高斯判别分析

**实验目的**

* 通过高斯判别及贝叶期算法对评论按照正负情感进行分类

**计算过程**

* 已知有一批已进行情感分类的评论数据，且输入特征x为连续型随机变量，求新输入的样本分类情况p(y|x)；
* 其概率密度函数如下：
* 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}))
* 通过极大似然函数计算\phi,\Sigma,\mu_{0},\mu_{1}的值
* \phi=\frac{1}{m}\sum_{i=1}^{m}{\#\{y^{(i)}=1\}}
* \mu_{0}=\frac{\sum_{i=1}^{m}\#\{y^{(i)}\}x^{(i)}}{\sum_{i=1}^{m}{\#\{y^{(i)}=0\}}}
* \mu_{1}=\frac{\sum_{i=1}^{m}\#\{y^{(i)}\}x^{(i)}}{\sum_{i=1}^{m}{\#\{y^{(i)}=1\}}}
* \Sigma=\frac{1}{m}\sum_{i=1}^{m}{(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^{T}}
* 其中表示p(y)出现的概率，是y=0的样本中特征均值，是y=1的样本中特征均值，是样本特征方差均值，是的行列式；
* 通过概率密度函数，可算出p(x|y)，根据贝叶斯公式，可求出MaxP(y|x)
* 由已知条件可得
* 计算完成

**样本空间**

* n{0<=n<1}，任意特征词语对于评论情感正负的概率

**基本事件**

* “天气”属于正面情感词语的概率

**独立事件**

* 词语A对于正面评论的概率，与词语B对于正面评论的概率相互独立

**备注**

* 判别模型
  + 有限的样本数量，通过判别函数和预测模型，预测实际数据的分类情况
* 生成模型
  + 无穷的样本数据，通过概率密度模型产生出模型，预测实际数据的分类情况
* 离散型随机变量
  + 有些随机变量，可能取到的值是有限多个，或者可列无限多个，则称这些变量为离散型随机变量（只出现可数型的实现集），如二项随机变量，泊松分布
* 连续型随机变量
  + 取任意实数a的概率为0，区间（出现不可数型的实现集，如[1~20]中间的任意数），如正态分布，指数分布，均匀分布
* 数学期望
  + 反映随机变量的平均取值大小
* 与Logistic区别
  + 如果确定数据分布情况，GDA优于logistic回归；否则logistic regession better than gaussian；GDA的假设比Logistic更加严格，满足GDA一定满足Logistic模型假设，反之不成立