Bayes Linear Classification 01 Background

Chen Gong

05 November 2019

数据集 $D = \{(x_i, y_i)\}_{i=1}^N$，其中 $x_i \in \mathbb{R}^p$，$y_i \in \mathbb{R}$。
数据矩阵为：(这样可以保证每一行为一个数据点)

$$X = (x_1, x_2, \cdots, x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{32} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}_{N \times P} \tag{1}$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N \times 1} \tag{2}$$

拟合函数我们假设为：$f(x) = w^T x = x^T w$。
预测值 $y = f(x) + \varepsilon$，其中 ε 是一个 Guassian Noise，并且 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$。
并且，x, y, ε 都是 Random variable。

1 最小二乘估计 (Least Square Estimation)

这实际上就是一个利用数据点的极大似然估计 (MLE)，并且有一个默认的隐含条件，也就是噪声 ε 符合 Gaussian Distribution。我们的目标是通过估计找到 w，使得：

$$w_{MLE} = argmax_w p(Data|w) \tag{3}$$

而如果仅仅只是这样来使用，很容易会出现过拟合的问题。所以，我们引入了 Regularized LSE，也就是正则化最小二乘法。同时也有一个默认的隐含条件，也是噪声 ε 符合 Gaussian Distribution。在 Liner Regression 中我们提到了有两种方法来进行思考，也就是 Lasso 和 Ridge Regression。在这里我们可以使用一个 Bayes 公式，那么：

$$p(w|Data) \propto p(Data|w)p(w) \tag{4}$$

$$w_{MAP} = argmax_w p(w|Data) = argmax_w p(Data|w)p(w) \tag{5}$$

那么假设 $p(w)$ 符合一个高斯分布 $\mathcal{N}(\mu_0, \Sigma_0)$ 时，这时是属于 Ridge（具体在线性回章节有介绍，也就是正则化的最小二乘估计 $\Leftrightarrow$ 先验服从高斯分布的极大后验估计）；而如果 $p(w)$ 符合一个 Laplace

分布，这是就是 Lasso。从概率的角度来思考和统计的角度来思想，我们其实获得的结果是一样的，这在 Linear Regression 中有证明。但是，我们只证明了 Ridge 的部分。

2 贝叶斯估计与频率派估计

其实在第一部分，我们讲的都是点估计，频率派估计的部分。因为在这些思路中，我们把参数 w 当成 a unknown random variable。这实际上就是一个优化问题。而在 Beyesian method 中，认为 w 是一个随机变量，也就是一个分布，那么我们求的 w 不再是一个数了，而是一个分布。下面我们将要进行 Bayes Linear Regression 的部分。