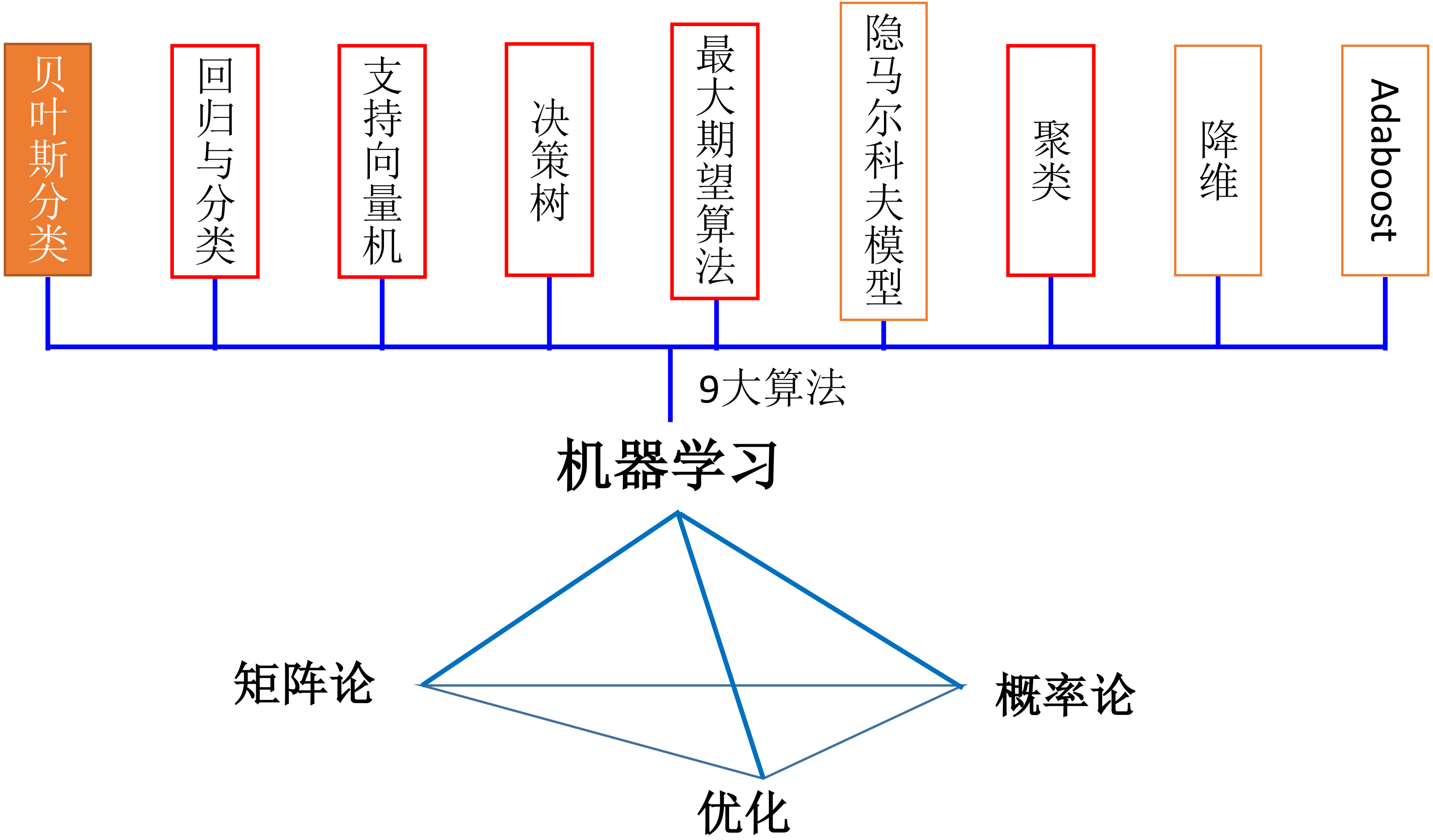
深蓝学院
shenlanxueyuan.com

贝叶斯分类

——贝叶斯决策与朴素贝叶斯

主讲人 霍博士


贝叶斯分类
回归与分类
支持向量机
决策树
最大期望算法
隐马尔科夫模型
聚类
降维
Adaboost
9大算法
机器学习
矩阵论
概率论
优化

垃圾邮件分类

广告分类

文档分类

食品分类

客户分类

……

分类问题

解决方案

?

