机器学习笔记–高斯判别模型 *(GDA)*

空修菜

1 高斯判别分析（Gaussian discriminant analysis，GDA）

1.1 一些理论

DGA 模型

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

具体写出来就是

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

在这个模型中，参数是 $\phi, \mu_0, \mu_1, \Sigma$, 模型的对数可能性（log-likelihood）为 ℓ

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}, \phi, \mu_0, \mu_1, \Sigma) \\ &= \log \prod_{i=1}^{n} p(x^{(i)}|y^{(i)}, \phi, \mu_0, \mu_1, \Sigma) p(y^{(i)}, \phi) \end{aligned}$$

对 ℓ 的参数求导数，就可以得到每一个参数的表达式。为了计算方便，分别令

$$\begin{aligned}\alpha &= 1/((2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}})\\ \eta_0 &= -\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\\ \eta_1 &= -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\end{aligned}$$

所以就可以得到

$$\begin{aligned}\ell(\phi,\mu_0,\mu_1,\Sigma) &= \log\prod_{i=1}^{n}\{\alpha\exp\eta_0\}^{1-y^{(i)}}\{\alpha\exp\eta_1\}^{y^{(i)}}\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}}\\ &= n\log\alpha+\sum_{i=1}^{n}(1-y^{(i)})\eta_0+\sum_{i=1}^{n}y^{(i)}\eta_1\\ &+\log\phi\sum_{i=1}^{n}y^{(i)}+\log(1-\phi)\sum_{i=1}^{n}(1-y^{(i)})\end{aligned}$$

求导，并令求导后的结果为 0，则有

$$0=\frac{\partial\ell}{\partial\phi}=\frac{1}{\phi}\sum_{i=1}^{n}y^{(i)}-\frac{1}{1-\phi}\sum_{i=1}^{n}(1-y^{(i)}).$$

整理上式则有

$$\phi=\frac{1}{n}\sum_{i=1}^{n}y^{(i)}.$$

在计算 $\frac{\partial\ell}{\partial\mu_0}$ 时，由链式法则我们可以得到如下表达式

$$\frac{\partial\ell}{\partial\mu_0}=\frac{\partial\ell}{\partial\eta_0}\frac{\partial\eta_0}{\partial\mu_0}=\sum_{i=1}^{n}(1-y^{(i)})\frac{\partial\eta_0}{\partial\mu_0}$$

注意到

$$\frac{\partial\eta_0}{\partial\mu_0}=\Big(\frac{\partial\eta_0}{\partial\mu_{01}},\frac{\partial\eta_0}{\partial\mu_{02}},...,\frac{\partial\eta_0}{\partial\mu_{0n}}\Big)^T$$

所以需要先计算 $\frac{\partial\eta_0}{\partial\mu_{0t}}$,$1\le t\le n$. 将 $x^{(i)}-\mu_0$ 的分量写出

$$x^{(i)}-\mu_0=(x_1^{(i)}-\mu_{01},x_2^{(i)}-\mu_{02},...,x_n^{(i)}-\mu_{0n})^T,$$

并假设

$$\Sigma^{-1} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

则由 η_0 的定义可以得到

$$\begin{aligned} \eta_0 &= -\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) \\ &= (x_1^{(i)}-\mu_{01},...,x_n^{(i)}-\mu_{0n}) \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \dots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \begin{pmatrix} x_1^{(i)}-\mu_{01} \\ \vdots \\ x_n^{(i)}-\mu_{0n} \end{pmatrix} \\ &= -\frac{1}{2}\sum_{s=1}^{n}\sum_{j=1}^{n} a_{js}(x_s^{(i)}-\mu_{0s})(x_j^{(i)}-\mu_{0j}) \end{aligned}$$

由上式子可知，在对 μ_{0t} 求导的时候要分四种情况 $s=t, s\neq t, j=t, j\neq t$。所以

$$\begin{aligned} \eta_0 &= -\frac{1}{2}\left\{\sum_{s\neq t}^{n}\sum_{j=1}^{n} a_{js}(x_s^{(i)}-\mu_{0s})(x_j^{(i)}-\mu_{0j}) + \sum_{j=1}^{n} a_{jt}(x_t^{(i)}-\mu_{0t})(x_j^{(i)}-\mu_{0j})\right\} \\ &= -\frac{1}{2}\left\{\sum_{s\neq t}^{n}\sum_{j\neq t}^{n} a_{js}(x_s^{(i)}-\mu_{0s})(x_j^{(i)}-\mu_{0j}) + a_{ts}(x_s^{(i)}-\mu_{0s})(x_t^{(i)}-\mu_{0t})\right\} \\ &\quad -\frac{1}{2}\left\{\sum_{j\neq t}^{n} a_{jt}(x_t^{(i)}-\mu_{0t})(x_j^{(i)}-\mu_{0j}) + a_{tt}(x_t^{(i)}-\mu_{0t})(x_t^{(i)}-\mu_{0t})\right\} \end{aligned}$$

