线性回归

假设数据集为：

$$\mathcal{D}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\} \tag{1}$$

后满我们记：

$$X=(x_1,x_2,\cdots,x_N)^T,Y=(y_1,y_2,\cdots,y_N)^T \tag{2}$$

线性回归假设：

$$f(w)=w^Tx \tag{3}$$

最小二乘法

对这个问题，采用二范数定义的平方误差来定义损失函数：

$$L(w)=\sum\limits_{i=1}^N||w^Tx_i-y_i||^2_2 \tag{4}$$

展开得到：

$$\begin{aligned}L(w)&=(w^Tx_1-y_1,\cdots,w^Tx_N-y_N)\cdot (w^Tx_1-y_1,\cdots,w^Tx_N-y_N)^T\\&=(w^TX^T-Y^T)\cdot (Xw-Y)=w^TX^TXw-Y^TXw-w^TX^TY+Y^TY\\&=w^TX^TXw-2w^TX^TY+Y^TY\end{aligned} \tag{5}$$

最小化这个值的 $\hat{w}$ ：

$$\begin{aligned}\hat{w}=\mathop{argmin}\limits_wL(w)&\longrightarrow\frac{\partial}{\partial w}L(w)=0\\&\longrightarrow2X^TX\hat{w}-2X^TY=0\\&\longrightarrow \hat{w}=(X^TX)^{-1}X^TY=X^+Y\end{aligned} \tag{6}$$

这个式子中 $(X^TX)^{-1}X^T$ 又被称为伪逆。对于行满秩或者列满秩的 X，可以直接求解，但是对于非满秩的样本集合，需要使用奇异值分解（SVD）的方法，对 X 求奇异值分解，得到

$$X=U\Sigma V^T \tag{7}$$

于是：

$$X^+=V\Sigma^{-1}U^T \tag{8}$$

在几何上，最小二乘法相当于模型（这里就是直线）和试验值的距离的平方求和，假设我们的试验样本张成一个 p 维空间（满秩的情况）：$X=Span(x_1,\cdots,x_N)$，而模型可以写成 $f(w)=X\beta$，也就是 $x_1,\cdots,x_N$ 的某种组合，而最小二乘法就是说希望 Y 和这个模型距离越小越好，于是它们的差应该与这个张成的空间垂直：

$$X^T\cdot(Y-X\beta)=0\longrightarrow\beta=(X^TX)^{-1}X^TY \tag{9}$$

噪声为高斯分布的 MLE

对于一维的情况，记 $y=w^Tx+\epsilon,\epsilon\sim\mathcal{N}(0,\sigma^2)$，那么 $y\sim\mathcal{N}(w^Tx,\sigma^2)$。代入极大似然估计中：

$$\begin{aligned}L(w)=\log p(Y|X,w)&=\log\prod_{i=1}^Np(y_i|x_i,w)\\&=\sum_{i=1}^N\log(\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(y_i-w^Tx_i)^2}{2\sigma^2}})\end{aligned} \tag{10}$$

$$\mathop{argmax}\limits_w L(w)=\mathop{argmin}\limits_w\sum_{i=1^N}(y_i-w^Tx_i)^2 \tag{11}$$

这个表达式和最小二乘估计得到的结果一样。

权重先验也为高斯分布的 MAP

取先验分布 $w\sim\mathcal{N}(0,\sigma_0^2)$。于是：

$$\begin{aligned}\hat{w}=\mathop{argmax}\limits_w p(w|Y)&=\mathop{argmax}\limits_w p(Y|w)p(w)\\&=\mathop{argmax}\limits_w\log p(Y|w)p(w)\\&=\mathop{argmax}\limits_w(\log p(Y|w)+\log p(w))\\&=\mathop{argmin}\limits_w[(y-w^Tx)^2+\frac{\sigma^2}{\sigma_0^2}w^Tw]\end{aligned} \tag{12}$$

这里省略了 X，$p(Y)$和 w 没有关系，同时也利用了上面高斯分布的 MLE的结果。

我们将会看到，超参数 σ_0 的存在和下面会介绍的 Ridge 正则项可以对应，同样的如果将先验分布取为 Laplace 分布，那么就会得到和 L1 正则类似的结果。

正则化

在实际应用时，如果样本容量不远远大于样本的特征维度，很可能造成过拟合，对这种情况，我们有下面三个解决方式：

1. 加数据
2. 特征选择（降低特征维度）如 PCA 算法。
3. 正则化

正则化一般是在损失函数（如上面介绍的最小二乘损失）上加入正则化项（表示模型的复杂度对模型的惩罚），下面我们介绍一般情况下的两种正则化框架。

$$L1:\mathop{argmin}\limits_w L(w)+\lambda||w||_1,\lambda\gt0 \tag{13}$$

$$L2:\mathop{argmin}\limits_w L(w)+\lambda||w||^2_2,\lambda \gt 0 \tag{14}$$

下面对最小二乘误差分别分析这两者的区别。

L1 Lasso

L1正则化可以引起稀疏解。

从最小化损失的角度看，由于 L1 项求导在0附近的左右导数都不是0，因此更容易取到0解。

从另一个方面看，L1 正则化相当于：

$$
\begin{aligned}
&\mathop{argmin}_w L(w)\\
&s.t. ||w||_1 < C
\end{aligned}
\tag{15}
$$

我们已经看到平方误差损失函数在 w 空间是一个椭球，因此上式求解就是椭球和 $||w||_1=C$的切点，因此更容易相切在坐标轴上。

L2 Ridge

$$
\begin{aligned}
\hat{w}=\mathop{argmin}_w L(w)+\lambda w^Tw &\longrightarrow \frac{\partial}{\partial w}L(w)+2\lambda w=0\\
&\longrightarrow 2X^TX\hat{w}-2X^TY+2\lambda\hat{w}=0\\
&\longrightarrow \hat{w}=(X^TX+\lambda\mathbb{I})^{-1}X^TY
\end{aligned}
\tag{16}
$$

可以看到，这个正则化参数和前面的 MAP 结果不谋而合。利用2范数进行正则化不仅可以是模型选择 w 较小的参数，同时也避免 X^TX不可逆的问题。

小结

线性回归模型是最简单的模型，但是麻雀虽小，五脏俱全，在这里，我们利用最小二乘误差得到了闭式解。同时也发现，在噪声为高斯分布的时候，MLE 的解等价于最小二乘误差，而增加了正则项后，最小二乘误差加上 L2 正则项等价于高斯噪声先验下的 MAP解，加上 L1 正则项后，等价于 Laplace 噪声先验。

传统的机器学习方法或多或少都有线性回归模型的影子：

1. 线性模型往往不能很好地拟合数据，因此有三种方案克服这一劣势：
   1. 对特征的维数进行变换，例如多项式回归模型就是在线性特征的基础上加入高次项。
   2. 在线性方程后面加入一个非线性变换，即引入一个非线性的激活函数，典型的有线性分类模型如感知机。
   3. 对于一致的线性系数，我们进行多次变换，这样同一个特征不仅仅被单个系数影响，例如多层感知机（深度前馈网络）。
2. 线性回归在整个样本空间都是线性的，我们修改这个限制，在不同区域引入不同的线性或非线性，例如线性样条回归和决策树模型。
3. 线性回归中使用了所有的样本，但是对数据预先进行加工学习的效果可能更好（所谓的维数灾难，高维度数据更难学习），例如 PCA 算法和流形学习。