Linear Classification 05

Chen Gong

03 November 2019

前面讲的方法都是概率判别模型，包括，Logistic Regression 和 Fisher 判别分析。接下来我们将要学习的是概率生成模型部分，也就是现在讲到的 Gaussian Discriminate Analysis。数据集的相关定义为：

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

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

那么，我们的数据集可以记为 $\{(x_i, y_i)\}_{i=1}^N$，其中，$x_i \in \mathbb{R}^p$，$y_i \in \{+1, -1\}$。我们将样本点分成了两个部分：

$$\begin{cases} C_1 = \{x_i | y_i = 1, i = 1, 2, \cdots, N_1\} \\ C_2 = \{x_i | y_i = 0, i = 1, 2, \cdots, N_2\} \end{cases} \tag{3}$$

并且有 $|C_1| = N_1$，$|C_2| = N_2$，且 $N_1 + N_2 = N$。

1 概率判别模型与生成模型的区别

什么是判别模型？所谓判别模型，也就是求

$$\hat{y} = argmax\ p(y|x) \qquad y \in \{0, 1\} \tag{4}$$

重点在于求出这个概率来，知道这个概率的值等于多少。而概率生成模型则完全不一样。概率生成模型不需要知道概率值具体是多大，只需要知道谁大谁小即可，举例即为 $p(y = 0|x)$ 和 $p(y = 1|x)$，谁大谁小的问题。而概率生成模型的求法可以用贝叶斯公式来进行求解，即为：

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x, y)}{p(x)} \propto p(x, y) \tag{5}$$

因为在这个公式中，比例大小 $p(x)$ 与 y 的取值无关，所以它是一个定值。所以，概率生成模型实际上关注的就是一个求联合概率分布的问题。那么，总结一下

$$p(y|x) \propto p(x|y)p(y) \propto p(x,y) \tag{6}$$

其中，$p(y|x)$ 为 Posterior function，$p(y)$ 为 Prior function，p(x|y) 为 Likehood function。所以有

$$\hat{y} = argmax_{y\in\{0,1\}} p(y|x) \propto argmax_{y\in\{0,1\}} p(x|y)p(y) \tag{7}$$

2 Gaussian Discriminate Analysis 模型建立

在二分类问题中，很显然可以得到，我们的先验概率符合，$p(y)$ ~Bernoulli Distribution。也就是，

y	1	0
p	φ	$1-\varphi$

表 1: Bernoulli 分布的概率分布表

所以，可以写出：

$$p(y) = \begin{cases} \varphi^y & y=1 \\ (1-\varphi)^{1-y} & y=0 \end{cases} \Rightarrow \varphi^y(1-\varphi)^{1-y} \tag{8}$$

而随后是要确定似然函数，我们假设他们都符合高斯分布。对于不同的分类均值是不同的，但是不同变量之间的协方差矩阵是一样的。那么我们可以写出如下的形式：

$$p(x|y) = \begin{cases} p(x|y=1) \sim \mathcal{N}(\mu_1, \Sigma) \\ p(x|y=0) \sim \mathcal{N}(\mu_2, \Sigma) \end{cases} \Rightarrow \mathcal{N}(\mu_1,\Sigma)^y \mathcal{N}(\mu_2,\Sigma)^{1-y} \tag{9}$$

那么我们的 Likehood function 可以被定义为：

$$\begin{aligned}
\mathcal{L}(\theta) &= \log \prod_{i=1}^{N} p(x_i, y_i) \\
&= \sum_{i=1}^{N} \log p(x_i, y_i) \\
&= \sum_{i=1}^{N} \log p(x_i|y_i)p(y_i) \\
&= \sum_{i=1}^{N} [\log p(x_i|y_i) + \log p(y_i)] \\
&= \sum_{i=1}^{N} \left[\log \mathcal{N}(\mu_1,\Sigma)^{y_i} \mathcal{N}(\mu_2,\Sigma)^{1-y_i} + \log \varphi_i^y (1-\varphi)^{1-y_i}\right] \\
&= \sum_{i=1}^{N} \log \mathcal{N}(\mu_1,\Sigma)_i^y + \sum_{i=1}^{N} \log \mathcal{N}(\mu_2,\Sigma)^{1-y_i} + \sum_{i=1}^{N} \log \varphi^{y_i} + \sum_{i=1}^{N} \log(1-\varphi)^{1-y_i}
\end{aligned} \tag{10}$$

