

Bayes Linear Classification 02 Inference

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)

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

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

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 Gaussian Noise, 并且 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ 。

并且, x, y, ε 都是 Random variable。

贝叶斯估计方法 (Bayesian Method), 可以分为两个步骤, 1.Inference, 2.Prediction。Inference 的关键在于估计 $\text{posterior}(w)$; 而 Prediction 的关键在于对于给定的 x^* 求出预测值 y^* 。

1 Bayesian Method 模型建立

首先我们需要对公式使用贝叶斯公式进行分解, 便于计算:

$$p(w|Data) = p(w|X, Y) = \frac{p(w, Y|X)}{p(Y|X)} = \frac{p(Y|X, w)p(w)}{\int_w p(Y|X, w)p(w)dw} \quad (3)$$

其中 $p(Y|X, w)$ 是似然函数 (likelihood function), $p(w)$ 是一个先验函数 (prior function)。实际这里省略了一个过程, $p(w, Y|X) = p(Y|X, w)p(w|X)$ 。但是很显然, $p(w|X)$ 中 X 与 w 之间并没有直接的联系。所以 $p(w|X) = p(w)$ 。

似然函数的求解过程为:

$$p(Y|X, w) = \prod_{i=1}^N p(y_i|x_i, w) \quad (4)$$

又因为 $y = w^T x + \varepsilon$, 并且 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ 。所以

$$p(y_i|x_i, w) = \mathcal{N}(w^T x_i, \sigma^2) \quad (5)$$

所以，

$$p(Y|X, w) = \prod_{i=1}^N p(y_i|x_i, w) = \prod_{i=1}^N \mathcal{N}(w^T x_i, \sigma^2) \quad (6)$$

而下一步，我们假设 $p(w) = \mathcal{N}(0, \Sigma_p)$ 。又因为 $p(Y|X)$ 与参数 w 无关，所以这是一个定值。所以，我们可以将公式改写为：

$$p(w|X, Y) \propto p(Y|w, X)p(w) \quad (7)$$

在这里我们将使用到一个共轭的技巧，因为 likelihood function 和 prior function 都是 Gaussian Distribution，所有 posterior 也一定是 Gaussian Distribution。所以，我们可以将公式改写为：

$$p(w|Data) \sim \mathcal{N}(\mu_w, \Sigma_w) \propto \prod_{i=1}^N \mathcal{N}(w^T x_i, \sigma^2) \mathcal{N}(0, \Sigma_p) \quad (8)$$

我们的目的就是求解 $\mu_w = ?, \Sigma_w = ?$ 。

2 模型的求解

对于 likelihood function 的化简如下所示：

$$p(Y|X, w) = \prod_{i=1}^N \frac{1}{(2\pi)^{\frac{1}{2}} \sigma} \exp \left\{ -\frac{1}{2\sigma^2} (y_i - w^T x_i)^2 \right\} \quad (9)$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - w^T x_i)^2 \right\} \quad (10)$$

下一步，我们希望将 $\sum_{i=1}^N (y_i - w^T x_i)^2$ 改写成矩阵相乘的形式，

$$\begin{aligned} \sum_{i=1}^N (y_i - w^T x_i)^2 &= \begin{bmatrix} y_1 - w^T x_1 & y_2 - w^T x_2 & \cdots & y_i - w^T x_i \end{bmatrix} \begin{bmatrix} y_1 - w^T x_1 \\ y_2 - w^T x_2 \\ \vdots \\ y_i - w^T x_i \end{bmatrix} \\ &= (Y^T - W^T X^T)(Y - XW)^T \\ &= (Y^T - W^T X^T)(Y - XW) \end{aligned} \quad (11)$$

所以，

$$\begin{aligned} p(Y|X, w) &= \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^N (Y^T - W^T X^T)(Y - XW) \right\} \\ &= \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} \exp \left\{ -\frac{1}{2} \sum_{i=1}^N (Y^T - W^T X^T) \sigma^{-2} I (Y - XW) \right\} \\ p(Y|X, w) &\sim \mathcal{N}(WX, \sigma^2 I) \end{aligned} \quad (12)$$

那么，将化简后的结果带入有：

$$p(w|Data) \sim \mathcal{N}(\mu_w, \Sigma_w) \propto \mathcal{N}(WX, \sigma^2 I) \mathcal{N}(0, \Sigma_p) \quad (13)$$

$$\begin{aligned}\mathcal{N}(WX, \sigma^2 I) \mathcal{N}(0, \Sigma_p) &\propto \exp \left\{ -\frac{1}{2} (Y - WX)^T \sigma^{-2} I (Y - WX) - \frac{1}{2} w^T \Sigma_p^{-1} w \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} (Y^T Y - 2Y^T XW + W^T X^T XW) - \frac{1}{2} W^T \Sigma_p^{-1} W \right\}\end{aligned}\quad (14)$$

那么这个公式长得怎么的难看我们怎么确定我们想要的 μ_w, Σ_w 。由于知道 posterior 必然是一个高斯分布，那么我们采用待定系数法来类比确定参数的值即可。对于一个分布 $p(x) \sim \mathcal{N}(\mu, \Sigma)$ ，他的指数部分为：

$$\exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} = \exp \left\{ -\frac{1}{2} (x^T \Sigma^{-1} x - 2\mu^T \Sigma^{-1} x + \Delta) \right\} \quad (15)$$

常数部分已经不重要了，对于我们的求解来说没有任何的用处，所以，我们直接令它为 Δ 。那么，我们类比一下就可以得到，

$$x^T \Sigma^{-1} x = W^T \sigma^{-2} X^T X W + W^T \Sigma_p^{-1} W \quad (16)$$

所以，我们可以得到 $\Sigma_w^{-1} = \sigma^{-2} X^T X + \Sigma_p^{-1}$ 。并且，我们令 $\Sigma_w^{-1} = A$ 。

从二次项中我们得到了 Σ_w^{-1} ，那么，下一步，我们期望可以从一次项中得到 μ_A 的值。我们将一次项提取出来进行观察，可以得到。

$$\mu^T A = \sigma^{-2} Y^T X \quad (17)$$

$$(\mu^T A)^T = (\sigma^{-2} Y^T X)^T \quad (18)$$

$$A^T \mu = \sigma^{-2} X^T Y \quad (19)$$

$$\mu = \sigma^{-2} (A^T)^{-1} X^T Y \quad (20)$$

有因为， Σ_w 是一个方差矩阵，那么他一定是对称的，所以 $A^T = A$ 。于是

$$\mu_m = \sigma^{-2} A^{-1} X^T Y \quad (21)$$

3 小结

我们利用贝叶斯推断的方法来确定参数之间的分布，也就是确定 $p(W|X, Y)$ 。我们使用 Bayes 的方法，确定为 $p(W|X, Y) \propto p(Y|W, X)p(W)$ 。并且确定一个噪声分布 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ 。那么，

$$p(Y|W, X) \sim \mathcal{N}(W^T X, \sigma^2) \quad (22)$$

$$P(W) \sim \mathcal{N}(0, \Sigma_p) \quad (23)$$

通过推导，我们可以得出，

$$p(W|X, Y) \sim \mathcal{N}(\mu_w, \Sigma_w) \quad (24)$$

其中，

$$\Sigma_w^{-1} = \sigma^{-2} X^T X + \Sigma_p^{-1} \quad \mu_m = \sigma^{-2} A^{-1} X^T Y \quad \Sigma_w^{-1} = A \quad (25)$$