# 小锅的机器学习笔记–线性模型

设:

数据集矩阵为: $X = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}_{n\times(p+1)}$ $\quad X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \\ 1 \end{bmatrix}_{(p+1)\times 1}$ 其中每一个$X_i$为一个样本

$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$ 其中每一个$y_i$为一个对应样本的观测值

$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \\ b \end{bmatrix}_{(p+1)\times 1}$ 其中 $b$ 为偏执项

# 线性回归:

## 最小二乘法:

损失函数即为预测值与观测值的误差:

$$L(W) = (W^TX^T - Y^T)(W^TX^T - Y^T)^T = W^TX^TXW - 2W^TX^TY + Y^TY \tag{1}$$

所以:

$$\frac{\mathrm{d}L(W)}{\mathrm{d}W} = \frac{\mathrm{d}[W^TX^TXW]}{\mathrm{d}W} - 2\frac{\mathrm{d}[W^TX^TY]}{\mathrm{d}W}$$

因为:

$$\frac{\mathrm{d}[W^TX^TXW]}{\mathrm{d}W} = \frac{\mathrm{d}[W^T]}{\mathrm{d}W}X^TXW + \frac{\mathrm{d}[W]}{\mathrm{d}W}(W^TX^TX)^T = 2X^TXW \tag{2}$$

令:

$$A = X^TY \quad A = [a_1, a_2, \ldots, a_{p+1}]^T_{(p+1)\times 1}$$

$$\frac{\mathrm{d}[W^TX^TY]}{\mathrm{d}W} = \begin{bmatrix} \dfrac{\partial W^TA}{\partial w_1} \\ \\ \dfrac{\partial W^TA}{\partial w_2} \\ \\ \vdots \\ \\ \dfrac{\partial W^TA}{\partial w_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \sum_{i=1}^{p}(w_ia_i) + b \cdot a_{p+1}}{\partial w_1} \\ \\ \dfrac{\partial \sum_{i=1}^{p}(w_ia_i) + b \cdot a_{p+1}}{\partial w_2} \\ \\ \vdots \\ \\ \dfrac{\partial \sum_{i=1}^{p}(w_ia_i) + b \cdot a_{p+1}}{\partial b} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{p+1} \end{bmatrix} = X^TY \tag{3}$$

所以可以知道:

$$\frac{\mathrm{d}L(W)}{\mathrm{d}W} = 2X^TXW - 2X^TY \tag{4}$$

我们令 $\dfrac{\mathrm{d}L(W)}{\mathrm{d}W} == 0$，且当 $X^TX$ 可逆时: 可以知道:

$$W = (X^TX)^{-1}X^TY \tag{5}$$

## 概率视角看待线性回归:

我们假设数据集可以通过线性完美拟合，但是对于每一个观察值 $y_i$，当我们在观察到它时总是会受到噪声 $\varepsilon$ 的影响:

$$y_i = y_i^* + \varepsilon_i \qquad \text{其中 } y_i^* \text{ 为真实值} \tag{6}$$

我们设:

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \text{且 } \varepsilon_i \text{ 服从均值为 } 0\text{，方差为 } \sigma^2 \text{ 的高斯分布，且 } \varepsilon_i \text{ 之间相互独立} \tag{7}$$

则有:

$$Y = W^TX^T + \varepsilon \tag{8}$$

所以有:

$$Y|W \quad \sim\sim \quad N(W^TX^T, \sigma^2) \tag{9}$$

所以:

$$\begin{aligned}W_{MAP} &= \underset{W}{argmax} P(Y|W) = \underset{W}{argmax} \sum_{i=1}^{n} \log P(y_i, X_i|w) \\ &= \underset{W}{argmax} \sum_{i=1}^{n} [\log(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2}(y_i - W^T X_i)^2] \\ &= \underset{W}{argmin} \sum_{i=1}^{n} (y_i - W^T X_i)^2 \\ &= (W^T X^T - Y^T)(W^T X^T - Y^T)^T\end{aligned}$$

可以观察到最终要优化的函数和最小二乘法所要优化的是一样的。

## 线性回归的正则化 (岭回归):

其实就是修改损失函数, 其中 $\lambda$ 是正则化系数:

$$J(W) = L(W) + \lambda W^T W == W^T X^T X W - 2W^T X^T Y + Y^T Y + \lambda W^T W \tag{10}$$

