四川大學
SICHUAN UNIVERSITY

机器学习-第四章 朴素贝叶斯

教师：胡俊杰 副教授

邮箱：hujunjie@scu.edu.cn

Sigmoid函数

$$h(x) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_n x^{(n)}$$

可以设$x^{(0)} = 1$，则

标量形式： $h(x) = w_0 x^{(0)} + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_n x^{(n)} = \sum_{k=0}^{n} w_k x^{(k)}$

向量形式： $h(x) = w^T x$

$$z = h(x) = w^T x \qquad z \in (-\infty, +\infty)$$

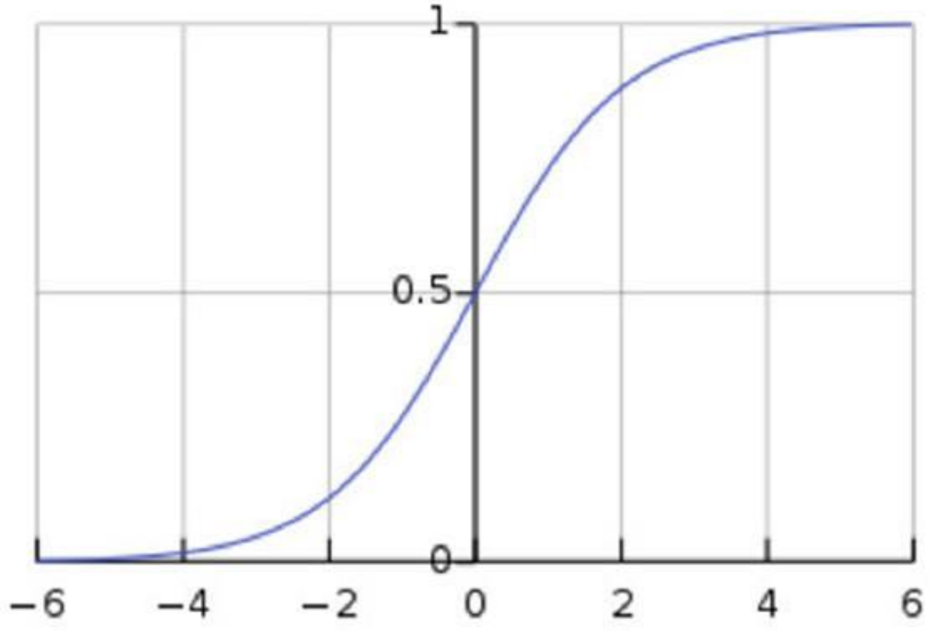
- 我们希望输出的值代表事件发生的概率，即$z \in [0,1]$
- $\sigma(z)$，对输出z作用一个函数，将z压缩至$[0,1]$区间

Sigmoid函数

Sigmoid 函数

$\sigma(z)$代表一个常用连续S形函数（Sigmoid function）或逻辑函数（ Logistic function ）

$$\sigma(z) = g(z) = \frac{1}{1+e^{-z}} \qquad z=w^T x$$



当$\sigma(z)$大于等于0.5时，预测 为1

当$\sigma(z)$小于0.5时，预测 为0

- 模型预测的类别不仅取决于模型的输出值，也依赖于设置的阈值

逻辑回归

假设一个二分类模型：

$$p(y=1|x;w)=h(x)$$

$$p(y=0|x;w)=1-h(x)$$

则：

$$p(y|x;w)=(h(x))^y(1-h(x))^{1-y}$$

逻辑回归模型的假设是： $h(x)=g(w^Tx)=g(z)$，其中$z=w^Tx$

$$g(z)=\frac{1}{1+e^{-z}} \qquad g'(z)=g(z)(1-g(z))$$

- 虽然称为逻辑回归，但用于解决分类问题

逻辑回归求解

$\hat{y}$ 表示模型的预测值$h(x)$

y 表示真实值(标签)