为了方便后续的推演过程，所以，我们将 Likehood function 写成，

$$\mathcal{L}(\theta) = ①+②+③$$

并且,我们令:$① = \sum_{i=1}^{N} \log \mathcal{N}(\mu_1, \Sigma)_i^y$,$② = \sum_{i=1}^{N} \log \mathcal{N}(\mu_2, \Sigma)^{1-y_i}$,$③ = \sum_{i=1}^{N} \log \varphi^{y_i} + \sum_{i=1}^{N} \log(1-\varphi)^{1-y_i}$。那么上述函数我们可以表示为:

$$\theta = (\mu_1, \mu_2, \Sigma, \varphi) \qquad \hat{\theta} = argmax_\theta \mathcal{L}(\theta) \tag{11}$$

3 Likehood functioon 参数的极大似然估计

Likehood function 的参数为 $\theta = (\mu_1, \mu_2, \Sigma, \varphi)$，下面我们分别用极大似然估计对这四个参数进行求解。下面引入几个公式:

$$tr(AB) = tr(BA) \tag{12}$$

$$\frac{\partial tr(AB)}{\partial A} = B^T \tag{13}$$

$$\frac{\partial |A|}{\partial A} = |A|A^{-1} \tag{14}$$

$$\frac{\partial In|A|}{\partial A} = A^{-1} \tag{15}$$

3.1 求解 φ

$$③ = \sum_{i=1}^{N} \log \varphi^{y_i} + \sum_{i=1}^{N} \log(1-\varphi)^{1-y_i} = \sum_{i=1}^{N} y_i \log \varphi + \sum_{i=1}^{N} (1-y_i) \log(1-\varphi)$$

$$\frac{\partial ③}{\partial \varphi} = \sum_{i=1}^{N} y_i \frac{1}{\varphi} + \sum_{i=1}^{N} (1-y_i) \frac{1}{1-\varphi} = 0 \tag{16}$$

$$\sum_{i=1}^{N} y_i(1-\varphi) + (1-y_i)\varphi = 0 \tag{17}$$

$$\sum_{i=1}^{N} (y_i - \varphi) = 0 \tag{18}$$

$$\hat{\varphi} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{19}$$

又因为 $y_i = 0$ 或 $y_i = 1$，所以 $\hat{\varphi} = \frac{1}{N} \sum_{i=1}^{N} y_i = \frac{N_1}{N}$。

3.2 求解 μ_1

$$① = \sum_{i=1}^{N} \log \mathcal{N}(\mu_1, \Sigma)^{y_i}$$
$$= \sum_{i=1}^{N} y_i \log \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}(x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1)\right\}$$

那么求解过程如下所示：由于到对 μ_1 求偏导，我们只需要关注公式中和 μ_1 有关的部分。那么我们可以将公式化简为:

$$\sum_{i=1}^{N} y_i \log exp\left\{-\frac{1}{2}(x_i - \mu_1)^T \Sigma^{-1} (x_i - \mu_1)\right\} \tag{20}$$

然后将 exp 和 log 抵消掉，再将括号打开，我们可以得到最终的化简形式：

$$-\frac{1}{2}\sum_{i=1}^{N} y_i \left\{x_i^T\Sigma^{-1}x_i - x_i^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x_i + \mu_1^T\Sigma^{-1}\mu_1\right\} \tag{21}$$