求导可得:

$$\frac{\partial J(w)}{\partial W} = 2(X^T X + \lambda I) - 2X^T Y \tag{11}$$

令倒数等于 0 可以得到:

$$W = (X^T X + \lambda I)^{-1} X^T Y \tag{12}$$

此时 $(X^T X + \lambda I)$ 一定可逆。

## 从贝叶斯角度看线性回归:

和概率视角一样

$$y_i = y_i^* + \varepsilon_i \qquad Y|W \quad \sim\sim \quad N(W^T X^T, \sigma^2)$$

所以有:

$$P(Y|W) = \prod_{i=1}^{n} P(y_i, X_i|W) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(y - W^T X_i)^2}{2\sigma^2}] \tag{13}$$

" 贝叶斯" 认为 $W$ 满足一个先验分布 $W \sim\sim N(0, \sigma_0^2)$，那么也就是:

$$P(W) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp[-\frac{W^T W}{2\sigma_0^2}] \tag{14}$$

$$
\begin{aligned}
W_{MAP} &= \underset{W}{argmax} \log P(W|Y) = \underset{W}{argmax} \log P(Y|W)P(W) \\
&= \underset{W}{argmax} \log P(Y|W) + \log P(W) \\
&= \underset{W}{argmax} \sum_{i=1}^{n} \log[\frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(y-W^TX_i)^2}{2\sigma^2}]] + \log \frac{1}{\sqrt{2\pi}\sigma_0} \exp[-\frac{W^TW}{2\sigma_0^2}] \\
&= \underset{W}{argmax} \sum_{i=1}^{n} -\frac{(y-W^TX_i)^2}{2\sigma^2} - \frac{W^TW}{2\sigma_0^2} \\
&= \underset{W}{argmin} \sum_{i=1}^{n} (y-W^TX_i)^2 + \frac{\sigma^2}{\sigma_0^2} W^TW
\end{aligned}
$$

其实就是相当于 $\lambda = \frac{\sigma^2}{\sigma_0^2}$ 的岭回归正则化。

# 线性分类

## 硬分类:

$$
Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad y_i = \begin{cases} +1 & if \text{ 第 } i \text{ 个样本是正样本} \\ -1 & if \text{ 第 } i \text{ 个样本是负样本} \end{cases}
$$

**感知机:**

基于错误驱动的思想, 假设数据是线性可分的:

那么模型为:

$$
f(X_i) = sign[W^TX_i] \qquad sign(a) = \begin{cases} +1 & if\ a > 0 \\ -1 & if\ a < 0 \end{cases} \tag{15}
$$

当样本被错误分类时，有:

$$
y_iW^TX_i < 0
$$

那么损失函数为:

$$
L(W) = \sum_{X_i \in D} -y_iW^TX_i \qquad \text{其中 } D \text{ 为被错误分类的样本集合} \tag{16}
$$

所以

$$
\frac{\partial L(W)}{\partial W} = \sum_{X_i \in D} -y_iX_i \tag{17}
$$

然后利用随机梯度下降法求解直至完全分类正确为止:

$$W^{t+1} = W^t - \lambda \frac{\partial L(W)}{\partial W} \qquad \lambda \text{ 为学习率} \tag{18}$$

**Fisher 判别分析**

设:

$$\text{数据集矩阵为:} \quad X = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}_{n \times p} \qquad X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}_{p \times 1}$$

首先我们设属于正负样本的集合分别为 $X_{c1}, X_{c2}$:

$$X_{c1} = \{X_i | y_i = +1\} \qquad X_{c2} = \{X_i | y_i = -1\}$$
$$|X_{c1}| = N_1 \qquad |X_{c2}| = N_2 \qquad \text{且} \quad N_1 + N_2 = N$$

Fisher 判别分析判别分析的本质是我们把一个多维样本投影一维空间中，使投影后正类和负类的类间距离大，类内方差小：

设投影的的方向为 $W$, 且 $W$ 是一个 $p$ 维列向量, 且有 $\|W\| = 1$：

那么第 $i$ 个样本在 $W$ 方向的投影为:

$$Z_i = \|X_i\| \cos(\theta) = \|X_i\| \|W\| \cos(\theta) = W^T X_i \tag{19}$$

那么正类样本和负类样本投影在一维空间的均值为:

$$\bar{Z}_{c1} = \frac{1}{N_1} \sum_{X_i \in X_{c1}} W^T X_i \qquad \bar{Z}_{c2} = \frac{1}{N_2} \sum_{X_i \in X_{c2}} W^T X_i \tag{20}$$

