云社区 > 云博客 > 博客详情

机器学习|朴素贝叶斯（Naive Bayes）算法总结

寻水的鱼 发表于 2018-01-31 10:34:02 01-31 10:34

寻水的鱼 发表于 2018-01-31 10:34:02 01-31 10:34
据算法 机器学习

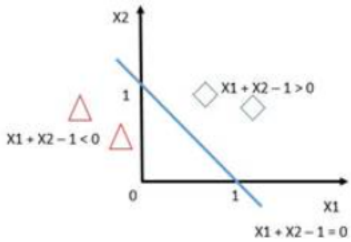
【摘要】 1. 简单定义朴素贝叶斯是一种简单快速的有监督学
法。那么如何解读上面这句定义？从描述定义的词汇入手。
类是一种确定数据类别的问题，例如天气阴晴，花开花谢，
目标是将具有相似或相同特征的对象聚集。而线性分类（器
性组合来坐出分类决定。也就是说不同类别之间的分类边界
图。线性算法的优点是训练和预测的效率比较高，但最终效

简单定义

朴素贝叶斯是一种简单快速的有监督学习类线性分类算法。

那么如何解读上面这句定义？从描述定义的词汇入手。

线性分类：分类是一种确定数据类别的问题，例如天气阴晴，花开花谢，水果酸甜等。其
的对象聚集。而线性分类（器）透过特征的线性组合来坐出分类决定。也就是说不同类别之
图。

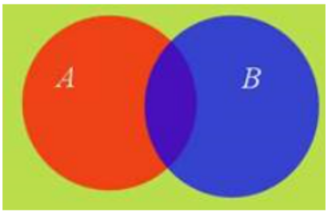


线性算法的优点是训练和预测的效率比较高，但最终效果对特征的依赖程度较高，需要数据

有监督学习：监督学习是指用已经标记好的数据，做训练来预测新数据的类型，或者是值
测一个值叫做回归。与无监督学习最简单的区分方法就是看训练样本数据有无标签，有即
习。

朴素：假设某个特征的出现与其它特征的出现是独立的。举个例子：想要根据颜色、形状
橙色、球形、味道酸甜的水果很可能是橘子。上述描述中，即使这些特征依赖于彼此或取决
特性均可以单独地提升该果实是橘子的可能性，这就是为什么它被称为朴素。

贝叶斯定理：直白的说，贝叶斯定理实际描述用于计算条件概率的方法和公式。最简单的
在事件B发生的情况下，另一事件A发生的概率，数学形式为P(A|B)。更直观的解释可借用下



在事件B发生的情况下，事件A发生的概率为：

$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

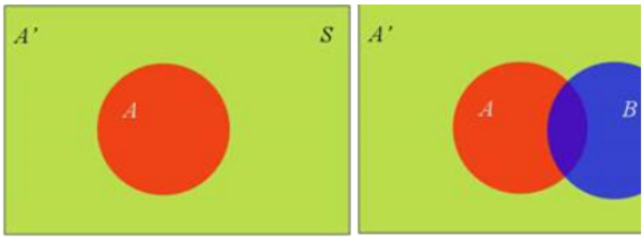
则：

$$P(A\cap B)=P(A|B)P(B)$$
$$P(A\cap B)=P(B|A)P(A)$$

得到的条件概率计算公式为，

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

此时需要对P(B)做一个引申，即需要知道事件B的全概率。则推广上述的场景至更广义一些。
事件A与A'的和，如图：



引入事件B后，事件B可划分为两个部分，1）与事件A重合的部分；2）与事件A'重合的部分，

$$P(B)=P(B\cap A)+P(B\cap A')$$

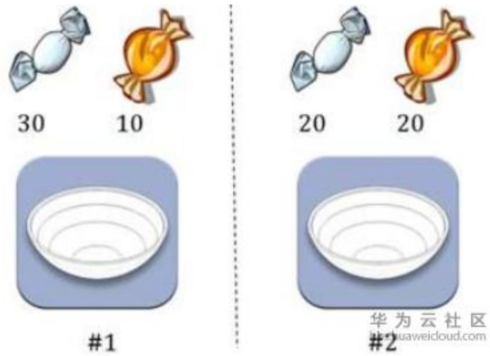
由上页的推导可得出：

$$P(B)=P(B|A)P(A)+P(B|A')P(A')$$

上述公式即为全概率公式。它的含义是，如果A和A'构成样本空间的一个划分，那么事件**B**的
别乘以**B**对这两个事件的条件概率的和。将这个公式代入上一节的条件概率公式，就得到了条