又因为 x_i 是 $px1$ 的矩阵，Σ^{-1} 是 pxp 的矩阵，并且 μ_1 也是 $px1$ 的矩阵。所以，$x_i^T\Sigma^{-1}\mu_1$ 和 $\mu_1^T\Sigma^{-1}x_i$ 都是一维的实数，所以 $x_i^T\Sigma^{-1}\mu_1 = \mu_1^T\Sigma^{-1}x_i$。所以：

$$① = -\frac{1}{2}\sum_{i=1}^{N} y_i \left\{x_i^T\Sigma^{-1}x_i - 2\mu_1^T\Sigma^{-1}x_i + \mu_1^T\Sigma^{-1}\mu_1\right\} \tag{22}$$

为了方便表示，我们令① $= \triangle$。所以，极大似然法求解过程如下：

$$\begin{aligned}
\frac{\partial\triangle}{\partial\mu_1} =& -\frac{1}{2}\sum_{i=1}^{N} y_i(-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu_1) = 0 \\
=& \sum_{i=1}^{N} y_i(\Sigma^{-1}x_i - \Sigma^{-1}\mu_1) = 0 \\
=& \sum_{i=1}^{N} y_i(x_i - \mu_1) = 0 \\
=& \sum_{i=1}^{N} y_i x_i = \sum_{i=1}^{N} y_i\mu_1 \\
\mu_1 =& \frac{\sum_{i=1}^{N} y_i x_i}{\sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} y_i x_i}{N_1}
\end{aligned} \tag{23}$$

3.3 求解 μ_2

μ_2 的求解过程与 μ_1 的基本保持一致性。区别点从公式 (22) 开始，我们有：

$$② = -\frac{1}{2}\sum_{i=1}^{N}(1-y_i) \left\{x_i^T\Sigma^{-1}x_i - 2\mu_1^T\Sigma^{-1}x_i + \mu_1^T\Sigma^{-1}\mu_1\right\} \tag{24}$$

极大似然法的求解过程如下所示：

$$\begin{aligned}
\frac{\partial\triangle}{\partial\mu_2} =& -\frac{1}{2}\sum_{i=1}^{N}(1-y_i)(-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu_1) = 0 \\
=& \sum_{i=1}^{N}(1-y_i)(x_i - \mu_1) = 0 \\
=& \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i = N\mu_1 - \sum_{i=1}^{N} y_i\mu_1 \\
\mu_2 =& \frac{\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i}{N - \sum_{i=1}^{N} y_i} = \frac{\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i}{N - N_1}
\end{aligned} \tag{25}$$

3.4 求解 Σ

如果要使用极大似然估计来求解 Σ，这只与 $\mathcal{L}(\theta)$ 中的①和②有关。并且①+②的表达式为：

$$\sum_{i=1}^{N} \log\mathcal{N}(\mu_1, \Sigma)^{y_i} + \sum_{i=1}^{N} \log\mathcal{N}(\mu_2, \Sigma)^{1-y_i} \tag{26}$$

那么，按照分类点的方法，我们可以将其改写为：

$$\hat{\Sigma} = argmax \sum_{x \in C_1} \log \mathcal{N}(\mu_1, \Sigma) + \sum_{x \in C_2} \log \mathcal{N}(\mu_2, \Sigma) \tag{27}$$

公式加号前后都是一样的，所以，为了方便计算我们暂时只考虑一半的计算：

$$\begin{aligned} \sum_{i=1}^{N} \log \mathcal{N}(\mu, \Sigma) =& \sum_{i=1}^{N} \log \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} exp\left\{ -\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) \right\} \\ =& -\sum_{i=1}^{N} \frac{p}{2} \log 2\pi - \sum_{i=1}^{N} \frac{1}{2} \log |\Sigma| + -\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) \\ =& C - \sum_{i=1}^{N} \frac{1}{2} \log |\Sigma| + -\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) \end{aligned} \tag{28}$$

