LiVE Group
视觉智能与学习研究中心

机器学习（第7讲）

主讲：张 磊

E-mail: leizhang@cqu.edu.cn
Lab Website: http://www.leizhang.tk

第七章：朴素贝叶斯学习

贝叶斯学习

- **Thomas Bayes的“逆概”问题。**

正向概率：假设袋子里面有N个白球，M个黑球，你伸手进去摸一把，摸出黑球的概率是多大？

逆向概率：如果我们事先并不知道袋子里面黑白球的比例，而是闭着眼睛摸出一个（或好几个）球，观察这些取出来的球的颜色之后，那么我们可以就此对袋子里面的黑白球的比例作出什么样的推测？

解决这个问题要靠什么？

一个字“猜”！

贝叶斯学习

- 贝叶斯学习是一种基于概率的学习方法，能够计算显式的假设概率。它基于（1）假设的先验概率，（2）给定假设下观察到不同数据的概率，以及（3）观察到数据本身的概率。

1. **P(h)**: 表示没有训练样本数据前，假设h拥有的初始概率，称为假设h的先验概率。如果，我们不清楚这个先验知识，在实际中通常简单地将每一种假设都赋予相同的概率。
2. **P(D)**: 表示观察到训练数据D的先验概率，通常作为常数处理，因为不同的假设拥有相同的P(D)。
3. **P(D|h)**: 表示假设h成立时观察到数据D的概率。
4. **P(h|D)**: 后验概率，即对于给定的一个训练样本D，假设h成立的概率。在机器学习中，我们感兴趣的是这个后验概率。

第七章：朴素贝叶斯学习

◘ 贝叶斯学习

贝叶斯公式：

$$P(h|D) = \frac{P(h) * P(D|h)}{P(D)}$$

这个公式是怎么来的？

例：一所学校里面有 60% 的男生，40% 的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。有了这些信息之后我们可以容易地计算“随机选取一个学生，他（她）穿长裤的概率和穿裙子的概率是多大”，这个就是前面说的“正向概率”的计算。然而，假设你走在校园中，迎面走来一个穿长裤的学生，你能够推断出他（她）是男生的概率是多大吗？

分析：如果知道有多少个穿长裤的学生，并知道穿长裤的人里面有多少男生就可以了。

设学校里面共有N人，其中：

- **穿长裤的男生的数量是N*P(男)*P(裤|男)**
- **穿长裤的女生的数量是N*P(女)*P(裤|女)**

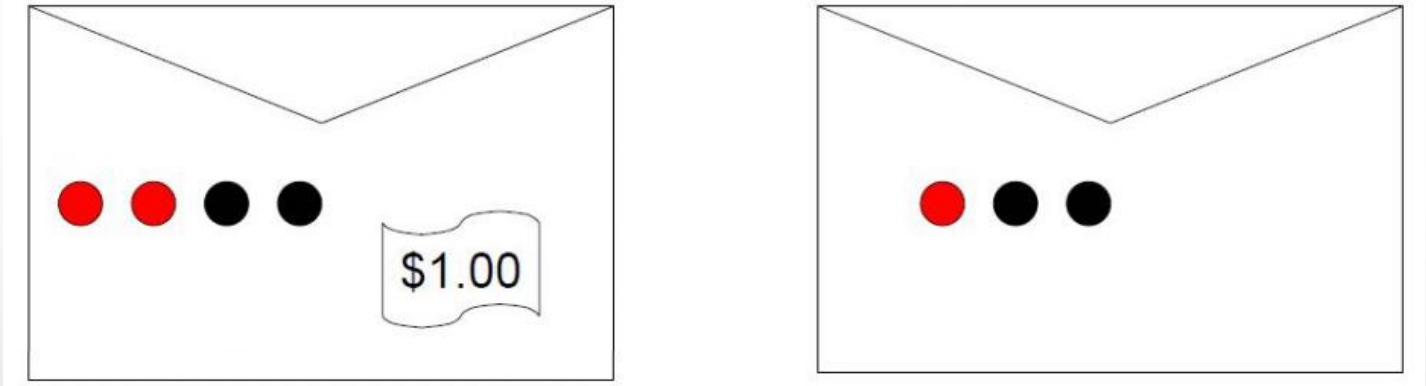
那么，P(男|裤)=N*P(男)*P(裤|男)/[N*P(男)*P(裤|男)+ N*P(女)*P(裤|女)]=P(男)*P(裤|男)/P(裤)

