朴素贝叶斯算法

SYSU & SPF 机器学习研修班系列讲义 3

主讲人：蔡淼
中山大学物理学院

一． 数学准备

1.1 样本空间的划分

设 Ω 为试验 E 的样本空间， $B_1,B_2\cdots B_n$ 为 E 的一组事件，若有：

1° $B_iB_j=\varnothing,i,j=1,2...,n$

2° $B_1\cup B_2\cup...\cup B_n=\Omega$

则称 $B_1,B_2\cdots B_n$ 为样本空间 Ω 的一个划分。（图 1）

1.2 全概率公式

设 Ω 为试验 E 的样本空间， A 为 E 的事件， $B_1,B_2\cdots B_n$ 为 Ω 的一个划分，且 $P(B_i)\neq 0$ ，则有公式

$$P(A)=\sum_{i=1}^{n}P(B)P(A|B_i) \tag{1}$$

(1)式又称全概率公式，其证明如下：

$$A=A\Omega=A\left(B_1\cup B_2\cup...\cup B_n\right)=AB_1\cup AB_2\cup...\cup AB_n$$

$$P(A)=P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+...+P(A|B_n)P(B_n)$$

$$=\sum_{i=1}^{n}P(B)P(A|B_i)$$

全概率公式体现了概率的**可加性**。

1.3 贝叶斯公式

设 Ω 为试验 E 的样本空间， A 为 E 的事件， $B_1,B_2\cdots B_n$ 为 Ω 的一个划分，且 $P(A)>0$ ， $P(B_i)>0(i=1,2,...,n)$ ，则

$$P\left(B_i|A\right)=\frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n}P\left(A|B_j\right)P(B_j)},i=1,2,...,n \tag{2}$$

(2)式即贝叶斯公式

二· 朴素贝叶斯分类器

朴素贝叶斯分类器的基本思想其实非常简单，即对于给定的待分类项 $X\{a_1,a_2,...,a_n\}$，求解在此项出现的条件下各个类别 y_i 出现的概率，哪个 $P(y_i|x)$ 最大，就将此待分类项归为哪个类别。

1. 算法思路

假令我们要考察的对象具有 n 个特征向量，则每一个样本都可以表示为一个 n 维向量，向量在每一个维度上的分量即相应的特征值，这个向量即特征向量。比如一个人的生理特征可以分为身高、体重和性别，那我们就可以把每一个人视作一个特征向量 $human=[height,weight,gender]$ 。

因此，输入空间 $\chi\subseteq R^n$ 为 n 维向量的集合，同时容易理解输出空间为类标记集合 $Y=\{c_1,c_2,...,c_k\}$。则贝叶斯分类器的输入为特征向量 $x\in\chi$ ，输出为类标记 $y\in Y$ 。即每个输入的特征向量都会被分为类空间中的某一个类。

$P(X,Y)$ 是 X 与 Y 的联合概率分布。

训练数据集为

$$T=\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$$

朴素贝叶斯法通过训练数据集学习联合概率分布 $P(X,Y)$，由概率论相关知识（见第一节数学准备）可知，

$$P(X=x,Y=c_k)=P(Y=c_k)P(X=x|Y=c_k)$$

故学习联合概率分布可以通过学习先验概率分布及条件概率分布来进行。其中，先验概率分布为

$$P(Y=c_k),k=1,2,...,K$$

条件概率分布为

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},...,X^{(n)}=x^{(n)}|Y=c_k),k=1,2,...,K$$

但是实际上，若认为 $X^{(1)},...,X^{(n)}$ 之间不是互相独立的，那么需要学习的参数具有指数级的数量（指数爆炸），而这在实际操作中是不可行的。实际上，假设 $x^{(j)}$ 可取值有 S_j 个， $j=1,2,..,n$ ，Y 可取值有 K 个，那么需学习的参数个数为 $K\prod_{j=1}^{n}S_j$ 。