识别=分类？

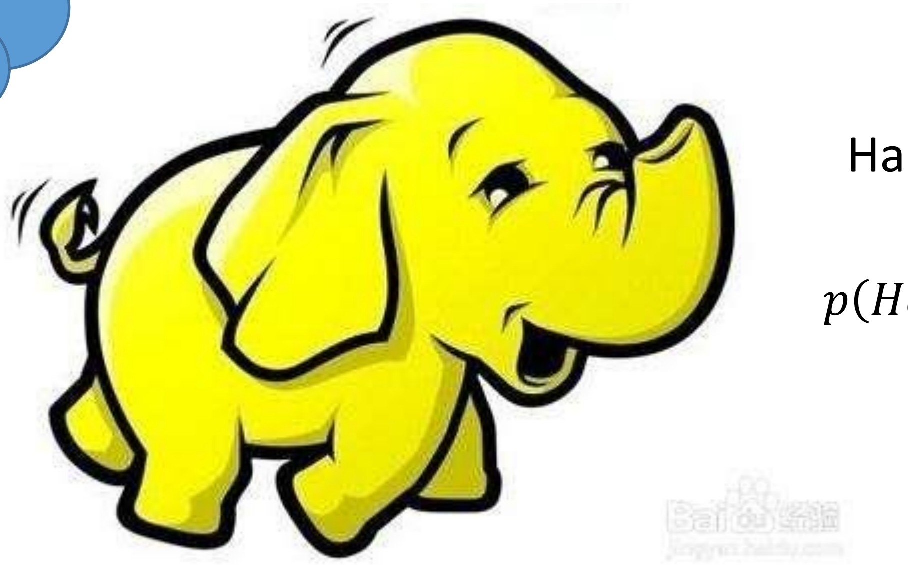
绳子？
柱子？
蒲扇？
……

What's this?

$p(绳子|细长)$

$p(柱子|粗壮)$

$p(蒲扇|宽阔)$



Hadoop？

$p(Hadoop|logo)$

$$p(大象|尾巴=细长，腿=粗壮，耳朵=宽阔，\ldots\ldots)$$

课程大纲

深蓝学院
shenlanxueyuan.com

- ✓ 贝叶斯决策理论基础
- ✓ 朴素贝叶斯分类器
- ✓ 鸢尾花分类实践

贝叶斯决策理论基础

深蓝学院 shenlanxueyuan.com

条件概率

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

全概率公式 $\bigcup_{i=1}^{n} B_i = \Omega$

- $P(A)=P(A|B_1)P(B_1)+\cdots+P(A|B_n)P(B_n)$

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

贝叶斯公式

- $P(B_i|A)=\frac{P(A|B_i)P(B_i)}{P(A)}=\frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$

贝叶斯决策理论基础

样本$\boldsymbol{x}$

类别集合$Y = \{c_1, c_2, \cdots, c_K\}$

条件风险(Conditional Risk)

$$Risk(c_i|\boldsymbol{x}) = \sum_{j=1}^{K} \lambda_{ij} P(c_j|\boldsymbol{x})$$

λ_{ij}：将一个真实类别为c_j的样本误分为c_i产生的期望损失

贝叶斯风险

$$R(f) = E_{\{x\}}[Risk(f(\boldsymbol{x})|\boldsymbol{x})]$$

贝叶斯判定准则(Bayesian Decision Rule)

为最小化总体风险，只需在每个样本上选择能使条件风险最小的类别标记

贝叶斯最优分类器

$$f^*(\boldsymbol{x}) = \underset{c \in Y}{\operatorname{argmin}}\, R(c|\boldsymbol{x})$$

