朴素贝叶斯分类器（Naive Bayes Classifier）

基本概念

贝叶斯分类器是一族分类算法的总称，该族算法均以贝叶斯定理为基础，统称为贝叶斯分类器。贝叶斯分类器的分类原理是通过先验概率利用贝叶斯公式计算出其后验概率，选择具有最大后验概率的类作为该对象所属的类别。
设S为实验E的样本空间，$B_1, B_2, \ldots, B_n$为E的一组事件，若：
1. $B_i \cap B_j = \emptyset, i \neq j, i, j = 1, 2, \ldots, n$
2. $B_1 \cup B_2 \cup \cdots \cup B_n = S$

则称$B_1, B_2, \ldots, B_n$为样本空间S的一个划分。对于每次试验，事件$B_1, B_2, \ldots, B_n$中有且仅有一个事件发生。
设A为实验E的事件，且$P(A) > 0, P(B_i) \geq 0$，则：

- 全概率公式：

$$\begin{aligned} P(A) &= P(A/B_1) * P(B_1) + P(A/B_2) * P(B_2) + \cdots + P(A/B_n) * P(B_n) \\ &= \sum_{i=1}^{n} P(A/B_i)P(B_i) \end{aligned} \tag{1}$$

- 贝叶斯定理：

$$P(B_j/A) = \frac{P(A/B_j)P(B_j)}{\sum_{i=1}^{n} P(A/B_i)P(B_i)} \tag{2}$$

假设有训练集D，样本$\vec{x} = (x_1, x_2, \ldots, x_n)$，类别标签$y = c_1, c_2, \ldots, c_k$。对于某一样本$x_i$，假设其类别标签为$c$，则朴素贝叶斯分类器的训练过程就是基于$D$来估计类别先验概率$P(c)$，并为每个属性估计条件概率$P(x_i|c)$。
朴素贝叶斯分类器（naive Bayes classifier）采用了“属性条件独立性假设”：对已知类别，假设所有的属性相互独立，也就是说，假设每个属性独立地对分类结果产生影响。

举例说明

下面以图1的示例来说明朴素贝叶斯分类器的决策过程：

Credit	Term	Income	y
excellent	3 yrs	high	safe
fair	5 yrs	low	risky
fair	3 yrs	high	safe
poor	5 yrs	high	risky
excellent	3 yrs	low	risky
fair	5 yrs	low	safe
poor	3 yrs	high	risky
poor	5 yrs	low	safe
fair	3 yrs	high	safe

图1

已知训练数据如图1所示，那么给定一样本$\{Credit = excellent, Term = 5yrs, Income = high\}$，它的类别标签$y$应该是多少呢？
对于该问题用朴素贝叶斯进行分类，我们只需要计算$P(y = safe|Credit = excellent, Term = 5yrs, Income = high)$和$P(y = risky|Credit = excellent, Term = 5yrs, Income = high)$并计算它们的大小即可，如果前者大则该样本的类别标签为$safe$，如果后者大则该样本的类别标签为$risky$。
由图1我们可得到一些先验概率：

$$\begin{cases} P(y = safe) = \frac{5}{9} \\ P(y = risky) = \frac{4}{9} \\ P(Credit = excellent|y = safe) = \frac{1}{9} \\ P(Term = 5yrs|y = safe) = \frac{2}{9} \\ P(Income = high|y = safe) = \frac{3}{9} \\ P(Credit = excellent|y = risky) = \frac{1}{9} \\ P(Term = 5yrs|y = risky) = \frac{2}{9} \\ P(Income = high|y = risky) = \frac{2}{9} \\ P(Credit = excellent) = \frac{2}{9} \\ P(Term = 5yrs) = \frac{4}{9} \\ P(Income = high) = \frac{5}{9} \end{cases}$$

{P(y=safe)=59P(y=risky)=49P(Credit=excellent|y=safe)=19P(Term=5yrs|y=safe)=29P(Income=high|y=safe)=39P(Credit=excellent|y=risky)=1