通式：P(B|A)=P(B)*P(A|B)/P(A)

贝叶斯学习

贝叶斯公式：

$$P(h|D) = \frac{P(h) * P(D|h)}{P(D)}$$



- 如果摸到一个红球，那么，这个信封有1美元的概率是？
- 如果摸到一个黑球，那么，这个信封有1美元的概率是？

逆概问题

第七章：朴素贝叶斯学习

贝叶斯学习

- c1、c2表示左右两个信封。
- P(R)，P(B)表示摸到红球、黑球的概率。
- P(R)=P(R|c1)*P(c1) + P(R|c2)*P(c2)：全概率公式
- 后验：P(c1|R)=P(R|c1)*P(c1)/P(R)
- P(R|c1)=2/4
- P(R|c2)=1/3
- P(c1)=P(c2)=1/2

如果摸到一个红球，那么，这个信封有1美元的概率是
P(c1|R)=P(R|c1)*P(c1)/P(R)=0.6

如果摸到一个黑球，那么，这个信封有1美元的概率是
P(c1|B)=P(B|c1)*P(c1)/P(B)=3/7

贝叶斯学习

贝叶斯公式：

$$P(h|D) = \frac{P(h) * P(D|h)}{P(D)}$$

贝叶斯学习的目的是对于训练样本D，找出一个最靠谱的猜测h (假设)，使得后验概率达到最大。

也就是计算下列模型的最优解。

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

其中H是各种“猜测”的集合。h_{MAP}是最佳的“最靠谱”的猜测，满足后验最大。

将上面的模型写完整：

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D) = \underset{h \in H}{\operatorname{argmax}} \frac{P(h) * P(D|h)}{P(D)} = \underset{h \in H}{\operatorname{argmax}} P(h) * P(D|h)$$

注：P(D)是不依赖于h的常量。

贝叶斯学习

最大后验假设：在许多学习场景中，学习器考虑候选假设集合H，并在其中寻找给定数据D时，可能性最大的假设h∈H,即为最大后验假设（Maximum a Posteriori, MAP），也称为MAP假设，定义为h_{MAP}。

确定MAP假设h_{MAP}的方法，是利用贝叶斯公式计算集合H中每个候选假设的后验概率。即

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D) = \underset{h \in H}{\operatorname{argmax}} \frac{P(h) * P(D|h)}{P(D)}$$

贝叶斯学习

贝叶斯模型：

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h) * P(D|h)$$

- 当设计者不清楚先验时，可将所有假设的先验P(h)均设为相同的值，即$P(h) = \frac{1}{H}$，此时，模型变成了$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(D|h)$，P(D|h)常被称为给定h时，数据D的似然，而使 P(D|h)最大的假设，称为极大似然假设。
- 可以看出，贝叶斯模型实际上是先验与数据似然函数的乘积。如何理解先验？

先验在很多时候，能够提供有力的“人为”帮助，提升模型的分类能力和鲁棒。

模型的抽象含义是：对于给定观测数据D，一个猜测h是好是坏，取决于“这个猜测本身的可能性大小（先验概率$P(h)$，Prior ）”和“这个猜测生成我们观测到的数据的可能性大小”（似然$P(D|h)$，Likelihood ）的乘积

贝叶斯学习

先验的理解

例：用户输入一个错误单词Thet，请判断用户实际真正想输入的单词：“They” or “That”？

分析：假设“They”的后验概率P(They|Thet)=P(Thet|They)*P(They)/P(Thet)；

假设“That”的后验概率P(That|Thet)=P(Thet|That)*P(That)/P(Thet);

其中，P(They)和P(That)为假设的先验概率。

1. **若根据实际使用频率，单词That的使用频率要高于They, 也就是说，用户实际想输入That的可能性更高，即 P(They)<P(That)**
2. **若根据键盘中字母的排列，字母t和字母y是紧密靠近的，因此，用户实际想输入they的可能性更高， 即P(They)>P(That)**