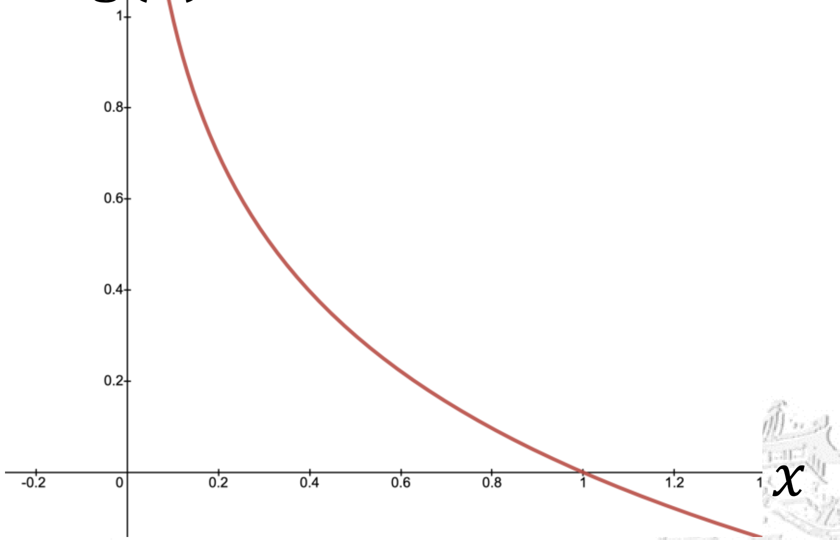
代价函数

$$J(w)=\frac{1}{m}\sum_{i=1}^{m}L(h(x_i),y_i)=\frac{1}{m}\sum_{i=1}^{m}L(\hat{y}_i,y_i)$$

$$h(x_i)=\frac{1}{1+e^{-w^Tx_i}},h(x_i)\in(0,1)$$

$$L(h(x_i),y_i)=\begin{cases}-\log(h(x_i),if\ y=1\\-\log\left(1-h(x_i)\right),if\ y=0\end{cases}$$

$-\log(x)$



- y只能等于0或1
- 当$x\in(0,1)$区间时， $-\log(x)>0$且单调递减
- 通过最小化$L(h(x_i),y_i)$，使得$h(x_i)\to y_i$

逻辑回归求解

代价函数

$$L(h(x_i), y_i) = \begin{cases} -\log(h(x_i), if\ y_i = 1 \\ -\log\left(1 - h(x_i)\right), if\ y_i = 0 \end{cases}$$

$\hat{y}$ 表示模型的预测值$h(x)$

y 表示真实值(标签)

$$L(h(x_i), y_i) = -y_i \log(h(x_i) - (1 - y_\mathrm{i}) \log\left(1 - h(x_i)\right)$$

$$J(w) = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i) = \frac{1}{m} \sum_{i=1}^{m} (-y_\mathrm{i} \log(h(x_i)) - (1 - y_i) \log(1 - h(x_i)))$$

似然函数（Likelihood function）

$L(\theta|x)$: 给定x时，关于参数θ的似然函数（Likelihood function）。代表给定数据x，参数θ生成该数据的可能性

极大似然估计（Maximum Likelihood Estimation, MLE）

- $\hat{\theta} = \operatorname{argmax} L(\theta|x_1, x_2, \ldots, x_n)$
- 在θ的所有可能取值中，寻找到$\hat{\theta}$使得似然函数最大，$\hat{\theta}$即称为θ的极大似然估计

逻辑回归求解

求解过程：

似然函数为： $L(w)=\prod_{i=1}^{m} P\left(y_i | x_i; w\right)=\prod_{i=1}^{m}\left(h\left(x_i\right)\right)^{y_i}\left(1-h\left(x_i\right)\right)^{1-y_i}$

似然函数两边取对数，则累乘号变成了累加号：

$$\boxed{\log L(w)=\sum_{i=1}^{m}\left(y_i \log\left(h\left(x_i\right)\right)+\left(1-y_i\right) \log\left(1-h\left(x_i\right)\right)\right)}$$

最大化

代价函数为：

$$J(w)=\boxed{-\frac{1}{m} \log L(w)=-\frac{1}{m} \sum_{i=1}^{m}\left(y_i \log\left(h\left(x_i\right)\right)+\left(1-y_i\right) \log\left(1-h\left(x_i\right)\right)\right)}$$

最小化

逻辑回归求解

梯度下降求解过程：

$$w_j := w_j - \alpha \frac{\partial J(w)}{\partial w_j}$$

$$J(w) = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i \log(h(x_i)) + (1-y_i)\log(1-h(x_i))\right)$$

$$\boxed{\frac{\partial J(w)}{\partial w_j} = \frac{1}{m}\sum_{i=1}^{m}\left(h(x_i) - y_i\right)x_i^{(j)}}$$

$$h(x_i) = \frac{1}{1+e^{-w^T x_i}}$$

$$w_j := w_j - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h(x_i) - y_i\right)x_i^{(j)}$$

