"""

Created on 2017/8/13

machine-learning-course

@author: DSG
"""

二、贝叶斯公式

条件概率: 设 A 与 B 是样本空间 Ω 中的两事件，若 $P(B)>0$ 则称 $P(A|B)=\frac{P(AB)}{P(B)}$ 为在 B 发生下 A 的条件概率，同样有 $P(B|A)=\frac{P(AB)}{P(A)}$。

性质：若 f 中的 $A_1,A_2,\cdots,A_n$ 互不相容，则 $P\left(\bigcup_{n=1}^{\infty}A_n|B\right)=\sum_{n=1}^{\infty}P(A_n|B)$。 f 为事件域。

乘法公式：若 $P(B)>0$，则 $P(AB)=P(A|B)P(B)$，

若 $P(A_1,A_2,\cdots,A_{n-1})>0$，则 $P(A_1,A_2,\cdots,A_n)=P(A_1)P(A_2|A_1)P(A_3|A_1A_2)\cdots P(A_n|A_1A_2\cdots A_{n-1})$

全概率公式：设 $B_1,B_2,\cdots,B_n$ 为样本空间 Ω 上的一个分割，即 $B_1,B_2,\cdots,B_n$ 互不相容，且 $\bigcup_{n=1}^{n}B_i=\Omega$，如果 $P(B_i)>0,i=1,2,\cdots,n$，则对任一事件 A 有 $P(A)=\sum_{n=1}^{\infty}P(B_i)P(A|B_i)$ 这里完整的写法 $P(A\Omega)=\sum_{n=1}^{\infty}P(B_i)P(A|B_i)$，解释：$A$ 在样本空间上发生的概率，就等于在所有分割上发生的概率总和。

贝叶斯公式设 $B_1,B_2,\cdots,B_n$ 为样本空间 Ω 上的一个分割，即 $B_1,B_2,\cdots,B_n$ 互不相容，且 $\bigcup_{n=1}^{n}B_i=\Omega$ 如果 $P(A)>0,P(B_i)>0,i=1,2,\cdots,n$，则：

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_j)},i=1,2,\cdots,n$$

推导：对 $P(B|A)=\frac{P(AB)}{P(A)}$ 分子用乘法公式，对分母用全概率公式

B_i 同常叫做导致事件 A 发生的概率。如果只对 $P(B|A)=\frac{P(AB)}{P(A)}$ 的分子用乘法公式，分母不动：

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{P(A)},i=1,2,\cdots,n$$

这个公式就是贝叶斯算法的核心公式。

$$\text{p(类别|特征)}=\frac{p(\text{特征|类别})\text{p(类别)}}{\text{p(特征)}}$$

二、朴素贝叶斯分类器的公式

假设某个体有 n 项特征（Feature），分别为 $f_1,f_2,\cdots,f_n$。现有 m 个类别（Category），分别为 $c_1,c_2,\cdots,c_m$。贝叶斯分类器就是计算某个数据在自己特征下概率最大的那个分类，也就是求下面这个算式的最大值：

$$P(c|f_1f_2\cdots f_n)=\frac{P(f_1f_2\cdots f_n|c)P(c)}{P(f_1f_2\cdots f_n)}$$

因为对于同一条数据来讲 $P(f_1f_2\cdots f_n)$ 是一样的。所以对于问题就会变成：

$$P(c|f_1f_2\cdots f_n)=P(f_1f_2\cdots f_n|c)P(c)$$

如果 n 项特征 $f_1,f_2,\cdots,f_n$ 彼此独立则上公式又可以变为：

$$P(c|f_1f_2\cdots f_n)=P(c)\prod_{n=1}^{n}P(f_n|c)$$

因为 $\ln f(x)$ 和 $f(x)$ 有相同的单调性和增减性所以为了避免数据下越结，同长会对上公式两边取自然对数那么问题就会变成比较下公式的大小：

$$\max\sum_{n=1}^{n}P(f_n|c)+P(c)$$

一般把上述 $f_1,f_2,\cdots,f_n$ 用向量 $\vec{w}$ 替换，w_n 表示向量的第 n 维：

$$P(c|\vec{w})=P(c)\prod_{n=1}^{n}P(w_n|c)$$

$$\sum_{n=1}^{n}P(w_n|c)+P(c)$$

对于特征为布尔型的算法应用，首先要算出每个类别的概率，还要知道每个类别中所有特征出现的概率（其

实就每个特征出现的次数占总特征数总和的比份，在文档分类器其中，就是每个词汇出现的次数占这个类别词汇总数的比份），然后找到测试数据的所有特征在每个类别中对应特征的概率，然后把他们想乘想加，在加上类别概率。

如果特征为数值型的数据，可以计算下式子的大小，不同的是所有的 w_c^n 就是每个特征在总特征总数下的比分，下式子和 $\sum_{n=1}^{n} P(w_n \mid c) + P(c)$ 有相同的单调性：值得注意的是期望和分布的峰值并不是在同一点取值。

$$(x^{(i)})^T w_c^n + P(c)$$