根据不同的先验知识，先验概率的值不同，因此先验知识的形式是多样的。

朴素贝叶斯学习

朴素贝叶斯到底“朴素”在哪？也就是朴素贝叶斯的假设是，当类假设h给定时，特征之间相互条件独立。

换句话说，在给定假设h的情况下，观察到联合的d1,d2,…,dn的概率等于单独特征（属性）的概率乘积：

$$P(d_1, d_2, \dots, d_n|h) = \prod_i P(d_i|h)$$

代入到贝叶斯模型中，有

$$h_{MAP} = \underset{h\in H}{\operatorname{argmax}} P(h|D) = \underset{h\in H}{\operatorname{argmax}} \frac{P(h) * \prod_i P(d_i|h)}{P(D)} = \underset{h\in H}{\operatorname{argmax}} P(h) * \prod_i P(d_i|h)$$

“朴素”即“特征独立假设”的好处就是大大降低了估计带来计算复杂度。

朴素贝叶斯学习

目标值PlayTennis的14个训练样例

Day	Outlook	Temperatu	Humidity	Wind	PlayTenni
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

问题：结合朴素贝叶斯方法，
对下面的新实例（4个特征属性）
进行分类：

Outlook=**Sunny**
Temperature=**Cool**
Humidity=**High**
Wind=**Strong**

PlayTennis=Yes or No?

第七章：朴素贝叶斯学习

- 朴素贝叶斯学习

$$h_{MAP} = \underset{h \in H=\{Yes,No\}}{\operatorname{argmax}} P(h) * \prod_i P(d_i|h)$$

分析：结合朴素贝叶斯分类器模型，分别把h=Yes和h=No两种假设下的后验概率计算出来，最大值对应的假设就是我们的答案。

- 当h=Yes时，

$$P(h) * \prod_i P(d_i|h) = P(Yes) * P(Sunny|Yes) * P(Cool|Yes) * P(High|Yes) * p(Strong|Yes)$$

$$= \frac{9}{14} * \frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} = 0.0053$$

- 当h=No时，

$$P(h) * \prod_i P(d_i|h) = P(No) * P(Sunny|No) * P(Cool|No) * P(High|No) * p(Strong|No)$$

$$= \frac{5}{14} * \frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} = 0.0206$$

显然，$P(No) * \prod_i P(d_i|No) = 0.0206 > P(Yes) * \prod_i P(d_i|Yes) = 0.0053$，所以，PlayTennis=No,今天的天气不适合打网球

贝叶斯学习方法的特性

1. 贝叶斯方法允许假设做出不确定性的预测（如：某一癌症病人有85%的机会康复）。
2. 先验知识可以与观察数据一起决定假设的最终概率。在贝叶斯学习中，先验知识的形式可以是：1）每个候选假设的先验概率；2）每个候选假设在观察数据上的概率分布。
3. 新的样本分类可以由多个假设一起作出预测，用它们的概率来加权（最优贝叶斯)。

◘ 采用朴素贝叶斯对垃圾邮件进行分类

1. **样本**：1000封邮件，每个邮件被标记为垃圾邮件或者正常邮件。设有400封垃圾邮件，600封正常邮件。
2. **分类目标（问题）**：给定第1001封邮件，确定它是垃圾邮件还是非垃圾邮件？
3. **方法**：朴素贝叶斯。

该如何建模和实现？

（一）分析：令待分类的这封邮件为$D=[d_1,d_2,\ldots,d_n]$，由n个单词组成。垃圾邮件定义为h1, 正常邮件定义为h2。设400封垃圾邮件中的单词总数为N1，600封正常邮件中的单词总数为N2。

采用朴素贝叶斯对垃圾邮件进行分类

（二）如果能够计算出垃圾邮件h1假设的后验概率和正常邮件h2假设的后验概率，两者概率最大的假设，即为该封邮件的类别。采用贝叶斯公式。

数学描述：$P(h_1|D)=\frac{P(D|h_1)P(h_1)}{P(D)}$, $P(h_2|D)=\frac{P(D|h_2)P(h_2)}{P(D)}$