因此，我们加入条件独立性假设，即

$$
\begin{aligned}
&P(X=x|Y=c_k)\\
&=P(X^{(1)}=x^{(1)},...,X^{(n)}=x^{(n)}|Y=c_k)\\
&=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)
\end{aligned}
\tag{1}
$$

由于这个假设是一个较强的假设，朴素贝叶斯法的朴素二字也因此得名。这一假设使朴素贝叶斯法变得简单，但有时也会牺牲一定的分类准确性。

具体而言，加入条件独立性后，需要学习的参数数量为$K\left(\sum_{j=1}^{n}S_j\right)$。

思考：两种情况下需要学习的参数数量如何得到？

在进行分类的过程中，对给定的输入 x，通过学习到的模型计算后验概率帆布$P(Y=c_k|X=x)$，将其中最大的类作为 x 的类输出。

后验概率的计算为

$$
P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)}
\tag{2}
$$

讲 1 式代入 2 式，得到

$$
P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)},k=1,2,...,K
$$

对上式取最大值即我们最后得到的分类输出，同时由于分母对于任意的c_k都是相同的，故对分类输出没有影响，所以可以只对分子取最大值，故最后的分类公式为

$$
y=\arg\max_{c_k}P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)
$$

2. 后验概率最大化

由第一小节我们可以知道，朴素贝叶斯分类的最后一步是在所有可能的类中取后验概率最大的类作为输出，这符合我们的直观感受，同时也可以通过数学方法来证明这种取法可以让我们的期望风险最小化。

选择 0-1 损失函数

$$
L(Y,f(X))=\begin{cases}1,Y\neq f(X)\\0,Y=f(X)\end{cases}
$$

考虑到这里的分布是离散概率分布，在联合分布概率为$P(X,Y)$的情况下，损失函数的期望函数是[2]

上式是连续的情况，在离散的情况下，容易得到：

$$
\begin{aligned}
&R_{\exp}(f)=E(L)\\
&=\sum_{k=1}^{K}\sum_{i=1}^{n}L(c_k,f(x_i))P(c_k\mid x_i)P(x_i)\\
&=E_X\sum_{k=1}^{K}L(c_k,f(X))P(c_k\mid X)
\end{aligned}
$$

由于 E_X 可视作常数，要使上式最小，只需让 $\sum_{k=1}^{K}L(c_k,f(X))P(c_k\mid X)$ 对 $X=x$ 逐个极小化，可得：

$$
\begin{aligned}
&f(x)\\
&=\arg\min_{y\in Y}\sum_{k=1}^{K}L(c_k,y)P(c_k\mid X=x)\\
&=\arg\min_{y\in Y}\sum_{k=1}^{K}L(y\neq c_k\mid X=x)\\
&=\arg\min_{y\in Y}\left(1-P(y=c_k\mid X=x)\right)\\
&=\arg\max_{y\in Y}P(y=c_k\mid X=x)\\
&=\arg\max_{y\in Y}P(c_k\mid X=x)
\end{aligned}
$$

故根据期望风险最小化准则可以得到后验概率最大化准则。

3. 算法过程[1]

由上两小节我们知道了朴素贝叶斯法的构建思路与数学原理，那么这一节就介绍朴素贝叶斯法的具体实现流程。

1）概念明晰

$T=\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$ 为训练数据集，其中 $x_i=(x_i^{(1)},x_i^{(2)},...,x_i^{(n)})^T$，$x_i^{(j)}$ 是第 i 个样本的第 j 个特征，$x_i^{(j)}\in\{a_{j1},a_{j2},...,a_{jS_j}\}$，$a_{il}$ 是第 j 个特征可能取的第 l 个值，$j=1,2,..,n,l=1,2,...,S_j,y_i\in\{c_1,c_2,...,c_k\}$