本章目录

1.贝叶斯方法

1.贝叶斯

- 贝叶斯 (Thomas Bayes)，英国数学家，曾做过神父，英国皇家学会会员。贝叶斯主要研究概率论。他提出了一种概率推理方法，后人称之为贝叶斯定理（Bayes’ theorem）。他的研究对统计推断、概率推理和决策分析等领域产生了重要影响。他去世后，理查德 普莱斯（Richard Price）于 1763 年整理并提交其论文《机会问题的解法》（An essay towards solving a problem in the doctrine of chances）给英国皇家学会，这对现代概率论和数理统计的发展产生了深远的影响

1.贝叶斯定理

联合概率： 联合概率是指多个随机变量同时满足各自条件的概率。X与Y的联合概率表示为$P(X,Y)$、 $P(XY)$ 或$P(X\cap Y)$

假设X和Y都服从正态分布，则$P(X<5,Y<0)$就是一个联合概率，表示 $X<5$ 和 $Y<0$同时发生的概率

1.贝叶斯定理

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

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

$$\longrightarrow P(Y|X)P(X) = P(X|Y)P(Y) \longrightarrow P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X|Y)P(Y) + P(X|\bar{Y})P(\bar{Y})}$$

全概率公式

1.贝叶斯方法-背景知识

贝叶斯分类：贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础，故统称为贝叶斯分类

先验概率
Prior probability: 根据以往经验和分析得到的概率，记为$P(Y)$。是在观测数据前，表达事件不确定性的概率分布，其代表经验知识，与观测数据无关

后验概率
Posterior probability: 给定观测数据X后，对事件Y发生概率的更新，记为$P(Y|X)$，它反映了在获取新数据X后，调整对Y发生可能性的评估

1.贝叶斯方法-一个简单示例

- 假设某种疾病在所有人群中的感染率是0.1%
- 医院现有技术对于该疾病检测准确率为 99%（已知患病情况下， 99% 的可能性可以检查出阳性；正常人 99% 的可能性检查为正常。）

问：从人群中随机抽一个人去检测，医院给出的检测结果为阳性，那么这个人实际得病的概率是多少?

99%？

1.贝叶斯方法-一个简单示例

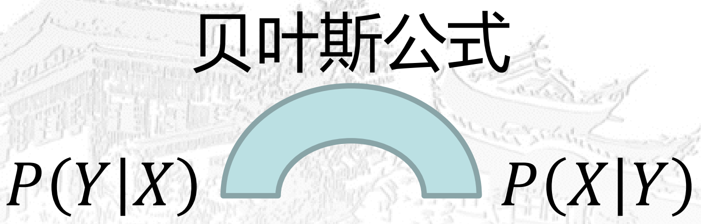
Y: 某人患有该疾病

X: 医院检测结果为阳性（检测结果显示患病）

- 医院现有的技术对于该疾病检测准确率为 99%：$P(X|Y)=99\%$

> 问：从人群中随机抽一个人去检测，医院给出的检测结果为阳性，那么这个人实际得病的概率是多少?

即求$P(Y|X)$



1.贝叶斯方法-一个简单示例

X:医院检测结果为阳性（检测结果显示患病）　　　　Y:某人患有该疾病

贝叶斯公式： $P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}=\frac{P(X|Y)P(Y)}{P(X|Y)P(Y)+P(X|\bar{Y})P(\bar{Y})}$

- 假设某种疾病在人群中的感染率是0.1%: $P(Y)=0.1\%,\ P(\bar{Y})=99.9\%$
- 医院现有的技术对于该疾病检测准确率为 99%: $P(X|Y)=99\%,\ P(X|\bar{Y})=\frac{\boxed{P(X,\bar{Y})}}{P(\bar{Y})}=\frac{0.01}{0.999}\approx 1\%$