投影后正负类样本之间的距离为:

$$Dis = \left( \bar{Z}_{c1} - \bar{Z}_{c2} \right)^2 \tag{21}$$

投影后正负类样本, 类内的方差为:

$$S_{c1} = \frac{1}{N_1} \sum_{X_i \in X_{c1}} \left( W^T X_i - \bar{Z}_{c1} \right)^2 \qquad S_{c2} = \frac{1}{N_2} \sum_{X_i \in X_{c2}} \left( W^T X_i - \bar{Z}_{c2} \right)^2 \tag{22}$$

定义目标函数 $J(W)$:

$$J(W) = \frac{\left( \bar{Z}_{c1} - \bar{Z}_{c2} \right)^2}{S_{c1} + S_{c2}} \tag{23}$$

我们只需要求:

$$W_{MLE} = \underset{W}{argmax} \left[ J(W) \right] \tag{24}$$

对于 $\left(\bar{Z}_{c1}-\bar{Z}_{c2}\right)^2$:

$$\begin{aligned}
\left(\bar{Z}_{c1}-\bar{Z}_{c2}\right)^2 &= \left(\frac{1}{N_1}\sum_{X_i\in C_1} W^T X_i - \frac{1}{N_2}\sum_{X_i\in C_2} W^T X_i\right)^2 \\
&= \left(W^T\left[\frac{1}{N_1}\sum_{X_i\in C_1} X_i - \frac{1}{N_2}\sum_{X_i\in C_2} X_i\right]\right)^2 \\
&= \left(W^T\left[\bar{X}_{C_1}-\bar{X}_{C_2}\right]\right)^2 \\
&= W^T\left(\bar{X}_{C_1}-\bar{X}_{C_2}\right)\left(\bar{X}_{C_1}-\bar{X}_{C_2}\right)^T W
\end{aligned}$$

对于 $S_{c1}$:

$$\begin{aligned}
S_{c1} &= \frac{1}{N_1}\sum_{i=1}^{N_1}\left(W^T X_i-\bar{Z}_{c1}\right)\left(W^T X_i-\bar{Z}_{c1}\right)^T \\
&= \frac{1}{N_1}\sum_{i=1}^{N_1} W^T\left(X_i-\bar{X}_{c1}\right)\left(X_i-\bar{X}_{c1}\right)^T W \\
&= W^T\frac{1}{N_1}\sum_{i=1}^{N_1}\left(X_i-\bar{X}_{c1}\right)\left(X_i-\bar{X}_{c1}\right)^T W
\end{aligned}$$

而 $\frac{1}{N_1}\sum_{i=1}^{N_1}\left(X_i-\bar{X}_{c1}\right)\left(X_i-\bar{X}_{c1}\right)^T$ 就是 $X_{c1}$ 的方差，我们记:

$$S_1 = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(X_i-\bar{X}_{c1}\right)\left(X_i-\bar{X}_{c1}\right)^T \tag{25}$$

所以有:

$$S_{c1} = W^T S_1 W \qquad S_1 \text{ 为 } X_{c1} \text{ 的方差} \tag{26}$$

同理有:

$$S_{c2} = W^T S_2 W \qquad S_2 \text{ 为 } X_{c2} \text{ 的方差} \tag{27}$$

所以损失函数 $J(W)$ 可以被写成:

$$J(W) = \frac{W^T\left(\bar{X}_{C_1}-\bar{X}_{C_2}\right)\left(\bar{X}_{C_1}-\bar{X}_{C_2}\right)^T W}{W^T\left(S_1+S_2\right)W} \tag{28}$$

我们引入两个符号 $X_{12}, S_{12}$, 它们均为 $p \times p$ 的矩阵:

$$X_{12} = \left(\bar{X}_{C_1} - \bar{X}_{C_2}\right)\left(\bar{X}_{C_1} - \bar{X}_{C_2}\right)^T \qquad S_{12} = S_1 + S_2$$

那么损失函数写成:

$$J(W) = \frac{W^T X_{12} W}{W^T S_{12} W} = W^T X_{12} W \left(W^T S_{12} W\right)^{-1}$$

对 $J(W)$ 求 $W$ 的导数:

$$\frac{\mathrm{d}J(W)}{\mathrm{d}W} = \frac{\mathrm{d}W^T X_{12} W}{\mathrm{d}W}\left(W^T S_{12} W\right)^{-1} + W^T X_{12} W \frac{\mathrm{d}\left(W^T S_{12} W\right)^{-1}}{\mathrm{d}W} \tag{29}$$

又因为:

$$\begin{aligned}\frac{\mathrm{d}J(W)}{\mathrm{d}W} &= \frac{\mathrm{d}W^T}{\mathrm{d}W} X_{12} W + \frac{\mathrm{d}W}{\mathrm{d}W}\left(W^T X_{12}\right)^T \\ &= X_{12} W + X_{12}^T W \\ &= 2X_{12} W\end{aligned}$$

