而两种方法的最终效果都类似。

References

[1] David Harrison Jr and Daniel L Rubinfeld. "Hedonic housing prices and the demand for clean air". In: *Journal of environmental economics and management* 5.1 (1978), pp. 81–102.

[2] Steven J. Miller. *The Method of Least Squares*. `https://web.williams.edu/Mathematics/sjmiller/public_html/BrownClasses/54/handouts/MethodLeastSquares.pdf`. Nov. 2018.