错检/误诊

$$P(Y|X)=\frac{0.99*0.001}{0.99*0.001+0.01*0.999}\approx 0.09$$

从人群中随机抽一个人去检测，医院给出的检测结果为阳性，实际真实得病的概率为9%

联系生活中的贝叶斯

1. 贝叶斯方法

贝叶斯公式

后验概率　　似然度　　先验概率

$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

边际似然度/证据

朴素贝叶斯法是典型的生成学习方法。生成方法由训练数据学习联合概率分布 $P(X,Y)$，然后求得后验概率分布$P(Y|X)$

具体来说，利用训练数据学习$P(X|Y)$和$P(Y)$的估计，得到联合概率分布：

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

2.朴素贝叶斯原理

2.朴素贝叶斯原理

判别模型和生成模型

监督学习模型可分为
判别模型（Discriminative model）和**生成模型**（Generative model）

判别模型（Discriminative model)	生成模型（Generative model）
由数据直接学习决策函数$Y = f(X)$或者条件概率分布$P(Y\|X)$ 的模型，即判别模型。基本思想是在有限样本条件下建立判别函数，不考虑样本的产生模型，直接研究预测模型。 即：直接估计$P(Y\|X)$	由训练数据学习联合概率分布 $P(X,Y)$，然后求得后验概率分布$P(Y\|X)$。具体来说，利用训练数据学习$P(X\|Y)$和$P(Y)$的估计，得到联合概率分布：$P(X,Y) = P(X\|Y)P(Y)$，再利用它进行分类。 即：先估计$P(X\|Y)$，然后推导$P(Y\|X)$
线性回归、逻辑回归、感知机、决策树、支持向量机……	朴素贝叶斯、HMM……

2.朴素贝叶斯原理

- 假设输入空间$\chi \in R^n$，即每个样本x是一个n维向量 $x \in \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(n)} \end{bmatrix}$

  注：每个维度可能有多种取值，记为S_j，即第j维可能有S_j种取值

- 假设输出空间$\mathcal{Y} = \{c_1, c_2, \dots, c_K\}$，即在分类任务中共有$K$个类别， c : class

- X是定义在输入空间χ上的随机变量，Y是定义在输出空间$\mathcal{Y}$上的随机变量

- $P(X,Y)$是X和Y的联合概率分布，训练数据集$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 由$P(X,Y)$独立同分布产生

2.朴素贝叶斯原理

$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

为随机变量X和Y赋值

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

$P(Y = c_k|X = x)$: 样本x属于第k个类别的概率

2.朴素贝叶斯原理

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

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

- 假设$x^{(j)}$可能的取值有S_j个，$j=1,2,\dots n$，Y可能值有K个，则$P(X=x|Y=c_k)$的可能情况有 $K\prod_{j=1}^{n}S_j$种，复杂度高
- 若假设在类别确定的条件下，各特征相互独立（条件独立），则

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}\middle|Y=c_k\right)$$

简化问题

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

2.朴素贝叶斯原理

- 若**假设**在类别确定的条件下，各特征相互独立，则

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\dots,X^{(n)}=x^{(n)}\middle|Y=c_k\right)$$

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

- 以上正是**朴素**贝叶斯方法的由来
- 以上假设使得朴素贝叶斯方法计算高效，且易于实现，但有时会牺牲一定的分类准确率（No free lunch, 没有免费的午餐）

计算精度 计算效率

2.朴素贝叶斯原理

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

贝叶斯公式： $$P(Y=c_k|X=x) = \frac{P(X=x|Y=c_k)P(Y=c_k)}{P(X=x)} = \frac{\boxed{P(X=x|Y=c_k)}P(Y=c_k)}{\sum_i \boxed{P(X=x|Y=c_i)}P(Y=c_i)}$$

$$P(Y=c_k|X=x) = \frac{P(Y=c_k)\prod_{j=1}^{n} P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_i P(Y=c_i)\prod_{j=1}^{n} P(X^{(j)}=x^{(j)}|Y=c_i)}$$

2.朴素贝叶斯原理

朴素贝叶斯分类的基本公式：

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_i P(Y=c_i)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_i)}$$

- 对于输入x，其可能的类别数为$k=1,2,\ldots,K$，选择$P(Y=c_k|X=x)$最大的那项即可

$$P(Y=c_1|X=x)=0.1$$

$$P(Y=c_2|X=x)=0.7$$

输入x对应c_2类

$$P(Y=c_3|X=x)=0.2$$

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

2.朴素贝叶斯原理

$$y = \underset{c_k}{\operatorname{argmax}} \frac{P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)}{\sum_i P(Y = c_i) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_i)}$$

- 对任意的c_k而言，以上公式的分母均相等，因此以上公式可简化为：

$$y = \underset{c_k}{\operatorname{argmax}} P(Y = c_k) \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)$$