且:

$$\begin{aligned}\frac{\mathrm{d}\left(W^T S_{12} W\right)^{-1}}{\mathrm{d}W} &= -\left(W^T S_{12} W\right)^{-2} \frac{\mathrm{d}W^T S_{12} W}{\mathrm{d}W} \\ &= -2\left(W^T S_{12} W\right)^{-2}\left(S_{12} W\right)\end{aligned}$$

所以有:

$$\frac{\mathrm{d}J(W)}{\mathrm{d}W} = 2X_{12} W\left(W^T S_{12} W\right)^{-1} - 2W^T X_{12} W\left(W^T S_{12} W\right)^{-2}\left(S_{12} W\right) \tag{30}$$

注意到 $W^T S_{12} W, W^T X_{12} W$ 是一个数，且 $W^T S_{12} W$, 不为 0, 所以我们令 $\frac{\mathrm{d}J(W)}{\mathrm{d}W} = 0$, 可以得到：

$$X_{12} W = \frac{W^T X_{12} W}{W^T S_{12} W} S_{12} W$$

注意到 $\frac{W^T X_{12} W}{W^T S_{12} W}$ 是一个数值，而且这个 Fisher 判别分析中，我们并不关心 $W$ 这个向量的取值，我们仅仅关心 $W$ 的方向上面那个等式中两端都是一个 $p$ 维列向量，$\frac{W^T X_{12} W}{W^T S_{12} W} > 0$ 不改变等式右端的方向，所以可以推断出如果等式成立，那么 $W$ 必须与 $\left(X_{12}\right)^{-1} S_{12} W$ 同向，我们将 $X_{12}$ 带入到等式中可以得到:

$$\left(\bar{X}_{C_1} - \bar{X}_{C_2}\right)\left(\bar{X}_{C_1} - \bar{X}_{C_2}\right)^T W \quad \textbf{同向} \quad S_{12} W \tag{31}$$

而 $\left(\bar{X}_{C_1} - \bar{X}_{C_2}\right)^T W$ 也只是一个不改变向量方向的数值所以如果要使 $\frac{\mathrm{d}J(W)}{\mathrm{d}W} == 0$，那么必须要让 $\bar{X}_{C_1} - \bar{X}_{C_2}$ 与 $S_{12} W$ 同向，所以我们最终得到了当 $\frac{\mathrm{d}J(W)}{\mathrm{d}W} == 0$ 时 $W$ 的取值:

$$W \quad \textbf{方向为} \quad \left(\bar{X}_{C_1} - \bar{X}_{C_2}\right)\left(S_{12}\right)^{-1} \tag{32}$$

且 $\|W\|_2 = 1$，这样我们也就求出来了投影的方向，在进行分类任务时，我们把验证集的样本通过 $W$ 投影到一维空间，然后计算这个样本投影后的值离 $\bar{Z}_{C_1}, \bar{Z}_{C_2}$ 的距离，离谁进就属于哪个类别。

## 软分类:

### 逻辑回归

我们规定数据集矩阵 $X$，投影向量 $W$ 分别为:

$$X = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}_{n\times p+1} \quad \textbf{其中：} \quad X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \\ 1 \end{bmatrix} \quad W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \\ b \end{bmatrix}$$

那么逻辑回归的原理就是通过一个投影向量 $W$ 把样本投影到一维轴上，如果投影后的值大于 0 则认为是正类，否则认为是负类。基于此我们规定如果 $X_i$ 是正类，那么 $y_i = 1$，否则 $y_i = 0$, 标签矩阵 $Y$ 被定义为:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

引入一个激活函数 $\sigma(x)$

$$\sigma(x) = \frac{1}{1+\exp(-x)} \in (0,1)$$

样本 $x_i$ 属于两个类别的概率分别为:

$$\begin{cases} p_1 = p(y_i = 1|X_i) = \sigma(W^T X_i) = \dfrac{1}{1+\exp(-W^T X_i)} & X_i \textbf{ 为正样本} \\ p_0 = p(y_i = 0|X_i) = 1 - \sigma(W^T X_i) = \dfrac{\exp(-W^T X_i)}{1+\exp(-W^T X_i)} & X_i \textbf{ 为负样本} \end{cases} \tag{33}$$

