《数据挖掘技术》

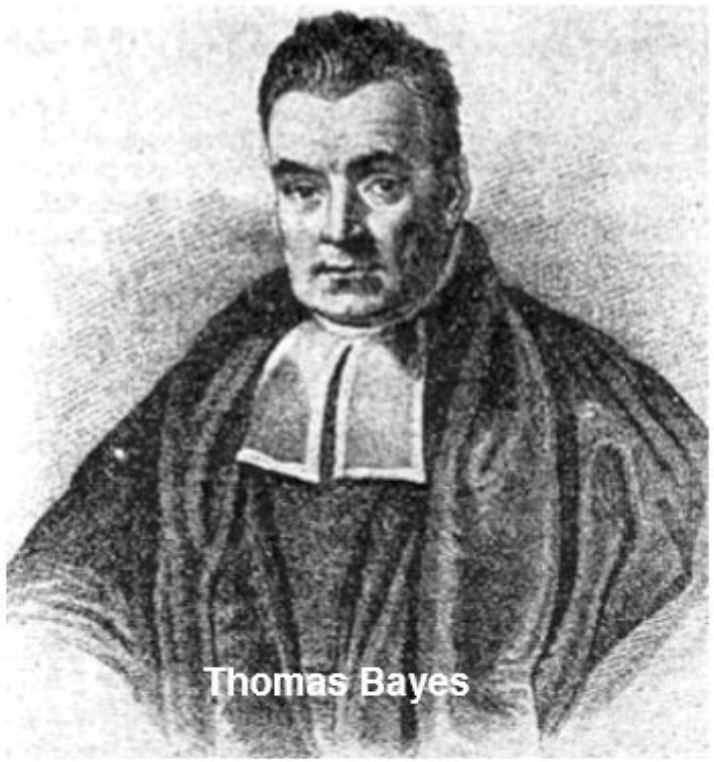
⭐ CH04 朴素贝叶斯法

➡️ Created by *Wang JingHui*

➡️ Version: *4.0*

About Bayes Theorem (Bayes‘ rule)

Bayes theorem named after Rev. Thomas Bayes. It works on **conditional probability**. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the conditional probability, we can calculate the probability of an event using its prior knowledge[1].



It tells us how often A happens given that B happens

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

- **P (A I B)** is “Probability of A given B”, the probability of A given that B happens.
- **P (A)** is Probability of A
- **P (B I A)** is “Probability of B given A”, the probability of B given that A happen
- **P (B)** is Probability of B

For example:

- P(Fire I Smoke) means how often there is fire when we see smoke.
- P(Smoke I Fire) means how often we see smoke when there is fire.

贝叶斯(约1701-1761)

- Thomas Bayes，英国数学家。约1701年出生于伦敦，做过神甫。
- 1742年成为英国皇家学会会员。1761年4月7日逝世。
- 贝叶斯在数学方面主要研究概率论。他首先将归纳推理法用于概率论基础理论，并创立了贝叶斯统计理论，对于统计决策函数、统计推断、统计的估算等做出了贡献。
- 他死后，理查德·普莱斯(Richard Price)于1763年将他的著作《机会问题的解法》(An essay towards solving a problem in the doctrine of chances)寄给了英国皇家学会，对于现代概率论和数理统计产生了重要的影响

主要内容

1. 朴素贝叶斯法的学习与分类
    i. 基本方法
    ii. 后验概率最大化的含义
2. 朴素贝叶斯法的参数估计
    i. 极大似然估计
    ii. 学习与分类算法
    iii. 贝叶斯估计

贝叶斯分类方法的思想

- 对于给出的特征，求解在此特征出现的条件下各个类别出现的概率中，哪个可能性最大，就认为此待分类的特征项属于哪个类别。
- 通俗来说，比如：你在学校看到一个长头发的人，你十有八九猜是个女生。为什么呢？因为长头发在女生的比率最高，当然人家也可能是男生。但在没有其它可用信息下，我们会选择条件概率最大的类别，这就是朴素贝叶斯的思想基础。

为了估计状态变量的条件分布, 利用贝叶斯法则, 有

$$\underbrace{P(X|Y)}_{posterior} = \frac{\overbrace{P(Y|X)}^{likelihood}\overbrace{P(X)}^{prior}}{\underbrace{P(Y)}_{evidence}}$$

其中$P(X|Y)$为后验概率(Posterior), $P(Y|X)$称为似然, $P(X)$称为先验(Prior).

