高斯判别分析

杨航锋

高斯判别分析属于机器学习算法中的分类算法，不妨假设样本数据为两种类别，它的大致思想是通过两个先验假设：一是样本数据的类别 y 在给定的情况下服从伯努利分布，二是不同类别中的样本数据分别服从多元高斯分布。首先估计出先验概率以及多元高斯分布的均值和协方差矩阵，然后再由贝叶斯公式求出一个新样本分别属于两类别的概率，预测结果取概率值大者。本文推导高斯判别分析算法的流程是，首先简单的推导出多元高斯分布、其次提出高斯判别分析算法的假设函数、然后构造损失函数、最后求解损失函数得出假设函数中的参数值。更多文章见GitHub地址

多元高斯分布

高中时期就学习过正态分布 $X \sim N(\mu, \sigma^2)$ ，它的概率密度函数 $\varphi(x)$ 为

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

假设 X_i 之间相互独立且$X_i \sim N(\mu_i, \sigma_i^2)$ $i = 1, 2, \cdots, n$ 令 $x = [x_1, x_2, \cdots, x_n]^T$; $\mu = [\mu_1, \mu_2, \cdots, \mu_n]^T$; $\sigma = [\sigma_1, \sigma_2, \cdots, \sigma_n]^T$ 则多元高斯分布的密度函数可以表示为

$$\begin{aligned}
\varphi(x) &= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_1-\mu_1}{\sigma_1})^2} \cdot \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_2-\mu_2}{\sigma_2})^2} \cdots \frac{1}{\sigma_n\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_n-\mu_n}{\sigma_n})^2} \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} \prod\limits_{i=1}^{n} \sigma_i} e^{-\frac{1}{2}\sum\limits_{j=1}^{n}(\frac{x_j-\mu_j}{\sigma_j})^2}
\end{aligned}$$

对于上述指数部分$\xi^2(x, \mu, \sigma) = \sum\limits_{j=1}^{n}(\frac{x_j-\mu_j}{\sigma_j})^2$可以表示为矩阵乘法的形式(联想下线性代数二次型的矩阵表示)

$$\begin{aligned}
\xi^2(x, \mu, \sigma) &= \sum_{j=1}^{n}(\frac{x_j-\mu_j}{\sigma_j})^2 \\
&= \sum_{j=1}^{n}(x_j-\mu_j)(x_j-\mu_j)(\frac{1}{\sigma_j})^2 \\
&= [\, x_1-\mu_1, x_2-\mu_2, \cdots, x_n-\mu_n \,] \begin{bmatrix} \frac{1}{\sigma^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma^2} \end{bmatrix} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_n-\mu_n \end{bmatrix} \\
&= (x-\mu)^T \Sigma^{-1} (x-\mu)
\end{aligned}$$

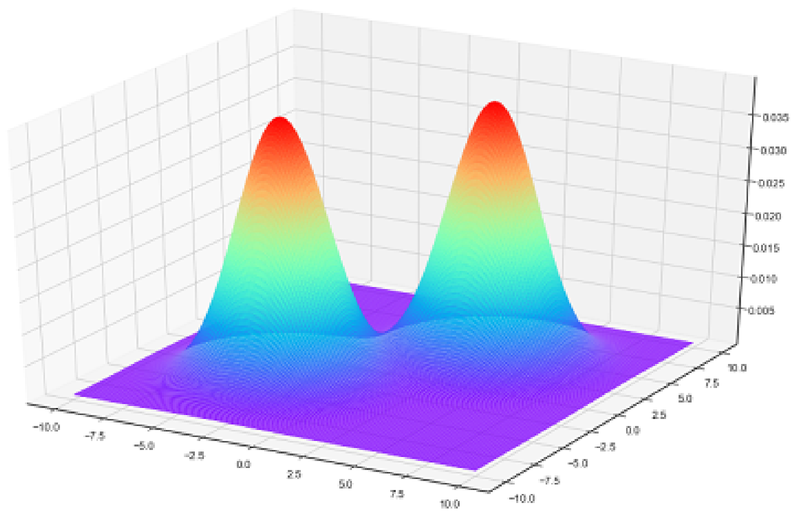
其中根据 X_i 相互独立知

$$
\begin{aligned}
\Sigma &= E\{(X-EX)(X-EX)^T\} \\
&= \begin{bmatrix} var(X_1) & cov(X_1,X_2) & \cdots & cov(X_1,X_n) \\ cov(X_2,X_1) & var(X_2) & \cdots & cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n,X_1) & cov(X_n,X_2) & \cdots & var(X_n) \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}
\end{aligned}
$$