2）计算先验概率及条件概率

计算先验概率及条件概率采用的是极大似然估计。

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N},k=1,2,...,K$$

$$P(X^{(j)}=a_{il}|Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{il},y_i=c_k)}{N}$$

$$j=1,2,..,n,l=1,2,...,S_j,k=1,2,...,K$$

3）对于所给的实例 $x_i=(x_i^{(1)},x_i^{(2)},...,x_i^{(n)})^T$，计算

$$P(Y=c_k)\prod_{j=1}^{n} P\left(X^{(j)}=x^{(j)}|Y=c_k\right),k=1,2,...,K$$

4）确定实例 x 的类

$$y=\arg\max_{c_k} P(Y=c_k)\prod_{j=1}^{n} P(X^{(j)}=x^{(j)}|Y=c_k)$$

思考 如何证明 3.2）中对先验概率和条件概率的估计是极大似然估计？

思考 极大似然估计可能会出现概率值为 0 的情况，此时会影响到后验概率的计算，如何解决这一问题？

思考 如果特征属性是连续的，还能使用朴素贝叶斯法吗？

一 高斯贝叶斯

三． 贝叶斯网络(Bayesian network)[3]

事实上，生活中大部分时候，不同的变量之间会有较强的关系，因此在实际操作中，朴素贝叶斯法的条件独立性假设难以成立，对此，有一种更为普适的基于贝叶斯公式的分类方法，就是贝叶斯网络。

贝叶斯网络本质上是以有向无环图形式表示不确定性的因果推理模型。具体而言，一个贝叶斯网络定义包括一个有向无环图（Directed acyclic graph,也称DAG）和一个条件概率表集合。

一般而言，贝叶斯网络的有向无环图中的节点表示随机变量，它们可以是可观察到的变量，抑或是隐变量、未知参数等。连接两个节点的箭头代表此两个随机变量是具有因果关系或是非条件独立的；而节点中变量间若没有箭头相互连接一起的情况就称其随机变量彼此间为条件独立。若两个节点间以一个单箭头连接在一起，表示其中一个节点是“因(parents)”，另一个是“果(descendants or children)”，两节点就会产生一个条件概率值。条件概率表中的每一个元素对应DAG 中唯一的节点，存储此节点对于其所有直接前驱节点的联合条件概率。