2.朴素贝叶斯原理

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

c_k类样本的数目

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

N: 训练样本数目

$$I(y_i = c_k) = \begin{cases} 1, if\ y_i = c_k \\ 0, else \end{cases}$$

Indicator function (指示函数)

属于c_k类，且输入特征为a_{jl}样本的数目

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

c_k类样本的数目

$j = 1,2, \dots, n$ 每个样本共有n维特征

$l = 1,2, \dots, S_j$ 第j维特征可能有S_j种取值

$k = 1,2, \dots, K$ 共有K个类别

$x_i^{(j)}$:第i个样本的第j个特征 a_{jl}: 第j个特征可能取的第l个值

2.朴素贝叶斯算法

输入：训练数据$T=\{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N))\}$，其中$x_i=\left(x_i^{(1)},x_i^{(2)},\dots,x_i^{(n)}\right)^T, y_i\in\{c_1,c_2,\dots,c_K\}$。

$x_i^{(j)}$代表第i个样本的第j个特征， $x_i^{(j)}\in\{a_{j1},a_{j2},\dots,a_{jS_j}\}$，$j=1,2,\dots,n,l=1,2,\dots,S_j$

输出：样本x的所属类别

步骤1：计算先验概率$P(Y=c_k)$和条件概率$P\left(X^{(j)}=a_{jl}\middle|Y=c_k\right)$

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

步骤2：对于给定的样本$x=\left(x^{(1)},x^{(2)},\dots,x^{(n)}\right)^T$，计算 $P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$

步骤3：确定样本x的类别y

$$y=\underset{c_k}{\operatorname{argmax}}P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)$$

3.朴素贝叶斯案例

3.朴素贝叶斯案例1

x y

共13个训练样本

样本	天气	温度	湿度	风速	是否打网球
1	阴	热	高	弱	是
2	雨	中	高	弱	是
3	雨	冷	正常	弱	是
4	阴	冷	正常	强	是
5	晴	冷	正常	弱	是
6	雨	中	正常	弱	是
7	晴	中	正常	强	是
8	阴	中	高	强	是
9	晴	热	高	弱	否
10	晴	热	高	强	否
11	雨	冷	高	强	否
12	晴	中	高	弱	否
13	雨	中	高	强	否

问：天气晴，温度冷，湿度高，风速强，是否适合打网球

3.朴素贝叶斯案例1

问：天气晴，温度冷，湿度高，风速强，是否适合打网球

$$y = \underset{c\in\{是,否\}}{\operatorname{argmax}} P(Y=c)P(天气=晴|c)P(温度=冷|c)P(湿度=高|c)P(风速=强|c)$$

$P(是)=8/13$ $\quad$ $P(否)=5/13$

$P(天气=晴|是)=2/8$ $\quad$ $P(天气=晴|否)=3/5$

$P(温度=冷|是)=3/8$ $\quad$ $P(温度=冷|否)=1/5$

$P(湿度=高|是)=3/8$ $\quad$ $P(湿度=高|否)=5/5$

$P(风速=强|是)=3/8$ $\quad$ $P(风速=强|否)=3/5$

$$P(是)P(天气=晴|是)P(温度=冷|是)P(湿度=高|是)P(风速=强|是)=0.0081$$

$$\boxed{P(否)P(天气=晴|否)P(温度=冷|否)P(湿度=高|否)P(风速=强|否)=0.027}$$

样本	天气	温度	湿度	风速	是否打网球
1	阴	热	高	弱	是
2	雨	中	高	弱	是
3	雨	冷	正常	弱	是
4	阴	冷	正常	强	是
5	晴	冷	正常	弱	是
6	雨	中	正常	弱	是
7	晴	中	正常	强	是
8	阴	中	高	强	是
9	晴	热	高	弱	否
10	晴	热	高	强	否
11	雨	冷	高	强	否
12	晴	中	高	弱	否
13	雨	中	高	强	否

3.朴素贝叶斯案例2

问：天气晴，温度冷，湿度正常，风速强，是否适合打网球

$$y = \underset{c\in\{是,否\}}{\operatorname{argmax}} P(Y=c)P(天气=晴|c)P(温度=冷|c)P(湿度=正常|晴)P(风速=强|c)$$