所以多元高斯分布$X \sim N(\mu, \Sigma)$的密度函数为

$$
\varphi(x,\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}\,|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}
$$

双二维独立高斯分布的图像为



高斯判别分析模型的假设函数

不妨假设含有m个样本数据$(x^{(1)},y^{(1)})$、$(x^{(2)},y^{(2)})$、$\cdots$、$(x^{(m)},y^{(m)})$，$y^{(i)} \in \{0,1\}$。当需要构建高斯判别分析模型 $p(x|y)$ 时，样本数据需满足以下给出的先验概率分布

$$
\begin{aligned}
y &\sim Bernoulli(\phi) \\
x|y=0 &\sim \mathcal{N}(\mu_0,\Sigma) \\
x|y=1 &\sim \mathcal{N}(\mu_1,\Sigma)
\end{aligned}
$$

写成分布函数的形式即

$$
\begin{aligned}
p(y) &= \phi^y(1-\phi)^{1-y} \\
p(x|y=0) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)} \\
p(x|y=1) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}
\end{aligned}
$$

上述模型中的未知参数为 ϕ、Σ、μ_0和μ_1 ，假设函数为 $p(y|x)=\dfrac{p(x|y)p(y)}{p(x)}$ 分别计算 $p(x|y=0)$和$p(x|y=1)$ 的概率，概率大者为样本数据所属类别。

高斯判别分析模型的损失函数

已知样本数据含有参数的概率分布，根据统计学的最大似然估计可以推导高斯判别分析模型的损失函数为

$$
\begin{aligned}
\mathcal{L}(\phi,\mu_0,\mu_1,\Sigma) &= \log\prod_{i=1}^{m} p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma) \\
&= \log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi) \\
&= \sum_{i=1}^{m}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)+\sum_{i=1}^{m}\log p(y^{(i)};\phi) \\
&= \sum_{i=1}^{m}\log\Big(p(x^{(i)}|y^{(i)}=1;\mu_1,\Sigma)^{y^{(i)}}\cdot p(x^{(i)}|y^{(i)}=0;\mu_0,\Sigma)^{1-y^{(i)}}\Big)+\sum_{i=1}^{m}\log p(y^{(i)};\phi) \\
&= \sum_{i=1}^{m} y^{(i)}\log p(x^{(i)}|y^{(i)}=1;\mu_1,\Sigma)+\sum_{i=1}^{m}(1-y^{(i)})\log p(x^{(i)}|y^{(i)}=0;\mu_0,\Sigma)+\sum_{i=1}^{m}\log p(y^{(i)};\phi)
\end{aligned}
$$

求解上述损失函数

对最大似然函数 $\mathcal{L}(\phi,\mu_0,\mu_1,\Sigma)$ 求偏导并令其相应偏导数为零即可求出参数

$$
\begin{aligned}
\nabla_\phi\mathcal{L}(\phi,\mu_0,\mu_1,\Sigma) &= \nabla_\phi\sum_{i=1}^{m}\log p(y^{(i)};\phi) \\
&= \nabla_\phi\sum_{i=1}^{m}\log\phi^{y^{(i)}}(1-\phi)^{(1-y^{(i)})} \\
&= \nabla_\phi\sum_{i=1}^{m}\{y^{(i)}\log\phi+(1-y^{(i)})\log(1-\phi)\} \\
&= \sum_{i=1}^{m}\{y^{(i)}\cdot\frac{1}{\phi}-(1-y^{(i)})\cdot\frac{1}{1-\phi}\} \\
&= \sum_{i=1}^{m}\{I(y^{(i)}=1)\cdot\frac{1}{\phi}-I(y^{(i)}=0)\cdot\frac{1}{1-\phi}\}
\end{aligned}
$$