贝叶斯决策理论基础

条件风险(Conditional Risk)

$$Risk(c_i|\boldsymbol{x}) = \sum_{j=1}^{K} \lambda_{ij} P(c_j|\boldsymbol{x})$$

若目标函数是**最小化分类错误率**，则

$$\lambda_{ij} = \begin{cases} 0, & \text{if } i = j \\ 1, & otherwise \end{cases}$$

$$Risk(c|\boldsymbol{x}) = 1 - P(c|\boldsymbol{x})$$

贝叶斯最优分类器

$$f^*(\boldsymbol{x}) = \operatorname*{argmin}_{c \in Y} R(c|\boldsymbol{x})$$

$$f^*(\boldsymbol{x}) = \operatorname*{argmax}_{c \in Y} \{P(c|\boldsymbol{x})\}$$

后验概率

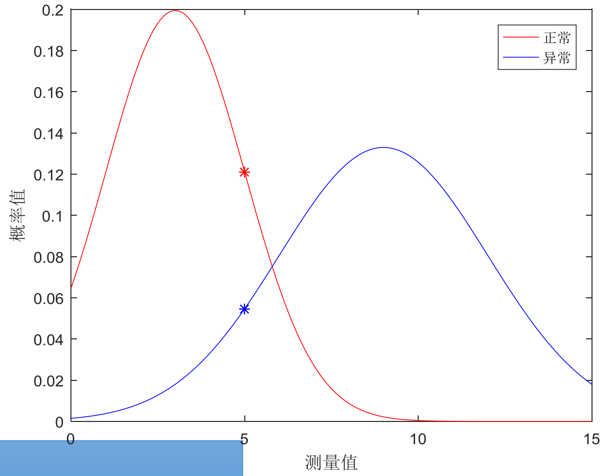
最小错误率的贝叶斯决策=选择具有最高概率的决策

贝叶斯决策理论基础

例 某体能指标正常和异常两类的先验概率分别为：

$$P(\text{正常}) = 0.95，P(\text{异常}) = 0.05.$$

小张的该项指标测量值$x = 5$，由每类的条件概率密度分布曲线可得，$P(x = 5|\text{正常}) = 0.12$，$P(x = 5|\text{异常}) = 0.05$. 小张的该项体能指标正常吗？



解：判断$P(\text{正常}|x = 5)$与$P(\text{异常}|x = 5)$的大小。

利用贝叶斯公式分别计算正常和异常的后验概率：

先验起主导作用

因此，$P(\text{正常}|x = 5) > P(\text{异常}|x = 5)$，根据贝叶斯决策规则，合理的决策是把小张该项体能指标归类于正常。

贝叶斯决策理论基础

条件风险(Conditional Risk)

$$Risk(c_i|\boldsymbol{x}) = \sum_{j=1}^{K} \lambda_{ij} P(c_j|\boldsymbol{x})$$

贝叶斯风险

$$R(f) = E_{\{x\}}[Risk(f(\boldsymbol{x})|\boldsymbol{x})]$$

贝叶斯最优分类器

$$f^*(\boldsymbol{x}) = \operatorname*{argmin}_{c \in Y} Risk(c|\boldsymbol{x})$$

最小风险的贝叶斯决策

贝叶斯决策理论基础

例 某体能指标正常和异常两类的先验概率分别为：

$$P(\text{正常}) = 0.95，P(\text{异常}) = 0.05.$$

小张的该项指标测量值$x = 5$，由每类的条件概率密度分布曲线可得，$P(x = 5|\text{正常}) = 0.12$， $P(x = 5|\text{异常}) = 0.05$. 小张的该项体能指标正常吗？

决策 \ 损失 \ 状态	正常	异常
正常	0	10
异常	1	0

解：后验概率：$P(\text{正常}|x = 5) \approx 0.98$， $P(\text{异常}|x = 5) \approx 0.02$

条件风险：$R(\text{正常}|x = 5) = \lambda_{11}P(\text{正常}|x = 5) + \lambda_{12}P(\text{异常}|x = 5) \approx 0.2$，

$R(\text{异常}|x = 5) = \lambda_{21}P(\text{正常}|x = 5) + \lambda_{22}P(\text{异常}|x = 5) \approx 0.98 > R(\text{正常}|x = 5)$

因此， 项体能
指标归

“损失”起主导作用

贝叶斯决策理论基础

生成式模型

先验概率

条件概率/似然

$$P(c|\boldsymbol{x})=\frac{P(\boldsymbol{x},c)}{P(\boldsymbol{x})} \xrightarrow{\text{贝叶斯定理}} =\frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})} \propto P(c)P(\boldsymbol{x}|c)$$

归一化因子

样本空间中各类样本所占的比例：$P(c)\xrightarrow{\text{大数定律}}\frac{N_c}{N}$

$P(\boldsymbol{x}|c)$?