如果一个样本本 $(X_i, y_i)$ 属于正类也就是 $y_i = 1$，模型应该使 $p_1$ 尽可能的大，如果属于负样本 $y_i = 0$, 模型应该使 $p_0$ 尽可能的大，所以对于每一个样本 $(X_i, y_i)$ 逻辑回归可以建模为:

$$p(y_i|X_i) = p_1^{y_i} \times p_0^{1-y_i} \tag{34}$$

所以我们只需要最大化：

$$P(Y|X) = \prod_{i=1}^{n} p(y_i|X_i) = \prod_{i=1}^{n} p_1^{y_i} \times p_0^{1-y_i}$$

我们使用极大释然估计法求 $W_{MLE}$:

$$\begin{aligned}W_{MLE} &= \underset{W}{argmax} \log P(Y|X) \\ &= \underset{W}{argmax} \sum_{i=1}^{n} \left[y_i \log p_1 + (1-y_i)\log p_0\right] \\ &= \underset{W}{argmax} \sum_{i=1}^{n} \left[y_i W^T X_i - \log\left[\exp(W^T X_i)+1\right]\right] \\ &= \underset{W}{argmin} \sum_{i=1}^{n} \left[\log\left[\exp(W^T X_i)+1\right] - y_i W^T X_i\right]\end{aligned}$$

也就是我们求一个 $W$ 使 $\sum_{i=1}^{n}\left[\log\left[\exp(W^T X_i)+1\right] - y_i W^T X_i\right]$ 最小化即可:
因为：

$$\frac{\partial\left[\log\left[\exp(W^T X_i)+1\right] - y_i W^T X_i\right]}{\partial W} = -y_i X_i + \frac{\exp(W^T X_i)}{\exp(W^T X_i)+1} X_i = \left[\frac{\exp(W^T X_i)}{\exp(W^T X_i)+1} - y_i\right] X_i \tag{35}$$

所以:

$$\nabla W = \frac{\partial \sum_{i=1}^{n}\left[\log\left[\exp(W^T X_i)+1\right] - y_i W^T X_i\right]}{\partial W} = \sum_{i=1}^{n}\left[\frac{\exp(W^T X_i)}{\exp(W^T X_i)+1} - y_i\right] X_i \tag{36}$$

最后利用梯度下降法求解即可:

$$W^{t+1} = W^t - \lambda \nabla W \tag{37}$$

**高斯判别分析**

高斯判别分析是一种概率生成模型，上面的逻辑回归是一种概率判别模型，两种的区别是概率判别模型直接求 $p(y=1|x)$，然后令 $p(y=0|x) = 1-p(y=1|x)$，判别模型直接求样本属于某一类的概率。而生成模型并不求具体的概率，他只是比较 $p(y=1|x), p(y=0|x)$ 谁大。依据贝叶斯公式，我们知道:

$$\begin{cases} p(y=1|x) = \dfrac{p(y=1,x)}{p(x)} = \dfrac{p(y=1)\times p(x|y=1)}{p(x)} \\ \\ p(y=0|x) = \dfrac{p(y=0,x)}{p(x)} = \dfrac{p(y=0)\times p(x|y=0)}{p(x)} \end{cases} \tag{38}$$

所以如果我们要比较 $p(y=1|x), p(y=0|x)$，只需要去比较 $p(y=1)\times p(x|y=1), p(y=0)\times p(x|y=0)$ 谁大即可。高斯判别分析假设 $y$ 服从伯努利分布，$y \sim B(\Theta)$, 可以写成:

$$p(y_i) = \Theta^{y_i}\cdot(1-\Theta)^{1-y_i}$$

并且假设:

$$\begin{aligned} x|y=1 &\quad \sim \quad N(u_1,\Sigma) \\ x|y=0 &\quad \sim \quad N(u_2,\Sigma) \end{aligned}$$

分别记 $x|y=1, x|y=0$ 的概率密度函数为:$f_1, f_2$, 那么 $p(x_i|y_i)$ 可以表示成:

$$p(x_i|y_i) = {f_1}^{y_i} \cdot {f_2}^{1-y_i} \tag{39}$$

我们要估计出一组参数 $\Theta, u_1, u_2, \Sigma$, 使 $p(Y|X)$ 最大，由于 $p(X)$ 在数据集确定时是一个定值，所以只需要使 $p(Y,X)$ 最大即可
我们的 $log$ 释然函数为:

$$\begin{aligned} L(\Theta,u_1,u_2,\Sigma) &= \log p(Y,X) \\ &= \log \prod_{i=1}^{n} p(y_i,x_i) \\ &= \sum_{i=1}^{n} \left[\log p(y_i) + \log p(x_i|y_i)\right] \\ &= \sum_{i=1}^{n} \left[\log\left[\Theta^{y_i}\cdot(1-\Theta)^{1-y_i}\right] + \log\left[{f_1}^{y_i}\cdot{f_2}^{1-y_i}\right]\right] \\ &= \sum_{i=1}^{n} \left[y_i\log\Theta + (1-y_i)\log\left[1-\Theta\right] + y_i\log f_1 + (1-y_i)\log f_2\right] \end{aligned}$$

