贝叶斯分类器

机器学习笔记 create by siwanghu v1.0

贝叶斯分类器是一种统计分类方法。

通过对象的**先验概率**，利用贝叶斯公式计算出其**后验概率**，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类

先验概率：事件发生前的预判概率，通常是基于历史数据的统计，可以由背景常识得出，也可以是人的主观观点给出。一般都是单独事件概率，如 P(x),P(y)。

后验概率：事件发生后求的反向条件概率，基于先验概率求得的反向条件概率。概率形式与条件概率相同。

条件概率：一个事件发生后另一个事件发生的概率。一般的形式为 P(x|y)表示 y 发生的条件下 x 发生的概率。

贝叶斯公式:

$$p(c_i \mid x,y) = \frac{p(x,y \mid c_i)\,p(c_i)}{p(x,y)}$$

P(c_i|x, y) 后验概率，求解的目标。

P(x, y|c_i) 是条件概率，又叫似然概率，一般是通过历史数据统计得到。

P(c_i) 是先验概率，一般都是人主观给出的。贝叶斯中的先验概率一般特指它。

P(x, y) 其实也是先验概率，只是在贝叶斯的很多应用中不重要，需要时往往用全概率公式计算得到

假设c_i是文章种类，是一个枚举值。x, y 是单词标签，表示文章中关键单词的出现次数。

在拥有训练集的情况下，显然除了后验概率 P(c_i|x, y)无法得到，P(x, y), P(c_i), P(x, y|c_i)都是可以在抽样集合上统计出的。

朴素贝叶斯分类器

朴素贝叶斯分类器是一个基于贝叶斯定理的比较简单的概率分类器，其中 naive（朴素）是指的对于模型中各个 feature（特征） 有强独立性的假设，并未将 feature 间的相关性纳入考虑中

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

求样本 X 的分类

$X = (Outlook = "Sunny", Temperature = Cool", Humidity = "High", Wind = "Strong")$

每个类的先验概率 $P(C_i)$可以根据训练样本计算：

P(Play_Tennis=”Yes”)=9/14=0.643

P(Play_Tennis=”No”)=5/14=0.357

为计算 $P(X/C_i)$，i=1,2，我们计算下面的条件概率：

P(Outlook=”sunny”|Play_Tennis=”Yes”)=2/9=0.222

P(Outlook=”sunny”|Play_Tennis=”No”)=3/5=0.600

P(Temperature=”Cool”|Play_Tennis=”Yes”)=3/9=0.333

P(Temperature=”Cool”|Play_Tennis=”No”)=1/5=0.200

P(Hudimity=”High”|Play_Tennis=”Yes”)=3/9=0.333

P(Hudimity=”High”|Play_Tennis=”No”)=4/5=0.800

P(Wind=”Strong”|Play_Tennis=”Yes”)=3/9=0.333

P(Wind=”Strong”|Play_Tennis=”No”)=3/5=0.600

我们得到：

P(X|Play_Tennis=”Yes”)=0.222×0.333×0.333×0.333=0.00823

P(X|Play_Tennis=”No”)=0.600×0.200×0.800×0.600=0.0576

P(X|Play_Tennis=”Yes”)P(Play_Tennis=”Yes”)=0.00823×0.643=0.0053

P(X|Play_Tennis=”No”)P(Play_Tennis=”No”)=0.0576×0.357=0.0206

因此，对于样本 X，朴素贝叶斯分类预测 Play_Tennis=”No”

朴素贝叶斯分类器算法步骤：

计算先验概率 P(c_l)和条件概率 P(x|c_l)

计算后验概率 P(c_l|x) =P(c_i)$\prod_{i=1}^{d}$ P($x_i|c_l$)

确定 x 的类 h(x)=argmax P(c_i)$\prod_{i=1}^{d}$ P($x_i|c_l$)