由公式(2)可知：

$$\begin{aligned} &P(y = safe|Credit = excellent, Term = 5yrs, Income = high) \\ &= \frac{P(Credit = excellent, Term = 5yrs, Income = high|y = safe)P(y = safe)}{P(Credit = excellent, Term = 5yrs, Income = high)} \\ &P(y = risky|Credit = excellent, Term = 5yrs, Income = high) \\ &= \frac{P(Credit = excellent, Term = 5yrs, Income = high|y = risky)P(y = risky)}{P(Credit = excellent, Term = 5yrs, Income = high)} \end{aligned}$$

P(y=safe|Credit=excellent,Term=5yrs,Income=high)=P(Credit=excellent,Term=5yrs,Income=high|y=safe)P(y=safe)P(Credit=excellent,Term=
P(y=risky|Credit=excellent,Term=5yrs,Income=high)=P(Credit=excellent,Term=5yrs,Income=high|y=risky)P(y=risky)P(Credit=excellent,Term

而

$$\begin{aligned} &P(Credit = excellent, Term = 5yrs, Income = high|y = safe) \\ &= P(Credit = excellent|y = safe) * P(Term = 5yrs|y = safe) \\ &\quad * P(Income = high|y = safe) \\ &= \frac{1}{9} * \frac{2}{9} * \frac{3}{9} = \frac{6}{729} \\ &P(Credit = excellent, Term = 5yrs, Income = high|y = risky) \\ &= P(Credit = excellent|y = risky) * P(Term = 5yrs|y = risky) \\ &\quad * P(Income = high|y = risky) \\ &= \frac{1}{9} * \frac{2}{9} * \frac{2}{9} = \frac{4}{729} \\ &P(Credit = excellent, Term = 5yrs, Income = high) \\ &= P(Credit = excellent) * P(Term = 5yrs) * P(Income = high) \\ &= \frac{2}{9} * \frac{4}{9} * \frac{5}{9} = \frac{40}{729} \end{aligned}$$

P(Credit=excellent,Term=5yrs,Income=high|y=safe)=P(Credit=excellent|y=safe)∗P(Term=5yrs|y=safe)∗P(Income=high|y=safe)=19∗29∗39=6
P(Credit=excellent,Term=5yrs,Income=high|y=risky)=P(Credit=excellent|y=risky)∗P(Term=5yrs|y=risky)∗P(Income=high|y=risky)=19∗29∗29=
P(Credit=excellent,Term=5yrs,Income=high)=P(Credit=excellent)∗P(Term=5yrs)∗P(Income=high)=29∗49∗59=40729

所以

$$P(y = safe|Credit = excellent, Term = 5yrs, Income = high) = \frac{\frac{6}{729} * \frac{5}{9}}{\frac{40}{729}} = \frac{15}{180}$$

$$P(y = risky|Credit = excellent, Term = 5yrs, Income = high) = \frac{\frac{4}{729} * \frac{4}{9}}{\frac{40}{729}} = \frac{8}{180}$$

P(y=safe|Credit=excellent,Term=5yrs,Income=high)=6729∗5940729=15180P(y=risky|Credit=excellent,Term=5yrs,Income=high)=4729∗4940

由此可知$P(y = safe|Credit = excellent, Term = 5yrs, Income = high)$>$P(y = risky|Credit = excellent, Term = 5yrs, Income = high)$，所以该样本的类别标签应预测为safesafe。

朴素贝叶斯分类器优缺点

优点：逻辑简单，只需要知道贝叶斯公司、全概率公式即可。
缺点：朴素贝叶斯分类器是建立在“属性条件独立性假设”基础上的，而现实中往往各个属性之间是有联系的并不是独立地对分类结果产生影响。

朴素贝叶斯分类器种类

依据样本服从的概率分布的不同， 朴素贝叶斯分类器可分为高斯贝叶斯分类器、多项式贝叶斯分类器、伯努利贝叶斯分类器等。

- 高斯贝叶斯分类器：假设属性的条件概率分布满足高斯分布。
- 多项式贝叶斯分类器：假设属性的条件概率分布满足多项式分布。

- 伯努利贝叶斯分类器：假设属性的条件概率分布满足二项分布。

参考文献

周志华. 机器学习 [D]. 清华大学出版社，2016.
华校专、王正林. Python大战机器学习 [D]. 电子工业出版社，2017.

图1

更多资料请移步github：
https://github.com/GarryLau/MachineLearning