所以，所谓我们需要估计的 $\Theta, u_1, u_2, \Sigma$ 可以写成:

$$(\Theta,u_1,u_2,\Sigma) = \underset{\Theta,u_1,u_2,\Sigma}{argmax} \quad L(\Theta,u_1,u_2,\Sigma) \tag{40}$$

首先我们估计 $\Theta$：

$$\frac{\partial L}{\partial \Theta} = \sum_{i=1}^{n}\left[\frac{y_i}{\Theta} - \frac{1-y_i}{1-\Theta}\right]$$

令 $\frac{\partial L}{\partial \Theta}=0$，求得:

$$\Theta_{MLE} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1-y_i) + \sum_{i=1}^{n} y_i} \tag{41}$$

我们记数据集中有正类 ($y_i=1$)，负类 ($y_i=0$) 的样本个数分别为:$N_1, N_2$，那么显然:

$$\Theta_{MLE} = \frac{N_1}{N_1+N_2} \tag{42}$$

我们将 $f_1, f_2$ 展开:

$$\begin{cases} f_1 = \dfrac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}} \exp\left[-\dfrac{1}{2}\,(x_i-u_1)^T\,\Sigma^{-1}\,(x_i-u_1)\right] \\ f_2 = \dfrac{1}{(2\pi)^{\frac{p}{2}}\,|\Sigma|^{\frac{1}{2}}} \exp\left[-\dfrac{1}{2}\,(x_i-u_2)^T\,\Sigma^{-1}\,(x_i-u_2)\right] \end{cases} \tag{43}$$

先求 $u_{1_{MEL}}$：

$$\begin{aligned}
\frac{\partial L}{\partial u_1} &= \frac{\partial \sum_{i=1}^{n} \frac{y_i}{2}\left[(x_i-u_1)^T\Sigma^{-1}(x_i-u_1)\right]}{\partial u_1} \\
&= \frac{\partial \sum_{i=1}^{n} \frac{y_i}{2}\left[u_1^T\Sigma^{-1}u_1 - 2u_1^T\Sigma^{-1}x_i\right]}{\partial u_1} \\
&= \sum_{i=1}^{n} y_i\left(\Sigma^{-1}u_1 - \Sigma^{-1}x_i\right)
\end{aligned}$$

注意到 $\Sigma$ 是可逆的矩阵，那么 $\Sigma$ 是列满秩矩阵，也就是可消去，如果我们要让 $\frac{\partial L}{\partial u_1}=0$，也就是要让：

$$u_{1_{MEL}} = \frac{1}{N_1}\sum_{x_i\in D_1} x_i \quad \textbf{式中 } D_1 \textbf{ 表示属于正类 } (y_i=1) \textbf{ 的所有样本集合} \tag{44}$$

同理可得:

$$u_{2_{MEL}} = \frac{1}{N_2}\sum_{x_i\in D_2} x_i \quad \textbf{式中 } D_2 \textbf{ 表示属于负类 } (y_i=0) \textbf{ 的所有样本集合} \tag{45}$$

最后我们求解 $\Sigma_{MLE}$：

$$\frac{\partial L}{\partial \Sigma} = -\frac{1}{2}\left[(N_1+N_2)\frac{\partial \log|\Sigma|}{\partial \Sigma} + \sum_{x_i\in D_1}\frac{\partial (x_i-u_1)^T\Sigma^{-1}(x_i-u_1)}{\partial \Sigma} + \sum_{x_i\in D_2}\frac{\partial (x_i-u_2)^T\Sigma^{-1}(x_i-u_2)}{\partial \Sigma}\right]$$

我们观察到其实 $(x_i-u_1)^T\Sigma^{-1}(x_i-u_1), (x_i-u_2)^T\Sigma^{-1}(x_i-u_2)$ 其实是一个数，那么一个数的迹等于它自己，所以有:

$$\begin{cases}
(x_i-u_1)^T\Sigma^{-1}(x_i-u_1) = tr((x_i-u_1)^T\Sigma^{-1}(x_i-u_1)) \\
(x_i-u_2)^T\Sigma^{-1}(x_i-u_2) = tr((x_i-u_2)^T\Sigma^{-1}(x_i-u_2))
\end{cases} \tag{46}$$