又因为，$-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)$ 是一个一维实数，判断法方法与前文的一样。所以，

$$\begin{aligned} \sum_{i=1}^{N} \log \mathcal{N}(\mu, \Sigma) =& C - \sum_{i=1}^{N} \frac{1}{2} \log |\Sigma| + -tr\left(\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu) \right) \\ =& C - \sum_{i=1}^{N} \frac{1}{2} \log |\Sigma| + -tr\left(\frac{1}{2}(x_i - \mu)^T (x_i - \mu) \Sigma^{-1} \right) \end{aligned} \tag{29}$$

而且，

$$S = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^T (x_i - \mu) \tag{30}$$

所以，

$$\sum_{i=1}^{N} \log \mathcal{N}(\mu, \Sigma) = C - \sum_{i=1}^{N} \frac{1}{2} \log |\Sigma| + -\frac{1}{2} N tr(S\Sigma^{-1}) \tag{31}$$

那么代入公式 (27) 中，我们可以得到：

$$\begin{aligned} \hat{\Sigma} =& argmax_\Sigma \; C - \sum_{i=1}^{N_1} \frac{1}{2} \log |\Sigma| - \frac{1}{2} N_1 tr(S_1 \Sigma^{-1}) + C - \sum_{i=1}^{N_2} \frac{1}{2} \log |\Sigma| - \frac{1}{2} N_2 tr(S_2 \Sigma^{-1}) \\ =& argmax_\Sigma \; C - \frac{1}{2} \sum_{i=1}^{N} \log |\Sigma| - \frac{1}{2} N_1 tr(S_1 \Sigma^{-1}) - \frac{1}{2} N_2 tr(S_2 \Sigma^{-1}) \end{aligned} \tag{32}$$

我们令函数 $C - \frac{1}{2} \sum_{i=1}^{N} \log |\Sigma| - \frac{1}{2} N_1 tr(S_1 \Sigma^{-1}) - \frac{1}{2} N_2 tr(S_2 \Sigma^{-1}) = \triangle$，那么对 Σ 求偏导可得：

$$\frac{\partial \triangle}{\partial \Sigma} = \frac{1}{2} \left(N\Sigma^{-1} - N_1 S_1^T \Sigma^{-2} - N_2 S_2^T \Sigma^{-2} \right) = 0 \tag{33}$$

又因为方差矩阵是对称矩阵，所以 $S_1^T = S_1$ 并且 $S_2^T = S_2$。所以：

$$\begin{aligned} \frac{\partial \triangle}{\partial \Sigma} =& N\Sigma^{-1} - N_1 S_1 \Sigma^{-2} - N_2 S_2 \Sigma^{-2} = 0 \\ =& N\Sigma - N_1 S_1 - N_2 S_2 = 0 \end{aligned} \tag{34}$$

解得：

$$\Sigma = \frac{N_1 S_1 + N_2 S_2}{N} \tag{35}$$

4 总结

下面对 Gaussian Discriminate Analysis 做一个简单的小结。我们使用模型为:

$$\hat{y} = argmax_{y\in\{0,1\}} p(y|x) \propto argmax_{y\in\{0,1\}} p(x|y)p(y) \tag{36}$$

$$\begin{cases} p(y) = \varphi^y(1-\varphi)^{1-y} \\ p(x|y) = \mathcal{N}(\mu_1, \Sigma)^y \mathcal{N}(\mu_2, \Sigma)^{1-y} \end{cases} \tag{37}$$

利用极大似然估计得到的结果为:

$$\theta = (\mu_1, \mu_2, \Sigma, \varphi) = \begin{cases} \hat{\varphi} = \frac{1}{N}\sum_{i=1}^{N} y_i \\ \mu_1 = \frac{\sum_{i=1}^{N} y_i x_i}{N_1} \\ \mu_2 = \frac{\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} y_i x_i}{N - N_1} \\ \Sigma = \frac{N_1 S_1 + N_2 S_2}{N} \end{cases} \tag{38}$$