贝叶斯分类器

1.极大似然估计原理

最大似然估计的目的就是：利用已知的样本结果，反推最有可能（最大概率）导致这样结果的参数值。

极大似然估计是建立在极大似然原理的基础上的一个统计方法。极大似然估计提供了一种给定观察数据来评估模型参数的方法，即："模型已定，参数未知"。通过若干次试验，观察其结果，利用试验结果得到某个参数值能够使样本出现的概率为最大，则称为极大似然估计。

由于样本集中的样本都是独立同分布，可以只考虑一类样本集D，来估计参数向量$\vec{\theta}$。记已知的样本集为：

$$D = \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$$

似然函数（likelihood function）：联合概率密度函数$p(D|\vec{\theta})$称为相对于$\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$的$\vec{\theta}$的似然函数。

$$l(\vec{\theta}) = p(D|\vec{\theta}) = p(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n|\vec{\theta}) = \prod_{i=1}^{n} p(\vec{x}_i|\vec{\theta})$$

如果$\hat{\vec{\theta}}$是参数空间中能使似然函数$l(\vec{\theta})$最大的$\vec{\theta}$值，则$\hat{\vec{\theta}}$应该是"最可能"的参数值，那么$\hat{\vec{\theta}}$就是θ的极大似然估计量。它是样本集的函数，记作：

$$\hat{\vec{\theta}} = d(D) = \arg\max_{\vec{\theta}} l(\vec{\theta})$$

$\hat{\vec{\theta}}(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n)$称为极大似然函数估计值。

2.贝叶斯分类器基本原理

贝叶斯决策论通过**相关概率已知**的情况下，利用**误判损失**来选择最优的类别分类。

假设有N种可能的分类标记，记为$Y = \{c_1, c_2, \ldots, c_N\}$，那对于样本$\boldsymbol{x}$，考虑其类别：

step 1. 算出样本$\boldsymbol{x}$属于第i个类的概率，即$P(c_i|x)$；

step 2. 通过比较所有的$P(c_i|\boldsymbol{x})$，得到样本$\boldsymbol{x}$所属的最佳类别。

step 3. 将类别c_i和样本$\boldsymbol{x}$代入到贝叶斯公式中，得到：

$$P(c_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c_i)P(c_i)}{P(\boldsymbol{x})}.$$

$P(c_i)$为先验概率，$P(\boldsymbol{x}|c_i)$为条件概率，$P(\boldsymbol{x})$是用于归一化的证据因子。对于$P(c_i)$可以通过训练样本中类别为c_i的样本所占的比例进行估计；此外，由于只需要找出最大的$P(\boldsymbol{x}|c_i)$，因此我们并不需要计算$P(\boldsymbol{x})$。

3.朴素贝叶斯分类器

假设样本$\boldsymbol{x}$包含d个属性，即$\boldsymbol{x} = \{x_1, x_2, \ldots, x_d\}$。于是有：

$$P(\boldsymbol{x}|c_i) = P(x_1, x_2, \cdots, x_d|c_i)$$

这个联合概率难以从有限的训练样本中直接估计到，于是朴素贝叶斯（Naive Bayesian）采用了“**属性条件独立性假设**”：对已知类别，假设所有属性相互独立，则有：

$$P(x_1, x_2, \cdots, x_d|c_i) = \prod_{j=1}^{d} P(x_j|c_i)$$

则可以推出相应的判定准则：

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c_i \in Y} P(c_i) \prod_{j=1}^{d} P(x_j|c_i)$$

条件概率$P(x_j|c_i)$的求解：

如果x_j是标签属性，那么我们可以通过**计数**的方法估计$P(x_j|c_i)$

$$P(x_j|c_i) = \frac{P(x_j, c_i)}{P(c_i)} \approx \frac{\#(x_j, c_i)}{\#(c_i)}$$

其中，$\#(x_j, c_i)$表示在训练样本中x_j与c_i共同出现的次数。

如果x_j是数值属性，通常我们假设类别中c_i的左右样本第j个属性服从正态分布。首先故居这个分布的均值μ和方差σ，然后计算x_j在这个分布中的概率密度$P(x_j|c_i)$