对于迹有运算 $tr(AB)=tr(BA)$ 成立: 所以有:

$$\begin{cases}
tr((x_i-u_1)^T\Sigma^{-1}(x_i-u_1)) = tr((x_i-u_1)(x_i-u_1)^T\Sigma^{-1}) \\
tr((x_i-u_2)^T\Sigma^{-1}(x_i-u_2)) = tr((x_i-u_2)(x_i-u_2)^T\Sigma^{-1})
\end{cases} \tag{47}$$

所以：

$$\begin{aligned}
\sum_{x_i\in D_1}\frac{\partial (x_i-u_1)^T\Sigma^{-1}(x_i-u_1)}{\partial \Sigma} &= \sum_{x_i\in D_1}\frac{\partial\left[tr((x_i-u_1)(x_i-u_1)^T\Sigma^{-1})\right]}{\partial \Sigma} \\
&= \frac{\partial\left[\sum_{x_i\in D_1} tr((x_i-u_1)(x_i-u_1)^T\Sigma^{-1})\right]}{\partial \Sigma} \\
&= \frac{\partial\left[tr(\left[\sum_{x_i\in D_1}(x_i-u_1)(x_i-u_1)^T\right]\Sigma^{-1})\right]}{\partial \Sigma}
\end{aligned}$$

$\dfrac{1}{N_1}\sum_{x_i\in D_1}(x_i-u_1)(x_i-u_1)^T$ 就是所有正类样本的协方差矩阵, 记所有正类样本的协方差矩阵为 $S_1$：所以有：

$$\sum_{x_i\in D_1}\frac{\partial(x_i-u_1)^T\Sigma^{-1}(x_i-u_1)}{\partial\Sigma}=N_1\frac{\partial\left[tr(S_1\Sigma^{-1})\right]}{\partial\Sigma} \tag{48}$$

同理:

$$\sum_{x_i\in D_2}\frac{\partial(x_i-u_2)^T\Sigma^{-1}(x_i-u_2)}{\partial\Sigma}=N_2\frac{\partial\left[tr(S_2\Sigma^{-1})\right]}{\partial\Sigma} \tag{49}$$

式中 $S_2$ 为所有负样本的协方差矩阵。
下面计算 $\frac{\partial\left[tr(S_1\Sigma^{-1})\right]}{\partial\Sigma}$：

$$\begin{aligned}\frac{\partial\left[tr(S_1\Sigma^{-1})\right]}{\partial\Sigma}&=\frac{\partial\left[tr(\Sigma^{-1}S_1)\right]}{\partial\Sigma}\\&=\frac{\partial\left[tr(\Sigma^{-1}S_1)\right]}{\partial\Sigma^{-1}}\frac{\partial\Sigma^{-1}}{\partial\Sigma}\\&\textbf{因为有公式:}\frac{\partial tr(AB)}{\partial A}=B^T\\&=-S_1^T\Sigma^{-2}\\&=-S_1\Sigma^{-2}\end{aligned}$$

同理可得:

$$\frac{\partial\left[tr(S_2\Sigma^{-1})\right]}{\partial\Sigma}=-S_2\Sigma^{-2}$$

再求:$\frac{\partial\log|\Sigma|}{\partial\Sigma}$:

$$\frac{\partial\log|\Sigma|}{\partial\Sigma}=\frac{1}{|\Sigma|}\frac{\partial|\Sigma|}{\partial\Sigma}$$

$$\textbf{因为有:}\frac{\partial|A|}{\partial A}=|A|A^{-1}$$

$$\begin{aligned}&=\frac{1}{|\Sigma|}|\Sigma|\Sigma^{-1}\\&=\Sigma^{-1}\end{aligned}$$

所以 $\frac{\partial L}{\partial\Sigma}$ 可以写成:

$$\frac{\partial L}{\partial\Sigma}=-\frac{1}{2}\left[(N_1+N_2)\Sigma^{-1}-N_1S_1\Sigma^{-2}-N_2S_2\Sigma^{-2}\right] \tag{50}$$

令 $\frac{\partial L}{\partial \Sigma} == 0$，可以得到:

$$\Sigma_{MLE} = \frac{N_1}{N_1+N_2}S_1 + \frac{N_2}{N_1+N_2}S_2 \tag{51}$$

当要预测样本 $x_i$ 时，我们利用 $\Theta_{MLE}, u_{1_{MLE}}, u_{2_{MLE}}, \Sigma_{MLE}$，分别计算 $p(y=1)\times p(x|y=1)$ 和 $p(y=0)\times p(x|y=0)$，谁大 $x_j$ 就属于哪一类。