贝叶斯决策理论基础

"未被观测到" $\neq$ "出现概率为0"

课程大纲

深蓝学院
shenlanxueyuan.com

- ✓ 贝叶斯决策理论基础
- ✓ 朴素贝叶斯分类器
- ✓ 鸢尾花分类实践

朴素贝叶斯分类器

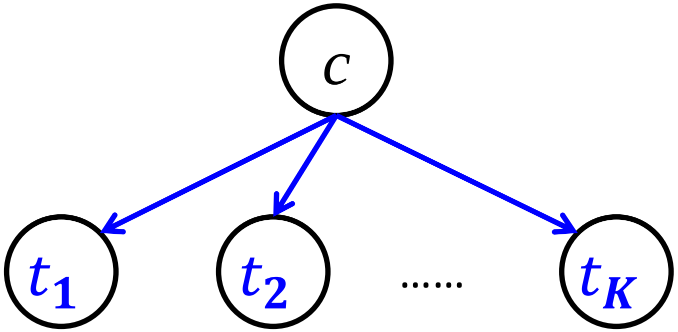
MAP？

$$P(c|\boldsymbol{x}) \propto P(c)P(\boldsymbol{x}|c)$$

$$\boldsymbol{x} = (t_1, t_2, \cdots, t_K)$$

属性条件独立性假设

$$P(c|\boldsymbol{x}) \propto P(c)\prod_{i=1}^{K} P(t_i|c)$$



朴素贝叶斯分类器
(Naïve Bayesian Classifier)

$$f_{nbc}(\boldsymbol{x}) = \underset{c \in Y}{\operatorname{argmax}} \left\{ P(c)\prod_{i=1}^{K} P(t_i|c) \right\}$$

朴素贝叶斯分类器

朴素贝叶斯分类器
(Naïve Bayesian Classifier)

$$f_{nbc}(\boldsymbol{x}) = \underset{c \in Y}{\operatorname{argmax}} \left\{ P(c) \prod_{i=1}^{K} P(t_i|c) \right\}$$

样本集合$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$，类别集合$Y = \{c_1, c_2, \cdots, c_K\}$

$$P(c) = \frac{N_c}{N}$$

$$\boxed{P(t_i|c) = \frac{N_{(c,t_i)}}{N_c}}$$

$N_{(c,x_i)}$表示第c类中在属性取值为t_i的样本个数

朴素贝叶斯分类器

深蓝学院 shenlanxueyuan.com

西瓜数据集 SL1.0

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否
测0	青绿	蜷缩	浊响	清晰	凹陷	硬滑	

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

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

- For each training sample
  - If 样本标签==c_i
    - c_i类别样本个数增加1
    - For each feature t_i
      - 如果特征t_i出现在样本中，该特征t_i在c_i类别出现的次数加1
- For each class
  - For each feature
    - 将该特征在该类别出现的次数除以该类别样本个数
- 返回每个类别的条件概率$P(t_i|c_i)$

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

周志华，机器学习，P84,151例子改编

朴素贝叶斯分类器

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

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

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

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

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

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

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

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

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

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

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

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

周志华，机器学习，P84,152

$P(好瓜 = 是) = \frac{8}{17} \approx 0.471$,

$P(好瓜 = 否) = \frac{9}{17} \approx 0.529$.

