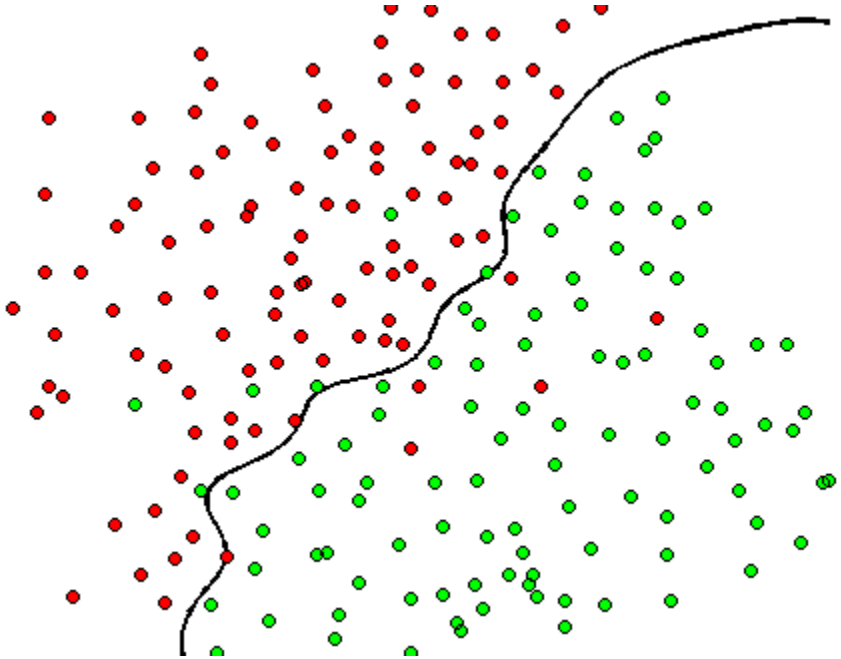
法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

关注 小象学院



分类模型(1)

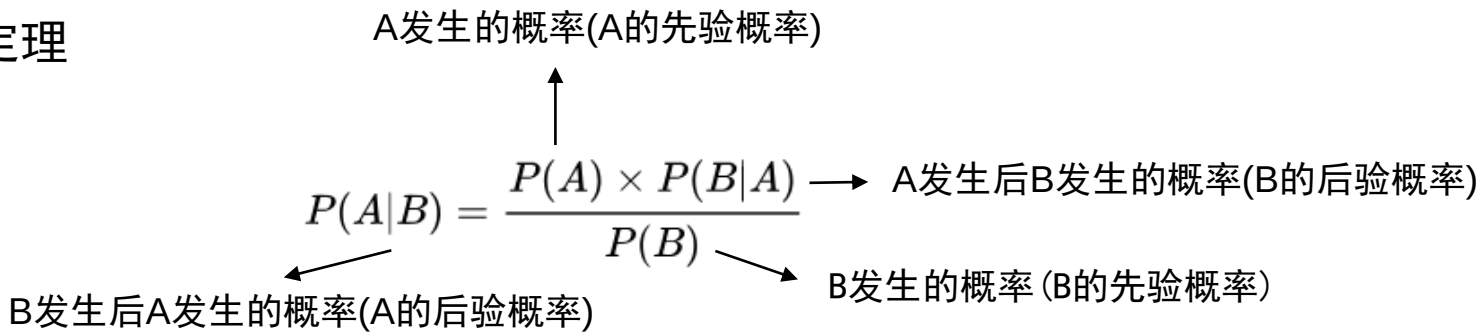
--Robin

目录

- kNN
- 决策树
- 朴素贝叶斯

朴素贝叶斯

- 贝叶斯定理



- 分类问题概率模型
  - 概率模型分类器是一个条件概率模型 $p(C|F_1,\dots,F_n)$
  - 独立的类别变量C，条件依赖于若干特征变量 $F_1,\dots,F_n$

- 根据贝叶斯定理，得到

$$p(C|F_1,\dots,F_n) = \frac{p(C)\ p(F_1,\dots,F_n|C)}{p(F_1,\dots,F_n)}.$$

朴素贝叶斯

朴素贝叶斯概率模型

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

- 不关心分母，因为分母不依赖于类别变量C而且特征***F***是给定的，等价于

$$p(C|F_1,\ldots,F_n) \quad \propto \quad p(C)\ p(F_1,\ldots,F_n|C)$$

- 只需要关心分子即可
- 假定样本每个特征与其他特征都不相关（特征的独立性），即样本所包含的属性在判定其是否为某一类时的概率分布上是独立的
- 分子等价于 $p(C)\ p(F_1,\ldots,F_n|C)$ ⇨ $p(C)p(F_1|C)\ldots p(F_n|C)$
- 上式中的每一项都可以从训练样本中计算出，由此可以得到某个样本在每个类别对应的概率，从中取出最大概率的那个类就是其预测分类

朴素贝叶斯

朴素贝叶斯概率模型

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

$$p(C|F_1,\ldots,F_n) \propto p(C)p(F_1|C)\ldots p(F_n|C)$$

- 构建分类器的简单方法，不是训练分类器的单一算法，而是一系列基于相同原理的算法
- 尽管有着这些朴素思想和过于简单的假设，但朴素贝叶斯分类器在很多复杂的现实情形中仍能取得相当好的效果

朴素贝叶斯

概率计算及参数估计

- 特征是离散变量，可以用频率来估计概率

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

如果是sunny，是否Play?

P(Yes | Sunny) => P(Sunny | Yes) * P(Yes)
= 3/9 * 9/14 = 0.2143

P(No | Sunny) => P(Sunny | No) * P(No)
= 2/5 * 5/14 = 0.1429

P(Yes | Sunny) > P(No | Sunny) => Play

朴素贝叶斯

概率计算及参数估计

- 特征是连续的，一种通常的假设是这些连续数值服从高斯分布
  - 首先对数据根据类别分类，然后计算每个类别中特征x的均值和方差。令mu_c表示为x在c类上的均值，令sigma^2_c为x在c类上的方差。在给定类中某个值的概率 P(x=v|c)，可以通过将v表示为均值为mu_c，方差为sigma^2_c的正态分布计算出来。

$$p(x=v|c)=\frac{1}{\sqrt{2\pi\sigma_c^2}}e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

朴素贝叶斯

sklearn.naive_bayes

- 根据特征分布的假设
- Gaussian Naive Bayes（高斯模型），适用于特征是连续的。
- Multinomial Naive Bayes（多项式模型），适用于特征是离散的。
- Bernoulli Naive Bayes（伯努利模型），用于离散特征的情况，且每个特征的取值只能是1和0
    - 如：文本分类中，某个单词在文档中出现过，则其特征值为1，否则为0

```
sklearn.naive_bayes.GaussianNB
sklearn.naive_bayes.MultinomialNB
sklearn.naive_bayes.BernoulliNB
```

朴素贝叶斯

- 朴素贝叶斯的优缺点

优点	缺点
• 易于实现 • 对小规模的数据表现很好 • 适合增量式训练（数据量超出内存时，我们可以一批批的去增量训练） • 对缺失数据不太敏感	• 在属性个数比较多或者属性之间相关性较大时，分类效果不好 • 由于是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率

联系我们

小象学院：互联网新技术在线教育领航者

– 微信公众号：**小象学院**