朴素贝叶斯法的学习与分类

基本方法

输入空间$\mathcal{X} \subseteq R^n$为$n$维向量的集合；

输出空间为类标记集合$y = \{c_1, c_2, ..., c_k\}$

- X是定义在输入空间$\mathcal{X}$上的随机向量;
- Y是定义在输出空间$\mathcal{Y}$上的随机变量;
- $P(X, Y)$是X, Y的联合概率分布。

目标（未知）

- 求出$P(Y|x)$的概率，即$P(c_1|x),...,P(c_k|x)$的概率，从中找到概率最大的$P(c_i|x)$，c_i作为预测目标。

先验概率分布（已知）

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

条件概率分布（已知）

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

$$k=1,2,...,K$$

输入

- 训练数据集$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$，$x_i \in \mathcal{X} \subseteq \mathbf{R^n}$ $y_i \in \mathcal{Y} = \{c_1, c_2, \ldots, c_k\}$;
- 实例特征向量x

输出:

- 学习到联合概率分布$P(X, Y)$或$P(Y|X)$

算法推导

朴素贝叶斯法是基于**贝叶斯定理**与**特征条件独立假设**的分类方法.

贝叶斯定理

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

$$P(A|B)P(B) = P(AB) = P(B|A)P(A)$$

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

换个表示方法

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

例： 一个特征情况，A 是类别，如{女生,男生},B是特征，如{长头发}

已知：

$$P(\text{长头发}|\text{女生}), P(\text{长头发}|\text{男生}), P(\text{女生}), P(\text{男生})$$

计算：

$$P(\text{女生}|\text{长头发}) = \frac{P(\text{长头发}|\text{女生})P(\text{女生})}{P(\text{长头发})}$$

$$P(\text{男生}|\text{长头发}) = \frac{P(\text{长头发}|\text{男生})P(\text{男生})}{P(\text{长头发})}$$

在贝叶斯定理中,每个名词都有约定俗成的**名称**:

- $P(A|B)$是已知B发生后A的条件概率,也由于得知B的取值而被称作A的后验概率;
- $P(A)$是A的先验概率(或边缘概率).之所以称为"先验"是因为它不考虑任何B方面的因素;
- $P(B|A)$是已知A发生后B的条件概率,也由于得知A的取值而成称作B的后验概率;
- $P(B)$是B的先验概率(或边缘概率).

条件独立假设

求$P(Y|X)$，其中$X \in \{X_1, X_2, \ldots, X_n\}$，条件独立假设这里给定$Y$的情况下：

1. 每一个X_i和其他的每个X_k是条件独立的
2. 每一个X_i和其他的每个X_k的子集是条件独立的

概率乘法定理

$$
\begin{aligned}
P(X|Y) &= P(X_1, X_2|Y) \\
&= P(X_1|X_2, Y)P(X_2|Y) \\
&= P(X_1|Y)P(X_2|Y)
\end{aligned}
$$

红色部分从上到下基于I.I.D.

条件独立假设等于是说用于分类的**特征**在**类确定**的条件下都是**条件独立**的.

条件独立性假设是:

$$P(X=x|Y=c_k)=P(X^{(1)},\dots,X^{(n)}|Y=c_k)$$
$$=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

条件独立性假设

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

即，用于分类的特征在类确定的条件下都是条件独立的。

例： 特征变成两个，出现条件独立。

$$P(\text{女生}\big|\text{长头发},\text{穿裙子}) = \frac{P(\text{长头发},\text{穿裙子}\big|\text{女生})P(\text{女生})}{P(\text{长头发},\text{穿裙子})} = \frac{P(\text{长头发}\big|\text{女生})P(\text{穿裙子}\big|\text{女生})P(\text{女生})}{P(\text{长头发},\text{穿裙子}|\text{女生}) + P(\text{长头发},\text{穿裙子}|\text{男生})}$$

$$P(\text{男生}\big|\text{长头发},\text{穿裙子}) = \frac{P(\text{长头发},\text{穿裙子}\big|\text{男生})P(\text{男生})}{P(\text{长头发},\text{穿裙子})} = \frac{P(\text{长头发}\big|\text{男生})P(\text{穿裙子}\big|\text{男生})P(\text{男生})}{P(\text{长头发},\text{穿裙子}|\text{女生}) + P(\text{长头发},\text{穿裙子}|\text{男生})}$$