$P(是)=8/13$ $\quad$ $P(否)=5/13$

$P(天气=晴|是)=2/8$ $\quad$ $P(天气=晴|否)=3/5$

$P(温度=冷|是)=3/8$ $\quad$ $P(温度=冷|否)=1/5$

$P(湿度=正常|是)=5/8$ $\quad$ $P(湿度=正常|否)=0/5$

$P(风速=强|是)=3/8$ $\quad$ $P(风速=强|否)=3/5$

不论天气、温度、风速如何变化，累乘后的概率均为0

样本	天气	温度	湿度	风速	是否打网球
1	阴	热	高	弱	是
2	雨	中	高	弱	是
3	雨	冷	正常	弱	是
4	阴	冷	正常	强	是
5	晴	冷	正常	弱	是
6	雨	中	正常	弱	是
7	晴	中	正常	强	是
8	阴	中	高	强	是
9	晴	热	高	弱	否
10	晴	热	高	强	否
11	雨	冷	高	强	否
12	晴	中	高	弱	否
13	雨	中	高	强	否

$$P(是)P(天气=晴|是)P(温度=冷|是)P(湿度=正常|是)P(风速=强|是)=0.021$$

$$\boxed{P(否)P(天气=晴|否)P(温度=冷|否)P(湿度=正常|否)P(风速=强|否)=0}$$

3.朴素贝叶斯案例2

拉普拉斯平滑是一种用于平滑分类数据的技术。引入拉普拉斯平滑法来解决零概率问题,通过应用此方法,先验概率和条件概率可以写为

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

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

其中K表示分类类别数量,S_j表示第j维可能的取值数量。

- 当$\lambda=0$时,即为一般的朴素贝叶斯方法
- 当$\lambda=1$时,即为加入了拉普拉斯平滑的朴素贝叶斯方法
- 加入拉普拉斯平滑之后,避免了出现概率为0的情况,又保证了每个值都在0到1的范围内

3.朴素贝叶斯案例2

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

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

a_{jl}: 第j个属性可能取的第l个值

对于所有类别， $$\sum_{k=1}^{K} P_\lambda(Y=c_k)=1$$

对于第j个属性， $$\sum_{l=1}^{S_j} P_\lambda(X^{(j)}=a_{jl}|Y=c_k)=1$$

加入拉普拉斯平滑后，朴素贝叶斯方法仍服从概率分布的性质

3.朴素贝叶斯案例2

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

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

问：天气晴，温度冷，湿度正常，风速强，是否适合打网球

$$y=\underset{c\in\{是,否\}}{\operatorname{argmax}} P(Y=c)P(天气=晴|c)P(温度=冷|c)P(湿度=正常|晴)P(风速=强|c)$$

$P(是)=(8+1)/(13+2)$=9/15　　$P(否)=(5+1)/(13+2)$=6/15

$P(天气=晴|是)=(2+1)/(8+3)$=3/11　　$P(天气=晴|否)=(3+1)/(5+3)$=4/8

$P(温度=冷|是)=(3+1)/(8+3)$=4/11　　$P(温度=冷|否)=(1+1)/(5+3)$=2/8

$P(湿度=正常|是)=(5+1)/(8+2)$=6/10　　$P(湿度=正常|否)=(0+1)/(5+2)$=1/7

$P(风速=强|是)=(3+1)/(8+2)$=4/10　　$P(风速=强|否)=(3+1)/(5+2)$=4/7

$P(是)P(天气=晴|是)P(温度=冷|是)P(湿度=正常|是)P(风速=强|是)=0.014$

$P(否)P(天气=晴|否)P(温度=冷|否)P(湿度=正常|否)P(风速=强|否)=0.0041$

样本	天气	温度	湿度	风速	是否打网球
1	阴	热	高	弱	是
2	雨	中	高	弱	是
3	雨	冷	正常	弱	是
4	阴	冷	正常	强	是
5	晴	冷	正常	弱	是
6	雨	中	正常	弱	是
7	晴	中	正常	强	是
8	阴	中	高	强	是
9	晴	热	高	弱	否
10	晴	热	高	强	否
11	雨	冷	高	强	否
12	晴	中	高	弱	否
13	雨	中	高	强	否

作业

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

问:
色泽=青绿
根蒂=蜷缩
敲声=沉闷
纹理=模糊
脐部=平坦
触感=硬滑

是否为好瓜？请给出计算过程

谢 谢！