测0	青绿	蜷缩	浊响	清晰	凹陷	硬滑

$$P(好瓜 = 是) \times P_{青绿|是} \times P_{蜷缩|是} \times P_{浊响|是} \times P_{清晰|是} \times P_{凹陷|是} \times P_{硬滑|是} \approx 0.041$$

$$P(好瓜 = 否) \times P_{青绿|否} \times P_{蜷缩|否} \times P_{浊响|否} \times P_{清晰|否} \times P_{凹陷|否} \times P_{硬滑|否} \approx 8.562e-4$$

$0.041>8.562e-4 \xrightarrow{朴素贝叶斯}$ 好瓜

朴素贝叶斯分类器

深蓝学院 shenlanxueyuan.com

表 4.3 西瓜数据集 3.0

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
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
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

$$p_{\text{密度: }0.697|\text{是}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{是})$$
$$= \frac{1}{\sqrt{2\pi}\cdot 0.129}\exp\left(-\frac{(0.697-0.574)^2}{2\cdot 0.129^2}\right) \approx 1.959\,,$$

$$p_{\text{密度: }0.697|\text{否}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{否})$$
$$= \frac{1}{\sqrt{2\pi}\cdot 0.195}\exp\left(-\frac{(0.697-0.496)^2}{2\cdot 0.195^2}\right) \approx 1.203\,,$$

$$p_{\text{含糖: }0.460|\text{是}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{是})$$
$$= \frac{1}{\sqrt{2\pi}\cdot 0.101}\exp\left(-\frac{(0.460-0.279)^2}{2\cdot 0.101^2}\right) \approx 0.788\,,$$

$$p_{\text{含糖: }0.460|\text{否}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{否})$$
$$= \frac{1}{\sqrt{2\pi}\cdot 0.108}\exp\left(-\frac{(0.460-0.154)^2}{2\cdot 0.108^2}\right) \approx 0.066\,.$$

最大似然估计

周志华，机器学习，P84,151,152

朴素贝叶斯分类器

深蓝学院
shenlanxueyuan.com

测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

$$P(\text{好瓜}=\text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}}$$

$$\times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \approx 0.063$$

$$P(\text{好瓜}=\text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}}$$

$$\times P_{\text{硬滑}|\text{否}} \times p_{\text{密度: 0.697}|\text{否}} \times p_{\text{含糖: 0.460}|\text{否}} \approx 6.80 \times 10^{-5} .$$

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”.

周志华，机器学习，P151,153

朴素贝叶斯分类器

拉普拉斯修正

测3	青绿	蜷缩	清脆	清晰	凹陷	硬滑	0.697	0.460	?

?

$$P(\text{好瓜}=\text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}}$$

$$\times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \cancel{\approx 0.063} \quad =0?$$

作用：避免其他属性所携带的信息被训练集中未出现的属性值“抹去”

$$\hat{P}(c) = \frac{N_c + 1}{N + K} \qquad P(t_i|c) = \frac{N_{(c,t_i)} + 1}{N_c + |t_i|}$$

t_i可能的取值数

朴素贝叶斯分类器

$P_{青绿|是} = P(色泽 = 青绿 \mid 好瓜 = 是) =$

$P_{青绿|否} = P(色泽 = 青绿 \mid 好瓜 = 否) =$

$P_{蜷缩|是} = P(根蒂 = 蜷缩 \mid 好瓜 = 是) =$

$P_{蜷缩|否} = P(根蒂 = 蜷缩 \mid 好瓜 = 否) =$

$P_{浊响|是} = P(敲声 = 浊响 \mid 好瓜 = 是) =$

$P_{浊响|否} = P(敲声 = 浊响 \mid 好瓜 = 否) =$

$P_{清晰|是} = P(纹理 = 清晰 \mid 好瓜 = 是) =$

$P_{清晰|否} = P(纹理 = 清晰 \mid 好瓜 = 否) =$

$P_{凹陷|是} = P(脐部 = 凹陷 \mid 好瓜 = 是) =$

$P_{凹陷|否} = P(脐部 = 凹陷 \mid 好瓜 = 否) =$

$P_{硬滑|是} = P(触感 = 硬滑 \mid 好瓜 = 是) =$

$P_{硬滑|否} = P(触感 = 硬滑 \mid 好瓜 = 否) =$

?

$P(好瓜 = 是) =$?