P(D)为与h无关的常量，因此，只需计算$P(D|h_1)P(h_1)$和$P(D|h_2)P(h_2)$即可。

◘ 采用朴素贝叶斯对垃圾邮件进行分类

（三）为了计算$P(D|h_1)P(h_1)$和$P(D|h_2)P(h_2)$，可以先计算先验概率$P(h_1)$和$P(h_2)$。

$$P(h_1)=\frac{400}{1000}=0.4;\ P(h_2)=\frac{600}{1000}=0.6$$

那如何计算$P(D|h_1)$和$P(D|h_2)$？

$$P(D|h_1)=P(d_1,d_2,\cdots,d_n|h_1)$$

然后呢？问题变得复杂。由于数据的稀疏性以及语言的多样性，完全出现两封完全相同的邮件（文本）是几乎不可能的。也就是说，n个单词$(d_1,d_2,\cdots,d_n)$在所有邮件中同时出现的概率几乎为0。那怎么办？

◘ 采用朴素贝叶斯对垃圾邮件进行分类

（四）当给定某假设时，采用特征的条件独立性假设，即朴素贝叶斯。

那如何计算$P(D|h_1)$和$P(D|h_2)$？

$$P(D|h_1) = P(d_1, d_2, \cdots, d_n|h_1) = P(d_1|h_1) * P(d_2|h_1) * \cdots * P(d_n|h_1)$$

$$= \prod_{i=1}^{n} P(d_i|h_1)$$

现在问题变得非常简单，计算$P(d_i|h_1)$的概率，只需要统计单词d_i在400封垃圾邮件中出现的频次，然后相乘。同理，$P(D|h_2)$的概率也是这么计算。

朴素贝叶斯分类器

方法的基本定义：

1. 设$\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_d\}$为一个待分类的样本，由d个特征属性构成；
2. 有类别集合$C = \{y_1, y_2, \ldots, y_C\}$；
3. 计算$P(y_1|\mathbf{x}), P(y_2|\mathbf{x}), \ldots, P(y_C|\mathbf{x})$；
4. 如果$P(y_k|\mathbf{x}) = \max\{P(y_1|\mathbf{x}), P(y_2|\mathbf{x}), \ldots, P(y_C|\mathbf{x})\}$，则x属于$y_k$这个类别。

在以上4个步骤中，只要成功完成第3步，即可实现样本x的分类。

如何实现第3步？

利用前面所学习的朴素贝叶斯学习算法！

朴素贝叶斯分类器

第3步计算$P(y_1|\mathbf{x}), P(y_2|\mathbf{x}),\ldots, P(y_C|\mathbf{x})$条件概率的具体步骤：

1. 已知类别的训练样本集（如14天的天气数据，以及是否打了网球）
2. 统计每个类别下，各个特征属性的条件概率（出现频次），即

$$P(x_1|y_1), P(x_2|y_1), \ldots, P(x_d|y_1);$$
$$P(x_1|y_2), P(x_2|y_2), \ldots, P(x_d|y_2);$$
$$\vdots$$
$$P(x_1|y_C), P(x_2|y_C), \ldots, P(x_d|y_C);$$

3. 计算各类别的先验概率$P(y_1), P(y_2), \ldots, P(y_C)$（统计各类别出现的频率）
4. 根据朴素贝叶斯方法，各个特征属性是条件独立的。利用贝叶斯定理(公式)，

我们要计算的$P(y_i|\mathbf{x}) = \frac{P(\mathbf{x}|y_i)P(y_i)}{P(\mathbf{x})} = \frac{P(x_1|y_i)\cdot P(x_2|y_i)\cdots P(x_d|y_i)\cdot P(y_i)}{P(\mathbf{x})} = \frac{P(y_i)\cdot\prod_{j=1}^{d} P(x_j|y_i)}{P(\mathbf{x})}$，

其中，P(x)是数据的先验，可以当做常数处理，不影响后验概率的比较。

朴素贝叶斯分类器

✓ Laplacian校验(零概率问题)

当特征属性为离散值时，计算条件概率$P(x_j|y_i)$即在类y_i条件下统计属性x_j出现的次数，可能会遇到出现次数为0，从而导致该条件概率为0，会影响分类器的性能（训练样本较少时会遇到该问题）。