如果特征独立，则：

$$P(\text{长头发},\text{穿裙子}) = P(\text{长头发}) * P(\text{穿裙子})$$

证明:

由

$$P\left(X=x, Y=c_k\right)=P\left(X=x \mid Y=c_k\right) P\left(Y=c_k\right)$$
$$P\left(X=x, Y=c_k\right)=P\left(Y=c_k \mid X=x\right) P\left(X=x\right)$$

类比于：$P(A|B)P(B)=P(AB)=P(B|A)P(A)$

得

$$P\left(X=x|Y=c_k\right)P\left(Y=c_k\right)=P\left(Y=c_k|X=x\right)P\left(X=x\right)$$

$$P\left(Y=c_k|X=x\right)=\frac{P\left(X=x|Y=c_k\right)P\left(Y=c_k\right)}{P\left(X=x\right)}$$

$$=\frac{P\left(X=x|Y=c_k\right)P\left(Y=c_k\right)}{\sum_Y P\left(X=x,Y=c_k\right)}$$

$$=\frac{P\left(X=x|Y=c_k\right)P\left(Y=c_k\right)}{\sum_Y P\left(X=x|Y=c_k\right)P\left(Y=c_k\right)}$$

$$=\frac{P\left(Y=c_k\right)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}{\sum_Y P\left(Y=c_k\right)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}$$

朴素贝叶斯分类器可表示为

$$
\begin{aligned}
y = f(x) &= \arg\max_{c_k} \frac{P(Y=c_k)\prod_{j=1}^{n} P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}{\sum_{Y} P(Y=c_k)\prod_{j=1}^{n} P\left(X^{(j)}=x^{(j)}|Y=c_k\right)} \\
&= \arg\max_{c_k} P(Y=c_k)\prod_{j=1}^{n} P\left(X^{(j)}=x^{(j)}|Y=c_k\right)
\end{aligned}
$$

后验概率最大化的含义

- 朴素贝叶斯法将实例分到**后验概率最大的类**中, 这等价于**期望风险最小化**. 只需对 $X = x$ 逐个极小化.

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

推导后验概率最大化准则:

$$
f(x) = \arg\max_{c_k} (P(c_k | X = x))
$$

朴素贝叶斯法的参数估计

极大似然估计

- 估计先验；
- 估计条件概率；

朴素贝叶斯算法：

输入： 线性可分训练数据集

$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中$x_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)^T$，

$x_i^{(j)}$是第i个样本的第j个特征，$x_i^{(j)}\in\left\{a_{j1},a_{j2},\cdots,a_{jS_j}\right\}$，$a_{jl}$是第$j$个特征可能取的第$l$个值，$j=1,2,\cdots,n;l=1,2,\cdots,S_j,y_i\in\{c_1,c_2,\cdots,c_K\}$；

实例x；

输出： 实例x的分类

1. 计算先验概率及条件概率

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

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

$$j=1,2,\cdots,n; \quad l=1,2,\cdots,S_j; \quad k=1,2,\cdots,K$$

2. 对于给定的实例

$$x=\left(x^{(1)}, x^{(2)}, \cdots, x^{(n)}\right)^{T}$$

计算

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

3. 确定实例x的类别

$$y=f(x)=\arg \max _{c_{k}} P\left(Y=c_{k}\right) \prod_{j=1}^{n} P\left(X^{(j)}=x^{(j)} | Y=c_{k}\right)$$

贝叶斯估计

- 对于x的某个特征的取值没有在先验中出现的情况, 如果用极大似然估计, 这种情况的可能性就是0;
- 但是出现这种情况的原因通常是因为数据集不能全覆盖样本空间;
- 出现未知的情况处理的策略就是做平滑.