$P(好瓜 = 否) =$

测2	青绿	蜷缩	清脆	清晰	凹陷	硬滑			?
测3	青绿	蜷缩	清脆	清晰	凹陷	硬滑	0.697	0.460	?

? $\xrightarrow{\text{拉普拉斯修正的朴素贝叶斯}}$?

周志华，机器学习，P84,152

$$P_{青绿|是} = \frac{3+1}{8+3} \approx 0.364$$

$$P_{青绿|否} = \frac{3+1}{9+3} \approx 0.333$$

$$P_{蜷缩|是} = \frac{5+1}{8+3} \approx 0.545$$

$$P_{蜷缩|否} = \frac{3+1}{9+3} \approx 0.333$$

$$P_{清脆|是} = \frac{0+1}{8+3} \approx 0.091$$

$$P_{清脆|否} = \frac{2+1}{9+3} = 0.25$$

$$P_{浊响|是} = \frac{6+1}{8+3} \approx 0.636$$

$$P_{浊响|否} = \frac{4+1}{9+3} \approx 0.417$$

$$P_{清晰|是} = \frac{7+1}{8+3} \approx 0.727$$

$$P_{清晰|否} = \frac{2+1}{9+3} = 0.25$$

$$P_{凹陷|是} = \frac{6+1}{8+3} \approx 0.636$$

$$P_{凹陷|否} = \frac{2+1}{9+3} = 0.25$$

$$P_{硬滑|是} = \frac{6+1}{8+2} = 0.7$$

$$P_{硬滑|否} = \frac{6+1}{9+2} \approx 0.636$$

$$\hat{P}(好瓜=是) = \frac{8+1}{17+2} \approx 0.474$$

$$\hat{P}(好瓜=否) = \frac{9+1}{17+2} \approx 0.526$$

测2	青绿	蜷缩	清脆	清晰	凹陷	硬滑

$$P(好瓜=是) \times P_{青绿|是} \times P_{蜷缩|是} \times P_{清脆|是} \times P_{清晰|是} \times P_{凹陷|是} \times P_{硬滑|是} \approx 0.003$$

$$P(好瓜=否) \times P_{青绿|否} \times P_{蜷缩|否} \times P_{清脆|否} \times P_{清晰|否} \times P_{凹陷|否} \times P_{硬滑|否} \approx 0.001$$

0.003>0.001 $\xrightarrow{\text{拉普拉斯修正的朴素贝叶斯}}$ 测2是好瓜

$$P_{青绿|是}=\frac{3+1}{8+3}\approx 0.364$$
$$P_{青绿|否}=\frac{3+1}{9+3}\approx 0.333$$
$$P_{蜷缩|是}=\frac{5+1}{8+3}\approx 0.545$$
$$P_{蜷缩|否}=\frac{3+1}{9+3}\approx 0.333$$
$$P_{清脆|是}=\frac{0+1}{8+3}\approx 0.091$$
$$P_{清脆|否}=\frac{2+1}{9+3}=0.25$$
$$P_{浊响|是}=\frac{6+1}{8+3}\approx 0.636$$
$$P_{浊响|否}=\frac{4+1}{9+3}\approx 0.417$$
$$P_{清晰|是}=\frac{7+1}{8+3}\approx 0.727$$
$$P_{清晰|否}=\frac{2+1}{9+3}=0.25$$
$$P_{凹陷|是}=\frac{6+1}{8+3}\approx 0.636$$
$$P_{凹陷|否}=\frac{2+1}{9+3}=0.25$$
$$P_{硬滑|是}=\frac{6+1}{8+2}=0.7$$
$$P_{硬滑|否}=\frac{6+1}{9+2}\approx 0.636$$

$$\hat{P}(好瓜=是)=\frac{8+1}{17+2}\approx 0.474$$

$$\hat{P}(好瓜=否)=\frac{9+1}{17+2}\approx 0.526$$

测3	青绿	蜷缩	清脆	清晰	凹陷	硬滑	0.697	0.460	?

