线性回归基本原理

【参考资料】

吴恩达机器学习笔记 http://www.ai-start.com/ml2014/html/week2.html

1. 基本形式

假设模型特征为$(x_1, x_1, \ldots, x_n)$，n代表特征的数量；$x^{(i)}$代表第i个训练实例，$y^{(i)}$为对应的标签，m为样本数量；$m \times n$维矩阵X为特征矩阵。

线性回归模型为：

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

引入$x_0 = 1$，则有

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

此时模型中的参数是一个$n+1$维的向量，任何一个训练实例也是一个$n+1$维的向量，特征矩阵X的维度是$m \times (n+1)$。因此公式整体可以转化为矩阵形式

$$h_\theta(x) = \theta^T X$$

2. 梯度下降

线性回归的目标函数为

$$J(\theta_0, \theta_1 \ldots \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)^2$$

所以梯度更新的公式为：

$$\begin{aligned}
\theta_\mathrm{j} &:= \theta_\mathrm{j} - \alpha \frac{\partial}{\partial \theta_j} \mathrm{J}(\theta_0, \theta_1, \ldots, \theta_\mathrm{n}) \\
&= \theta_\mathrm{j} - \alpha \frac{\partial}{\partial \theta_\mathrm{j}} \frac{1}{2\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\mathrm{h}_\theta\left(\mathrm{x}^{(\mathrm{i})}\right) - \mathrm{y}^{(\mathrm{i})} \right)^2 \\
&= \theta_\mathrm{j} - \alpha \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\left(\mathrm{h}_\theta\left(\mathrm{x}^{(\mathrm{i})}\right) - \mathrm{y}^{(\mathrm{i})} \right) \cdot \mathrm{x}_\mathrm{j}^{(\mathrm{i})} \right)
\end{aligned}$$

3. 正规方程

除了梯度下降之外，线性回归模型的参数还可以通过正规方程法求解，即：

$$\theta = \left(X^T X\right)^{-1} X^T y$$

推导过程如下：

首先将目标函数

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)^2$$

改写为矩阵形式

$$J(\theta)=\frac{1}{2}(X\theta-y)^2$$

展开上式

$$\begin{aligned}
&J(\theta)=\tfrac{1}{2}(X\theta-y)^T(X\theta-y)\\
&=\tfrac{1}{2}\left(\theta^TX^T-y^T\right)(X\theta-y)\\
&=\tfrac{1}{2}\left(\theta^TX^TX\theta-\theta^TX^Ty-y^TX\theta-y^Ty\right)
\end{aligned}$$

接下来，对$J(\theta)$求偏导，需要用到以下几个矩阵的求导法则：

$$\begin{aligned}
\tfrac{dAB}{dB}&=A^T\\
\tfrac{dX^TAX}{dX}&=2AX
\end{aligned}$$

所以有

$$\begin{aligned}
&\tfrac{\partial J(\theta)}{\partial\theta}=\tfrac{1}{2}\left(2X^TX\theta-X^Ty-\left(y^TX\right)^T-0\right)\\
&=\tfrac{1}{2}\left(2X^TX\theta-X^Ty-X^Ty-0\right)\\
&=X^TX\theta-X^Ty
\end{aligned}$$

令$\frac{\partial J(\theta)}{\partial\theta}=0$，有

$$\theta=\left(X^TX\right)^{-1}X^Ty$$

注意，对于$\left(X^TX\right)^{-1}$不可逆的情况，比如特征之间相互不独立（即特征矩阵不满秩），正规方程法是无法使用的，不过在大多数情况下，$\left(X^TX\right)^{-1}$都是可逆的。

【梯度下降 v.s. 正规方程】

对比如下：

梯度下降	正规方程
需要选择学习率	不需要
需要多次迭代	一次运算得出
当特征数量n大时也能较好适用	需要计算$\left(X^T X\right)^{-1}$,如果特征数量n较大则运算代价大，因为矩阵逆的计算时间复杂度为$O\left(n^3\right)$，通常来说当n小于10000 时还是可以接受的
适用于各种类型的模型	只适用于线性模型，不适合逻辑回归模型等其他模型

总结一下，只要特征变量的数目并不大，标准方程是一个很好的计算参数θ的替代方法。