高斯判别模型_阳畅_20160401

一、分类原理
高斯判别模型属于生成模型，先建立p(x|y)的高斯判别模型，再根据贝叶斯公式求出p(y|x)。

二、模型说明
如果输入特征x是连续型随机变量，且服从混合正态分布（每个特征符合正态分布），那么可以使用高斯判别分析模型来确定p(x|y)，即求出模型中参数的极大似然估计。
假设为二分类问题，那么将新的样本x带入进建立好的模型中，计算出p(y=1|x)、p(y=0|x)，选取概率更大的结果为正确的分类。

三、数学实现
为了简化模型，假设特征值为二分类，分类结果服从0-1分布。（如果为多分类，分类结果就服从二项分布）

模型基于这样的假设：

$$y \sim Bernoulli(\phi)$$

$$x|y=0 \sim N(\mu_0, \Sigma)$$

$$x|y=1 \sim N(\mu_1, \Sigma)$$

他们的概率（密度）函数分别为：

$$p(y) = \phi^y(1-\phi)^{1-y}$$

$$p(x|y=0) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))$$

$$p(x|y=1) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))$$

模型的待估计参数为$\phi, \Sigma, \mu_0, \mu_1$，通常模型有两个不同的期望，而有一个相同的协方差。

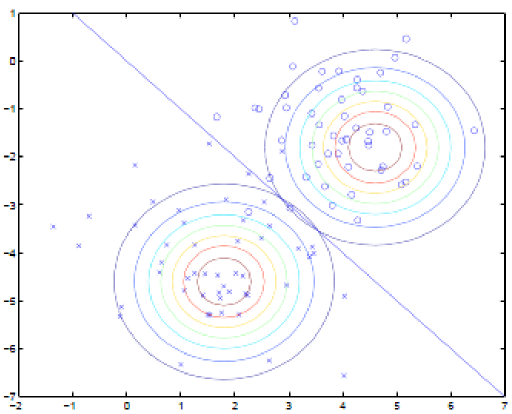
该模型的极大似然对数方程为：

$$l(\phi, \mu_0, \mu_1, \Sigma)$$

$$= log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$$

$$= log \prod_{i=1}^{m} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi)$$

这里的参数有两个μ，表示在不同的结果模型下，特征均值不同，但假设协方差相同。反映在图上就是不同模型中心位置不同，但形状相同。这样就可以用直线来进行分隔判别。



直线两边的y值不同，但协方差矩阵相同，因此形状相同。μ不同，因此位置不同。

求导后，得到参数估计公式：

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

Φ是训练样本中结果y=1占有的比例。μ_0是y=0的样本中特征均值。μ_1是y=1的样本中特征均值。Σ是样本特征方差均值。

在对$\phi, \Sigma, \mu_0, \mu_1$计算完成之后，将新的样本x带入进建立好的模型中，计算出$p(y = 1|x)$、$p(y = 0|x)$，选取概率更大的结果为正确的分类。