朴素贝叶斯模型参数的贝叶斯估计

1. 条件概率的贝叶斯估计

$$P_\lambda\left(X^{(j)}=a_{jl}|Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(x_i^{(j)}=a_{jl},y_i=c_k\right)+\lambda}{\sum_{i=1}^{N}I\left(y_i=c_k\right)+S_j\lambda}$$

式中$\lambda\geq 0$。当$\lambda=0$时，是极大似然估计；当$\lambda=1$时，称为拉普拉斯平滑,拉普拉斯平滑相当于给未知变量给定了先验概率.。

2. 先验概率的贝叶斯估计

$$P\left(Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(y_i=c_k\right)+\lambda}{N+K\lambda}$$

例子

Day	Outlook	Temperature	Humidity	Wind	Play Tennis ?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

there are 5 cases of not being able to play a game, and 9 cases of being able to play a game.

If we were given a new instance :

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

We want to know if we can play a game or not ?

Example : Can we play tennis today ?

If X = (Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong)， then

P (Play=Yes I X) = P (Play=Yes | Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong)

$$= \frac{P(\text{Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong I Play=Yes}) * P(\text{Play=Yes})}{P(\text{Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong})}$$

$$= \frac{P(\text{Outlook = Sunny I Play=Yes}) * P(\text{Temperature = Cool I Play=Yes}) * P(\text{Humidity = High I Play=Yes}) * P(\text{Wind = Strong I Play=Yes}) * P(\text{Play=Yes})}{P(\text{Outlook=Sunny}) * P(\text{Temperature=Cool}) * P(\text{Humidity=High}) * P(\text{Wind=Strong})}$$

$$= \frac{(2/9) * (3/9) * (3/9) * (3/9) * (9/14)}{(5/14) * (4/14) * (7/14) * (6/14)}$$

$$= \frac{0.0053}{0.02186} = \mathbf{0.2424}$$

P (Play= No I X) = P (Play= NO | Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong)

$$= \frac{(3/5) * (1/5) * (4/5) * (3/5) * (5/14)}{(5/14) * (4/14) * (7/14) * (6/14)} = \frac{0.0206}{0.02186} = \mathbf{0.9421}$$

- P(Play=Yes I X) = 0.2424
- P(Play=No I X) = 0.9421

Since 0.9421 is greater than 0.2424 then the answer is 'no', we cannot play a game of tennis today.

Outlook	Play = Yes	Play = No	Total
Sunny	2/9	3/5	5/14
Overcast	4/9	0/5	4/14
Rain	3/9	2/5	5/14

Temperature	Play = Yes	Play = No	Total
Hot	2/9	2/5	4/14
Mild	4/9	2/5	6/14
Cool	3/9	1/5	4/14

Humidity	Play = Yes	Play = No	Total
High	3/9	4/5	7/14
Normal	6/9	1/5	7/14

Wind	Play = Yes	Play = No	Total
Strong	3/9	3/5	6/14
Weak	6/9	2/5	8/14

If we were given a new instance :

X = (**Outlook** = Sunny, **Temperature** = Cool, **Humidity** = High, **Wind** = Strong), can we play the game ?

Firstly we look at the probability that we can play the game

- P (Outlook=Sunny | Play=Yes) = 2/9
- P (Temperature=Cool | Play=Yes) = 3/9
- P (Humidity=High | Play=Yes) = 3/9
- P (Wind=Strong | Play=Yes) = 3/9
- P (Play=Yes) = 9/14

Next we consider the fact that we cannot play a game:

- P (Outlook=Sunny | Play=No) = 3/5
- P (Temperature=Cool | Play=No) = 1/5
- P (Humidity=High | Play=No) = 4/5
- P (Wind=Strong | Play=No) = 3/5
- P (Play=No) = 5/14

参考

1. [^1]: [视觉SLAM十四讲, 高翔](## 参考)
2. [^2]: Generative and discriminative classifiers: Naive Bayes and logistic regression
3. [^3]: Machine Learning New Chapter
4. [^4]: An Introduction to Conditional Random Fields for Relational Learning

Enjoy your machine learning!

https://github.com/wjssx/

E-mail: csr_dsp@sina.com