其中 $I(x)$ 为示性函数，当 x 为真时 $I(x)$ 的值为 1 ,当 x 为假时 $I(x)$ 的值为 0 ，令 $\nabla_\phi\mathcal{L}(\phi,\mu_0,\mu_1,\Sigma)=0$

$$\phi = \frac{\sum\limits_{i=1}^{m} I(y^{(i)} = 1)}{\sum\limits_{i=1}^{m} \{I(y^{(i)} = 1) + I(y^{(i)} = 0)\}} = \frac{\sum\limits_{i=1}^{m} I(y^{(i)} = 1)}{m}$$

同样地对 μ_0 求偏导可得

$$\begin{aligned}
\nabla_{\mu_0} \mathcal{L}(\phi, \mu_0, \mu_1, \Sigma) &= \nabla_{\mu_0} \sum_{i=1}^{m} (1 - y^{(i)}) \log p(x^{(i)} | y^{(i)} = 0; \mu_0, \Sigma) \\
&= \nabla_{\mu_0} \sum_{i=1}^{m} (1 - y^{(i)}) \cdot \log \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x^{(i)} - \mu_0)^T \Sigma^{-1} (x^{(i)} - \mu_0)} \\
&= \sum_{i=1}^{m} (1 - y^{(i)}) \Sigma^{-1} (x^{(i)} - \mu_0) \\
&= \sum_{i=1}^{m} (I(y^{(i)}) = 0) \Sigma^{-1} (x^{(i)} - \mu_0)
\end{aligned}$$

令 $\nabla_{\mu_0} \mathcal{L}(\phi, \mu_0, \mu_1, \Sigma) = 0$

$$\mu_0 = \frac{\sum\limits_{i=1}^{m} I(y^{(i)} = 0) x^{(i)}}{\sum\limits_{i=1}^{m} I(y^{(i)} = 0)}$$

根据对称性可知

$$\mu_1 = \frac{\sum\limits_{i=1}^{m} I(y^{(i)} = 1) x^{(i)}}{\sum\limits_{i=1}^{m} I(y^{(i)} = 1)}$$

最后对 Σ 求偏导可得

$$
\begin{aligned}
\nabla_{\Sigma}\mathcal{L}(\phi,\mu_0,\mu_1,\Sigma) &= \nabla_{\Sigma}\left(\sum_{i=1}^{m} y^{(i)}\log p(x^{(i)}|y^{(i)}=1;\mu_1,\Sigma)+\sum_{i=1}^{m}(1-y^{(i)})\log p(x^{(i)}|y^{(i)}=0;\mu_0,\Sigma)\right) \\
&= \nabla_{\Sigma}\left(\sum_{i=1}^{m} y^{(i)}\cdot\log\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)}+\right. \\
&\left.\sum_{i=1}^{m}(1-y^{(i)})\cdot\log\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0)}\right) \\
&= \nabla_{\Sigma}\left(\sum_{i=1}^{m}\log\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}-\frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})\right) \\
&= \nabla_{\Sigma}\left(\sum_{i=1}^{m}(-\frac{n}{2}\log 2\pi-\frac{1}{2}\log|\Sigma|)-\frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})\right) \\
&= -\frac{1}{2}\sum_{i=1}^{m}\frac{1}{|\Sigma|}|\Sigma|\Sigma^{-1}-\frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\cdot\nabla_{\Sigma}\Sigma^{-1} \\
&= -\frac{m}{2}\Sigma^{-1}-\frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T(-\Sigma^{-2})
\end{aligned}
$$

其中直接利用了下面的结论

$$
\begin{aligned}
\nabla_{\Sigma}|\Sigma| &= |\Sigma|\Sigma^{-1} \\
\nabla_{\Sigma}\Sigma^{-1} &= -\Sigma^{-2}
\end{aligned}
$$

$\nabla_{\Sigma}\mathcal{L}(\phi,\mu_0,\mu_1,\Sigma)=0$ 从而推导出

$$
\Sigma=\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T
$$

总结

上面的推导看上去有些复杂，但求解出的结果却是非常简洁。通过上述公式，所有的未知参数都已经估计出来了，当需要判断一个新样本 $x^{(i)}$ 时，可分别求出 $p(y^{(i)}=0|x^{(i)})$ 和 $p(y^{(i)}=1|x^{(i)})$ ，取概率更大的那个类。