$$P(好瓜=是)\times P_{青绿|是}\times P_{蜷缩|是}\times P_{清脆|是}\times P_{清晰|是}\times P_{凹陷|是}\times P_{硬滑|是}\times P_{0.697|是}\times P_{0.460|是}\approx 0.005$$

$$P(好瓜=否)\times P_{青绿|否}\times P_{蜷缩|否}\times P_{清脆|否}\times P_{清晰|否}\times P_{凹陷|否}\times P_{硬滑|否}\times P_{0.697|否}\times P_{0.460|否}\approx 7.0*10^{-5}$$

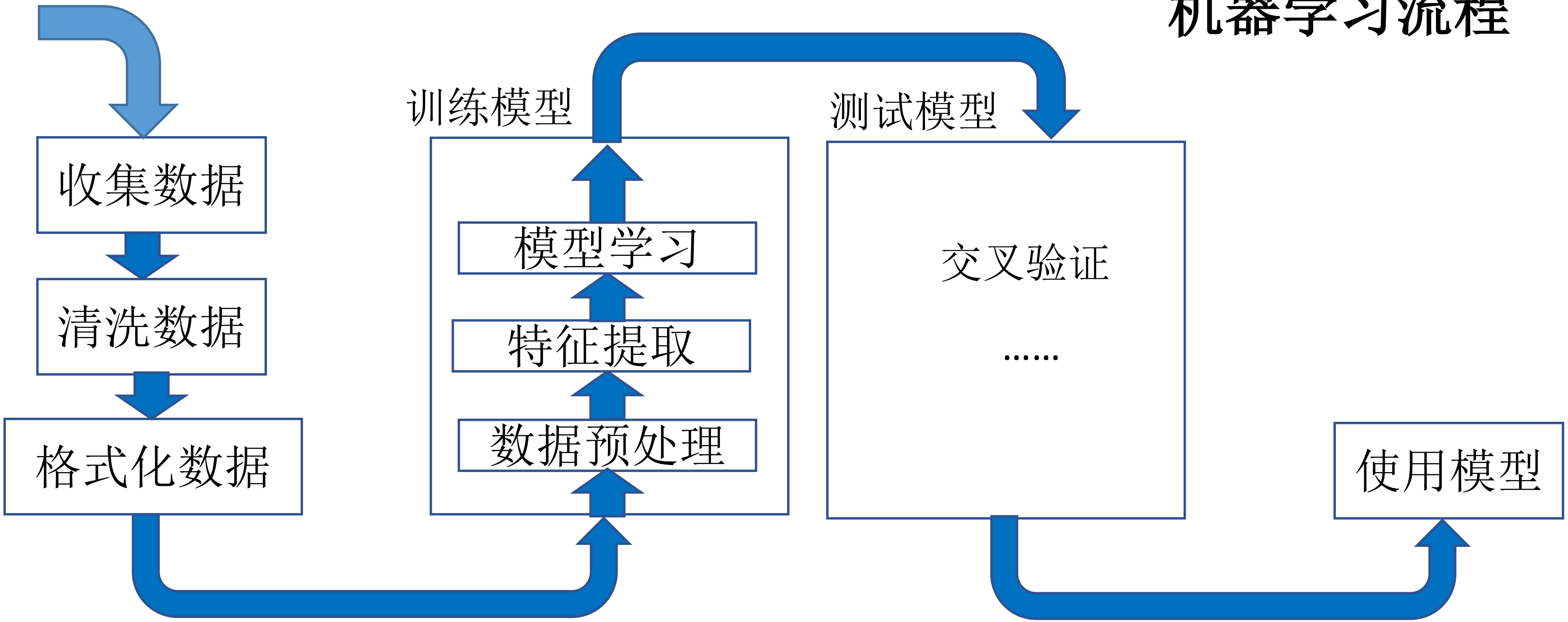
$0.005>7.0*10^{-5}$ $\xrightarrow{\text{拉普拉斯修正的朴素贝叶斯}}$ 测3是好瓜