这种情况下，可以利用Laplacian校验，即统计每一类下的某特征属性出现的次数时，全部计数加1，可避免0概率的出现。

当训练样本数量充分大时，不影响分类器的结果。

朴素贝叶斯分类器

前面的例子中，考虑了特征值属性为离散值时的情况，在实际问题中，通常是特征值属性为连续值的情况。

那该如何计算完成上述条件概率的计算？即

$$P(x_1|y_i), P(x_2|y_i), \dots, P(x_d|y_i), \forall i = 1, \dots, C$$

- 当特征值为连续值时，通常假设特征属性服从高斯分布，即

$$P(\mathbf{x}|y_i) = N(\mu_i, \sigma_i)$$

因此，$P(x_j|y_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_j-\mu_i)^2}{2\sigma_i^2}}$

- 如何计算每一类各特征属性的均值以及方差？即μ_i和$\sigma_i (i = 1, \dots, C)$

直接利用期望和方差公式即可。

朴素贝叶斯分类器

最小错误率贝叶斯分类器

利用后验概率$P(y_i|\mathbf{x})$，当$P(y_k|\mathbf{x})$为所有后验概率中最大值时，待分类样本$\mathbf{x}$属于y_k类；即$P(y_k|\mathbf{x}) = \max\limits_{i=1,\ldots,c} P(y_i|\mathbf{x})$

最大似然比贝叶斯分类器

对于两类问题，当 $P(\mathbf{x}|y_i)P(y_i) > P(\mathbf{x}|y_j)P(y_j)$时，判决$\mathbf{x}$属于$y_i$类；也就是当 $\frac{P(\mathbf{x}|y_i)}{P(\mathbf{x}|y_j)} > \frac{P(y_j)}{P(y_i)}$时，判决$\mathbf{x}$属于$y_i$类。其中，$\frac{P(\mathbf{x}|y_i)}{P(\mathbf{x}|y_j)}$为似然比。$\frac{P(y_j)}{P(y_i)}$为判决门限。

最小风险贝叶斯分类器

朴素贝叶斯分类器

最小风险贝叶斯分类器

顾名思义，对于待分类样本，有多种决策，具有最小风险的决策（分类）即为最终的分类。作如下定义：

1. **决策**ω_i：把待识别样本x分类到y_i类中；
2. **损失**ρ_{ij}：把真实属于y_i类的样本x错误分类到y_j类中；
3. **条件风险**$R(\omega_i|\mathbf{x})$: 对待识别样本x采取决策ω_i，产生的可能风险；条件风险的计算公式如下：$R(\omega_i|\mathbf{x}) = \sum_{j=1}^{c} \rho_{ij} P(y_j|\mathbf{x})$
4. 最小风险贝叶斯分类器就是计算具有最小风险的决策ω_k

$$R(\omega_k|\mathbf{x}) = \min_{i=1,\ldots,c} R(\omega_i|\mathbf{x})$$

那么待识别样本x采取决策ω_k，**也就是将被分类到**y_k**类中。**

朴素贝叶斯分类器

最小风险贝叶斯分类器

当$\rho_{ij}=\begin{cases}0, i=j\\1, i\neq j\end{cases}$时，

有

$$\min_{i=1,\ldots,c} R(\omega_i|\mathbf{x})=\min_{i=1,\ldots,c}\sum_{j=1}^{C}\rho_{ij}P(y_j|\mathbf{x})=\min_{i=1,\ldots,c}\sum_{j=1,j\neq i}^{C}P(y_j|\mathbf{x})$$

$$=\min_{i=1,\ldots,c} 1-P(y_i|\mathbf{x})$$

注：$\sum_{j=1}^{C}P(y_j|\mathbf{x})=1$

$$=\max_{i=1,\ldots,c} P(y_i|\mathbf{x})$$

注：最小错误率贝叶斯分类器

可见，最小错误率分类器是最小风险贝叶斯分类器的一种特例。

朴素贝叶斯分类器

分类器性能评价

在机器学习中，基本上分类器的评价是通过分类器正确率(识别率)进行衡量。

分类器正确率：指被正确分类的样本数占所有样本数的比率。

分类正确率的计算通常是基于一部分新的测试样本集，而非训练样本集。

因为训练集往往会因为过拟合，而得出较高的正确率，但不能作为分类器的评价指标。