由于 Σ^{-1} 是对称的，即 $a_{ts}=a_{st}$，且是可逆的，所以

$$\begin{aligned} \frac{\partial\eta_0}{\partial\mu_{0t}} &= -\frac{1}{2}\left\{\sum_{s\neq t}^{n} -a_{ts}(x_s^{(i)}-\mu_{0s}) + \sum_{j\neq t}^{n} -a_{jt}(x_t^{(i)}-\mu_{0t}) - 2a_{tt}(x_t^{(i)}-\mu_{0t})\right\} \\ &= \frac{1}{2}\left\{\sum_{j\neq t}^{n} 2a_{jt}(x_j^{(i)}-\mu_{0j}) + 2a_{tt}(x_t^{(i)}-\mu_{0t})\right\} \\ &= \sum_{j=1}^{n} a_{jt}(x_j^{(i)}-\mu_{0j}) = \sum_{j=1}^{n} a_{tj}(x_j^{(i)}-\mu_{0j}) \end{aligned}$$

由上式可知, 它是由 Σ^{-1} 的第 t 行与列向量 $x^{(i)} - \mu_0$ 相乘而得。将 t 从 1 到 n 取遍，就得到了 $\frac{\partial \eta_0}{\partial \mu_0}$,

$$\frac{\partial \eta_0}{\partial \mu_0} = \Sigma^{-1}(x^{(i)} - \mu_0).$$

所以根据梯度下降法以及矩阵乘法规则可得

$$0 = \frac{\partial \ell}{\partial \mu_0} = \sum_{i=1}^{n}(1 - y^{(i)})\Sigma^{-1}(x^{(i)} - \mu_0) = \Sigma^{-1}\sum_{i=1}^{n}(1 - y^{(i)})(x^{(i)} - \mu_0),$$

由于 Σ 是可逆的, 将上式左右同乘 Σ 可得

$$0 = \sum_{j=1}^{n}(1 - y^{(i)})(x^{(i)} - \mu_0),$$

所以我们就得到了 μ_0 的表达式

$$\mu_0 = \frac{\sum_{j=1}^{n}(1 - y^{(i)})x^{(i)}}{\sum_{j=1}^{n}(1 - y^{(i)})}.$$

通过同样的方法，我们可以得到

$$\mu_1 = \frac{\sum_{i=1}^{n} y^{(i)}x^{(i)}}{\sum_{i=1}^{n} y^{(i)}}$$

利用 indicate function 我们可以得到如下形式的表达式:

$$\begin{aligned}
\phi &= \frac{1}{n}\sum_{i=1}^{n} 1\{y^{(i)} = 1\} \\
\mu_0 &= \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}} \\
\mu_1 &= \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}} \\
\Sigma &= \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T
\end{aligned}$$

1. 在得到上述参数后，根据贝叶斯公式计算每一类的概率，即

$$p(y = 0 \mid x) = \frac{p(x \wedge y = 0)}{p(x)} = \frac{p(y = 0)p(x \mid y = 0)}{p(y = 0)p(x \mid y = 0) + p(y = 1)p(x \mid y = 1)},$$

$$p(y=1 \mid x)=\frac{p(x \wedge y=1)}{p(x)}=\frac{p(y=1)p(x \mid y=1)}{p(y=0)p(x \mid y=0)+p(y=1)p(x \mid y=1)}.$$

哪一个的概率值大，x 就是哪一类。

2. 上面得到的是一个二维的概率向量，实际上，可以类似于 softmax 模型，将高斯判别模型推广到 n 维。

3. 高斯判别模型实际上是罗辑回归模型的一种。两者的共同假设是：y 都服从伯努利分布。高斯判别分析的假设很强：样本点服从高斯分布，且两个不同类型的样本点的协方差矩阵相同，均值向量不同。所以，罗辑回归往往比高斯分布的效果更好一些。