朴素贝叶斯

【朴素贝叶斯的假设】

样本属性之间独立。当属性之间不能做相互独立假设的时候，需要数据预处理，减少属性间相关性或将冗余属性摒除。

【原理】

1、给定一个训练样本，每个样本有多个属性X，每个样本可以用属性的向量来表示：$X = (x_1, x_2, .., x_n)$

2、假设有m个类别需要分类。$C_1, C_2, ... C_3$

3、朴素贝叶斯分类器需要计算样本给定属性下属于某一类别的后验概率。即$P(C_i|X)$

后验概率可以通过先验概率与似然度的乘积来计算，即

$$P(C_i \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid C_i)P(C_i)}{P(\mathbf{X})}$$

假设属性之间是相互独立的，则有，

$$P(\mathbf{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

【优点】

1、朴素贝叶斯很容易实现。

2、效果明显

【缺点】

1、属性之间需要相互独立

2、如果某个属性的先验概率为0，会使整个后验概率为0，这个属性的主导作用特别强。

【文本分析的伯努利模型】

Bernoulli Model (伯努利模型): a document is represented by a feature vector with binary elements taking value 1 if the corresponding word is present in the document and 0 if the word is not present.

$$p(x_k|e_i) = \frac{n_{e_i}(x_k)}{N_{e_i}} \quad p(e_i) = \frac{N_{e_i}}{N}$$

where $n_{e_i}(x_k)$ is the number of documents of emotion e_i in which x_k is observed, and N_{e_i} and N is the number of documents with emotion e_i and total documents, respectively.

Multinomial Model (多项式模型): a document is represented by a feature vector with integer elements whose value is the frequency of that word in the document.

$$p(x_k|e_i) = \frac{nw_{e_i}(x_k)}{nw_{e_i}} \quad p(e_i) = \frac{N_{e_i}}{N}$$

where $nw_{e_i}(x_k)$ is the number of times word x_k occurs in documents with emotion e_i, and nw_{e_i} is the total number of words occurs in documents with emotion e_i.