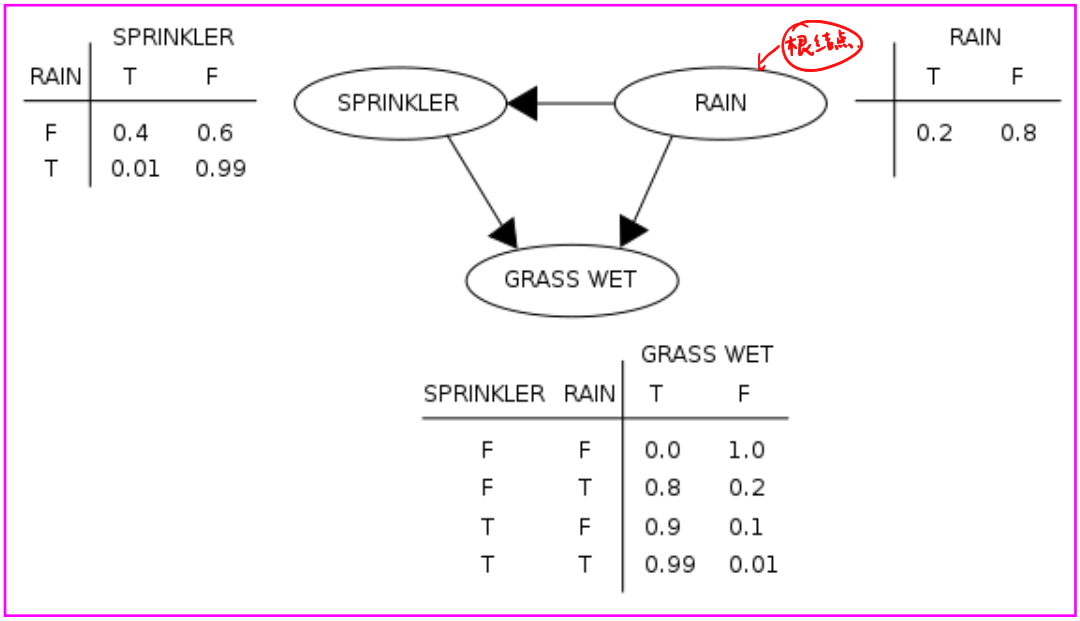
这个定义比较抽象，我们来结合一个实例理解一下。

EXAMPLE

我们考虑三个事件：洒水器(SPRINKLER)洒水,下雨(RAIN),草地湿了(GRASS WET)。然后我们要研究的是，当草地处于不同的状态时，判断洒水器的状态和天气是否下雨。

如果按照朴素贝叶斯方法，我们应当假设这三者之间互相独立，但是实际上这三者之间会具有明显的相关性，那么这时候就不再采用朴素贝叶斯方法，而采用贝叶斯网络进行分类。根据经验我们可以画出如下的 DAG 图，其中箭头的指向代表了因果关系的影响。用自然语言解释一下图中的因果链就是：下雨和洒水器洒水都很可能会导致草地湿，而下雨的时候洒水器一般而言不会打开。

（实际上贝叶斯网络并不一定要求是因果关系，理论上只要非条件独立就可以，但是在实际操作中大部分贝叶斯网络都采用因果关系进行解释。）



RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

图 1

一般情况下，多变量非独立联合条件概率分布有如下求取公式：

$$P(x_1,x_2,...,x_n)=P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)...P(x_n|x_1,x_2,...,x_{n-1})$$

为了减少计算量，贝叶斯网络在此处引入了一个重要的假设，或者说性质：

每一个节点在其直接前驱节点的值制定后，这个节点条件独立于其所有非直接前驱前辈节点。

由此，任意随机变量的联合条件概率分布简化为：

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

其中 Parents 表示 x_i 的直接前驱节点（因节点）的联合，因此可以通过每个节点的条件概率表求得。

为了理解这一公式，我们仍用图 1 的例子来加以说明。

EXAMPLE

图 1 中列出了每个节点相对于因节点的条件概率值表，假设已知草地是湿的，我们现在利用上述公式来判断天气为下雨的概率。

$$\Pr(R=T|G=T)=\frac{\Pr(G=T,R=T)}{\Pr(G=T)}=\frac{\sum_{S\in\{T,F\}}\Pr(G=T,S,R=T)}{\sum_{S,R\in\{T,F\}}\Pr(G=T,S,R)}$$

分子和分母可以分别用条件概率公式求出，比如

$$\begin{aligned}\Pr(G=T,S=T,R=T)&=\Pr(G=T|S=T,R=T)\Pr(S=T|R=T)\Pr(R=T)\\&=0.99\times0.01\times0.2\\&=0.00198.\end{aligned}$$

于是最后可以得到：

$$\Pr(R=T|G=T)=\frac{0.00198_{TTT}+0.1584_{TFT}}{0.00198_{TTT}+0.288_{TTF}+0.1584_{TFT}+0.0_{TFF}}=\frac{891}{2491}\approx35.77\%$$

上面这个例子是非常简单的，一是因为它的节点少，二是因为它的结构明确，三是因为它不存在隐藏节点。对于最后一点，含有隐藏节点的 DAG，由于隐藏节点无法观测，故无法直接列出其条件概率表，此时需要借助更为复杂的数学方法（如梯度下降搜索法）进行计算，在这里就不加以详细说明，有兴趣的同学可以自行查阅相关资料。

思考 除了贝叶斯网络，还有没有其他可以用于解决非条件独立情况的办法？

- 半朴素贝叶斯算法.
- 将相关的特征强行归为一类.

Reference
参考阅读
[1]李航，统计学习方法[M].北京：清华大学出版社，2012:48-53
[2]李航，统计学习方法[M].北京：清华大学出版社，2012:7-8
[3] https://en.wikipedia.org/wiki/Bayesian_network