

Naive Bayes

Mia Feng

2018 年 4 月 20 日

1 概述

Naive Bayes: 监督学习，生成式模型。用于分类，朴素指的是各特征条件独立 [?]。用于垃圾邮件分类等。

求解目标：分类，以后验概率最大时对应的类别作为预测分类结果。

$$y = \arg \max_{c_k} P(Y = c_k) \prod_{j=1}^n P(X_j = x_j | Y = c_k) \quad (1)$$

求解思路：最大化后验概率，或者说省略分母不看后最大化似然，取后验概率最大或者似然值最大对应的类标作为预测类标。Concretely, 分别计算各类别出现概率 $P(Y = c_k), k = 1, 2, \dots, m$; 分别计算各类别下对应特征出现的概率 $P(X = x_j | Y = c_k), j = 1, 2, \dots, n, k = 1, 2, \dots, m$; 按公式 (1) 预测类标。求解方法：MAP 或者最大化似然。

1.1 推导

推导 取 I 为示性函数。 a_l 表示 X 的第 l 个特征。样本有 n 个。类标 m 个，特征 s 个。

$$P(Y = c_k) = \frac{\sum_{i=1}^n I(y_i = c_k)}{n}, k = 1, 2, \dots, m \quad (2)$$

$$P(X_j = a_{jl} | Y = c_k) = \frac{\sum_{i=1}^n I(x_i^j = a_{jl}, y_i = c_k)}{\sum_{i=1}^n I(y_i = c_k)} \quad (3)$$

其中, $i = 1, 2, \dots, n, l = 1, 2, \dots, s, k = 1, 2, \dots, m$

改进 为了避免分母为 0，进行了拉普拉斯平滑，即在分母上加了类数目。

$$P(X_j = a_{jl} | Y = c_k) = \frac{\sum_{i=1}^n I(x_i^j = a_{jl}, y_i = c_k) + 1}{\sum_{i=1}^n I(y_i = c_k) + m} \quad (4)$$

2 算法实现

注意实现时取了拉普拉斯平滑，见公式 (4)，且为了防止下溢取对概率值取了对数。

3 Implementation

聚类测试：训练数据由 naivebayes.py 中的 loadDataSet() 生成

```
Connected to pydev debugger (build 173.4301.16)
['dog', 'my', 'is', 'cute'] classified as : 0
['stupid', 'garbage'] classified as : 1
['Haha', 'I', 'really', 'Love', 'You'] classified as : 0
['This', 'is', 'my', 'dog'] classified as : 0
```

图 1: 分类结果