$$P(A|B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A')P(A')}$$

为了更直观的理解，在此举一个例子：



两个一模一样的碗，一号碗有30颗水果糖和10颗巧克力糖，二号碗有水果糖和巧克力糖各20
中摸出一颗糖，发现是水果糖。请问这颗水果糖来自一号碗的概率有多大？
我们假定，W_1表示一号碗，W_2表示二号碗。由于这两个碗是一样的，所以$P(W_1)=P(W_2)$
前，这两个碗被选中的概率相同。因此，$P(W_1)=0.5$，没有做实验之前，来自一号碗的概率
再假定，T表示水果糖，所以问题就变成了在已知T的情况下，来自一号碗的概率有多大，即

$$P(W_1|T)=P(W_1)\frac{P(T|W_1)}{P(T)}$$

已知，$P(W_1)$等于0.5，$P(T|W_1)$为一号碗中取出水果糖的概率，等于0.75，那么求出P(T)就
得到：

$$P(T)=P(T|(W_1)P(W_1)+P(T|(W_2)P(W_2)=0.75\times 0.5+0.5\times 0.5=$$

$$P(W_1|T)=0.5\times\frac{0.75}{0.625}$$

上述解释只是针对空间内总样本数为2的情况，一般在机器学习中对训练样本和测试样本的数
分类，就是把标签为**0**和**1**的样本分别取出，建立两种训练模型**$model_0$**和**$model_1$**。在判别测
分别代入$model_0$和$model_1$，选取拟合更好的模型。因此再推广朴素贝叶斯的公式至二分类

$$P(y|(x_1,x_2,\dots,x_n))=\frac{P((x_1,x_2,\dots,x_n)|y=1)P(y=}{P(x_1,x_2,\dots,x_n)}$$

由全概率公式可知：

则：

$$P(y|(x_1,x_2,\dots,x_n)) = \frac{P((x_1,x_2,\dots,x_n)|y=1)P(y=1)}{P((x_1,x_2,\dots,x_n)|y=1)P(y=1)+P((x_1,x_2,\dots,x_n)|y}$$

根据朴素贝叶斯性质得知，若x_1、x_2…….x_n之间相互独立，则有：

$$P((x_1,x_2,\dots,x_n)|y=1) = \prod_{i=1}^{n} P(x_i|y=1)$$

（举个例子便于理解，在下雨天时小明穿板鞋的概率是0.9，下雨天时小明吃米饭的概率为(
吃米饭的概率为0.5 x 0.9 =0.45。）
代入上页公式可得：

$$P((x_1,x_2,\dots,x_n)|y=1) = \frac{P(y=1)\prod_{i=1}^{n} P(x_i|y=1)}{P(y=1)\prod_{i=1}^{n} P(x_i|y=1)+P(y=0)\prod_{i=1}^{n} P}$$

$P(x_i|y=1)$表示所有的标签为1的样本中，$x_i=1$的样本所占的比率，$P(x_i|y=0)$表示所有标签为(
何得知P(y=1)和P(y=0)呢？其实很容易理解，**P(y=1)**是训练样本中标签为**1**的样本比率，**P(y**
，在样本空间中找出标签为1的样本与总样本数m的比值即为P(y=1)，同理可得P(y=0)。

解释说明与分析

逻辑解释

$$P(y|(x_1,x_2,\dots,x_n)) = \frac{P((x_1,x_2,\dots,x_n)|y=1)P(y}{P(x_1,x_2,\dots,x_n)}$$