**朴素贝叶斯分类器**

我们假设是一个二分类的问题，对于样本 $x_i$，$x_i=[x_{i1}, x_{i2}, \ldots x_{ip}]^T$, 也就是这个样本共有 $p$ 个特征，这个样本所属的类别 $y_i=1$ 或者 $y_i=0$，那么朴素贝叶斯分类器认为各个样本之间相互独立，并且样本之间的特征也是相互独立的。

当给出一个样本 $x_i$ 时，与高斯判别分析一样，我们去比较 $p(y_i=1|x_i), p(y_i=0|x_i)$ 谁大，因为:

$$\begin{cases} p(y_i=1|x_i) = \dfrac{p(y_i=1,x_i)}{p(x_i)} = \dfrac{p(y_i=1)\times p(x_i|y_i=1)}{p(x_i)} \\ \\ p(y_i=0|x_i) = \dfrac{p(y_i=0,x_i)}{p(x_i)} = \dfrac{p(y_i=0)\times p(x_i|y_i=0)}{p(x_i)} \end{cases} \tag{52}$$

我们只需要比较这两个 $p(y_i=1)\times p(x_i|y_i=1)$ 和 $p(y_i=0)\times p(x_i|y_i=0)$。

我们通常会假设 $y$ 服从伯努利分布，即 $y \sim B(\Theta)$, 我们可以利用极大释然估计法求出 $\Theta_{MLE} = \dfrac{N_1}{N_1+N_2}$,$N_1, N_2$ 分别为样本数据集中属于正类和负类的数量, 这样我们就可以知道:

$$\begin{cases} p(y_i=1) = \Theta_{MLE} \\ p(y_i=0) = 1-\Theta_{MLE} \end{cases} \tag{53}$$

由于假设样本之间的特征是独立的，所以 $p(x_i|y_i=1)$ 和 $p(x_i|y_i=0)$ 可以展开成：

$$\begin{cases} p(x_i|y_i=1) = \prod_{j=1}^{p} p(x_{ij}|y_i=1) \\ p(x_i|y_i=0) = \prod_{j=1}^{p} p(x_{ij}|y_i=0) \end{cases} \tag{54}$$

我们记 $Z_j \quad j=1,2,\ldots,p$ 所有样本的第 $j$ 个特征，那么我们通常认为:

$$\begin{cases} Z_j|y=1 \quad \sim\sim \quad N(u,\sigma) & \text{如果 } Z_j \text{ 是连续型随机变量} \\ Z_j|y=1 \quad \sim\sim \quad Categorical(\alpha_1,\alpha_2,\ldots,\alpha_m) & \text{如果 } Z_j \text{ 是离散型随机变量} \end{cases} \tag{55}$$

式中 $\alpha_i \quad i=1,\ldots,m$ 为 $Z_j$ 属于第 $i$ 个类别的概率，且 $\sum_{i=1}^{m}\alpha_i=1$, 我们对所有的特征 $Z_j|y=1 \quad j=1,2,\ldots,p$ 都建模，利用极大释然估计法得到 $Z_j|y=1$ 对应的分布，那么也就是得到了概率 $p(Z_j|y=1)$，那么 $p(x_i|y_i=1)$ 可以写成:

$$p(x_i|y_i=1) = \prod_{j=1}^{p} p(x_{ij}|y_i=1) = \prod_{j=1}^{p} p(Z_j=x_{ij}|y_i=1) \tag{56}$$

**其实就是把 $x_{ij}$ 的值直接作为 $Z_j$ 的取值往分布函数里面丢得到概率**

同理我们可以利用极大释然估计法得到 $Z_j|y=0$ 对应的分布, 然后计算 $p(x_i|y_i=0)$

$$p(x_i|y_i=0)=\prod_{j=1}^{p}p(x_{ij}|y_i=0)=\prod_{j=1}^{p}p(Z_j=x_{ij}|y_i=0) \tag{57}$$

所以有:

$$\begin{cases} p(y_i=1)\times p(x_i|y_i=1)= & \Theta_{MLE} & \prod_{j=1}^{p}p(Z_j=x_{ij}|y_i=1) \\ p(y_i=0)\times p(x_i|y_i=0)= & (1-\Theta_{MLE}) & \prod_{j=1}^{p}p(Z_j=x_{ij}|y_i=0) \end{cases} \tag{58}$$

谁大 $x_i$ 就属于那个类别