4. 半朴素贝叶斯分类器

朴素贝叶斯采用了“属性条件独立性假设”，半朴素贝叶斯基本想法时适当考虑一部分属性间的相互依赖信息。“**独依赖估计**”（One—Dependence Estimator，ODE）是半朴素贝叶斯最常用的一种策略，即：假设每个属性在类别之外最多依赖一个其他属性。

$$P(\boldsymbol{x}|c_i) = \prod_{j=1}^{d} P(x_j|c_i, \mathrm{pa}_j)$$

其中pa_j为属性x_i所依赖的属性，成为x_i的父属性。假设父属性pa_j已知，那么可以使用下面的公式估计$P(x_j|c_i,\mathrm{pa}_j)$

$$P(x_j|c_i,\mathrm{pa}_j)=\frac{P(x_j,c_i,\mathrm{pa}_j)}{P(c_i,\mathrm{pa}_j)}$$

5.举例理解朴素贝叶斯分类器

使用经典的西瓜训练集如下：

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

对下面的测试例“测1”进行 分类：

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

首先，估计类先验概率$P(c_j)$，有

$$P(\text{好瓜} = \text{是}) = \frac{8}{17} = 0.471$$
$$P(\text{好瓜} = \text{否}) = \frac{9}{17} = 0.529$$

然后，为每个属性估计条件概率（这里，对于连续属性，假定它们服从正态分布）

$$P_{\text{青绿}|\text{是}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375$$

$$P_{\text{青绿}|\text{否}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333$$

$$P_{\text{蜷缩}|\text{是}} = P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{是}) = \frac{5}{8} = 0.625$$

$$P_{\text{蜷缩}|\text{否}} = P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{否}) = \frac{3}{9} = 0.333$$

$$P_{\text{浊响}|\text{是}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{浊响}|\text{否}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{否}) = \frac{4}{9} \approx 0.444$$

$$P_{\text{清晰}|\text{是}} = P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{是}) = \frac{7}{8} = 0.875$$

$$P_{\text{清晰}|\text{否}} = P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222$$

$$P_{\text{凹陷}|\text{是}} = P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{凹陷}|\text{否}} = P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222$$

$$P_{\text{硬滑}|\text{是}} = P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750$$

$$P_{\text{硬滑}|\text{否}} = P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{否}) = \frac{6}{9} \approx 0.667$$

$$
\begin{aligned}
\rho_{\text{密度：}0.697|\text{是}} &= \rho(\text{密度} = 0.697|\text{好瓜} = \text{是}) \\
&= \frac{1}{\sqrt{2\pi} \times 0.129} exp\left(-\frac{(0.697-0.574)^2}{2 \times 0.129^2}\right) \approx 1.959
\end{aligned}
$$

$$
\begin{aligned}
\rho_{\text{密度：}0.697|\text{否}} &= \rho(\text{密度} = 0.697|\text{好瓜} = \text{否}) \\
&= \frac{1}{\sqrt{2\pi} \times 0.195} exp\left(-\frac{(0.697-0.496)^2}{2 \times 0.195^2}\right) \approx 1.203
\end{aligned}
$$

$$
\begin{aligned}
\rho_{\text{含糖：}0.460|\text{是}} &= \rho(\text{密度} = 0.460|\text{好瓜} = \text{是}) \\
&= \frac{1}{\sqrt{2\pi} \times 0.101} exp\left(-\frac{(0.460-0.279)^2}{2 \times 0.101^2}\right) \approx 0.788
\end{aligned}
$$

$$
\begin{aligned}
\rho_{\text{含糖：}0.460|\text{否}} &= \rho(\text{密度} = 0.460|\text{好瓜} = \text{是}) \\
&= \frac{1}{\sqrt{2\pi} \times 0.108} exp\left(-\frac{(0.460-0.154)^2}{2 \times 0.108^2}\right) \approx 0.066
\end{aligned}
$$

于是有

$$
\begin{aligned}
&P(\text{好瓜} = \text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}} \\
&\quad \times P_{\text{硬滑}|\text{是}} \times p_{\text{密度：}0.697|\text{是}} \times p_{\text{含糖：}0.460|\text{是}} \approx 0.063
\end{aligned}
$$

$$
\begin{aligned}
&P(\text{好瓜} = \text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}} \\
&\quad \times P_{\text{硬滑}|\text{否}} \times p_{\text{密度：}0.697|\text{否}} \times p_{\text{含糖：}0.460|\text{否}} \approx 6.80 \times 10^{-5}
\end{aligned}
$$

由于$0.063 > 6.80 \times 10^{-5}$，因此，朴素贝叶斯分类器将测试样本“测1”判别为“好瓜”。