在二分类一般形式的朴素贝叶斯公式中，可以将等式右边分为P(y=1)和 $\frac{P(x_1,x_2,\dots,x_n|y=1)}{P(x_1,x_2,\dots,x_n)}$
的乘积，我们将P(y=1)称为先验概率；

$$\frac{P(x_1,x_2,\dots,x_n|y=1)}{P(x_1,x_2,\dots,x_n)}$$

称为可能性函数，即调整因子；而等式左边的

$$P(y=1|x_1,x_2,\dots,x_n)$$

称为后验概率。那么公式就可以理解为：

后验概率 = 先验概率 x 调整因子

这就是贝叶斯推断的含义。我们先预估一个"先验概率"，然后加入实验结果，看这个实验Ⅰ
率"，由此得到更接近事实的"后验概率"。在这里，如果调整因子

$$\frac{P(x_1,x_2,\dots,x_n|y=1)}{P(x_1,x_2,\dots,x_n)} > 1$$

意味着"先验概率"被增强，事件A的发生的可能性变大；如果调整因子**=1**，意味着B事件无Ⅰ
调整因子**<1**，意味着"先验概率"被削弱，事件A的可能性变小。在某种意义上讲，这个公式「

普适推广至多分类下的朴素贝叶斯
在更多的情况下，样本的分类情况远远多于二分，标签的数量不只是2。那么一般化多分类Ⅰ

$$P(y_j|x_1,x_2,\dots,x_n) = \frac{P(x_1,x_2,\dots,x_n|y_j)P(y_j)}{P(x_1,x_2,\dots,x_n)}$$

我们可定义 v_{MAP} 为最大后验值，$\arg max$ 函数定义为求出使目标函数取得最大值对应的

$$v_{MAP} = \arg max_{y_j \epsilon Y} P(y_j|x_1,x_2,\dots,x_n)$$

由于x_1、x_2……..x_n之间相互独立，则有：

由于 $P(x_1,x_2,...,x_n)$ 给定输入为常数，则有：

$$v_{MAP} \propto P(y_j)\prod_{i=1}^{n} P(x_i|y_j)$$

实际上，上式右边就是朴素贝叶斯分类器公式：

$$v_{NB} = \arg max_{y_j \epsilon Y} P(y_j)\prod_{i=1}^{n} P(x_i|y_j)$$

朴素贝叶斯优缺点

优点：
这是一个相对容易构建和理解的算法。
使用该算法比许多其他分类算法能更快地预测类。
使用小数据集也可以容易地训练数据。
类条件特征独立分布意味着每个类分布可以独立地估计为一维分布。这又有助于缓解数据降

缺点：
如果给定没有出现过的类和特征，则该类别的条件概率估计将出现0，该问题被称为“零条件
为它会擦除其他概率中的所有信息。
另一个缺点是它的特征之间独立的假设非常强。 在现实生活中几乎很难找到这样的数据集。
间相关性较大时，朴素贝叶斯分类模型的分类效率低。而在属性相关性较小时，朴素贝叶斯

改进
拉普拉斯平滑
拉普拉斯平滑是为了解决上述“零条件概率问题”。考虑一种特殊情况，如果样本中变量x_i的耳
式子 $\prod_{i=1}^{n} P(x_i|y_j)$ 的结果也为0，显然不合理。因此需要对0概率进行调整。对于此问题，
滑法，以基础的二分类为例：

$$\begin{aligned} P(x_i|y=1) &= \frac{1+\sum_{i=1}^{m} I(x_j^{(i)}=1, y^{(i)}=1)}{1+\sum_{i=1}^{m} I(x_j^{(i)}=1, y^{(i)}=1)+1+\sum_{i=1}^{m} I(x_j^{(i)}=0, y} \\ &= \frac{1+\sum_{i=1}^{m} I(x_j^{(i)}=1, y^{(i)}=1)}{2+\sum_{i=1}^{m} I(y^{(i)}=1)} \end{aligned}$$

更一般情况下，加入x_i有k个不同的取值，当样本数量足够大时加入拉普拉斯平滑后并不会影

$$P(x_i|y=k_i) = \frac{1+\sum_{i=1}^{m} I(x_j^{(i)}=k_i, y^{(i)}=1)}{k+\sum_{i=1}^{m} I(y^{(i)}=1)}$$

作者|王天鸿

数据算法 机器学习

展开全部内容

评论文章 (1) (0)

上一篇：如何成为一名优秀的JA... 下一篇：中国