课程大纲

- ✓ 贝叶斯决策理论基础
- ✓ 朴素贝叶斯分类器
- ✓鸢尾花分类实践

鸢尾花分类实践

机器学习流程



鸢尾花分类实践

Iris Data Set

http://archive.ics.uci.edu/ml/assets/MLimages/Large53.jpg

数目：150=50*3

特征/属性（cm）：

1. sepal length —花萼长度
2. sepal width —花萼**宽**度
3. petal length —花瓣长度
4. petal width —花瓣**宽**度

类别:

-- Iris Setosa —山鸢尾

-- Iris Versicolour —多彩鸢尾

-- Iris Virginica —弗吉尼亚鸢尾

数据下载地址：http://archive.ics.uci.edu/ml/datasets/Iris

鸢尾花分类实践

深蓝学院
shenlanxueyuan.com

Documentation

Search Documentation

CONTENTS Close

Classification

Open This Example

This example shows how to perform classification using discriminant analysis, naive Bayes classifiers, and decision trees. Suppose you have a data set containing observations with measurements on different variables (called predictors) and their known class labels. If you obtain predictor values for new observations, could you determine to which classes those observations probably belong? This is the problem of classification.

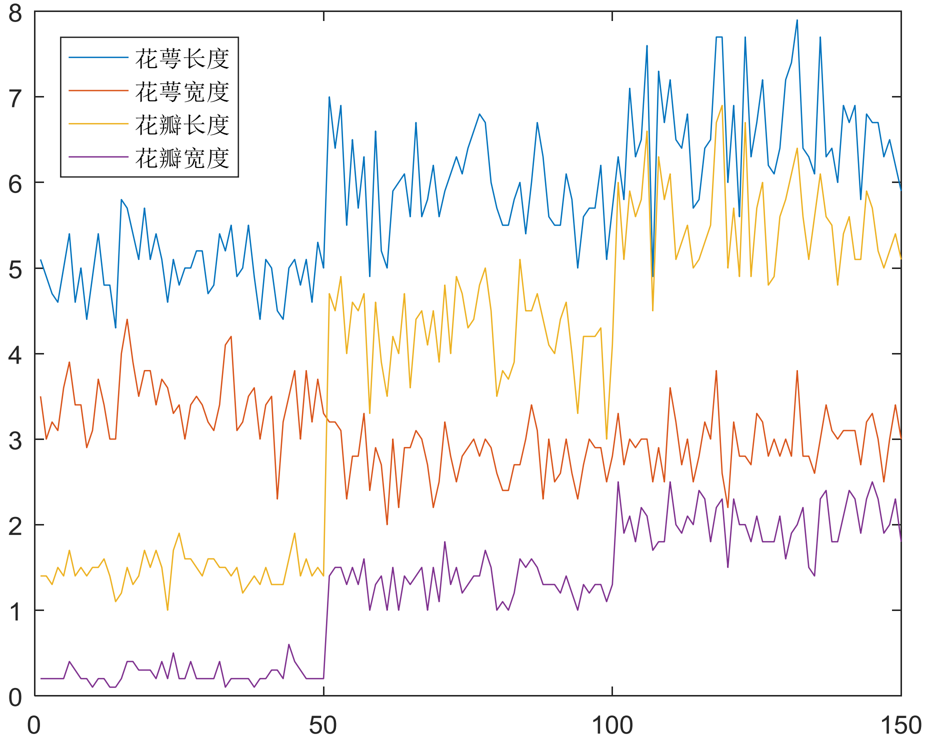
Fisher's Iris Data

Fisher's iris data consists of measurements on the sepal length, sepal width, petal length, and petal width for 150 iris specimens. There are 50 specimens from each of three species. Load the data and see how the sepal measurements differ between species. You can use the two columns containing sepal measurements.

```
load fisheriris
gscatter(meas(:,1), meas(:,2), species,'rgb','osd');
xlabel('Sepal length');
ylabel('Sepal width');
N = size(meas,1);
```

鸢尾花分类实践

```
clear all;close all;clc;
rng('default');
%%
% 导入Fisher's Iris data（鸢尾花数据）
load fisheriris;
% 显示特征取值
figure;
plot(meas);
legend('花萼长度','花萼宽度','花瓣长度','花瓣宽度','Location','NorthWest');
```



鸢尾花分类实践

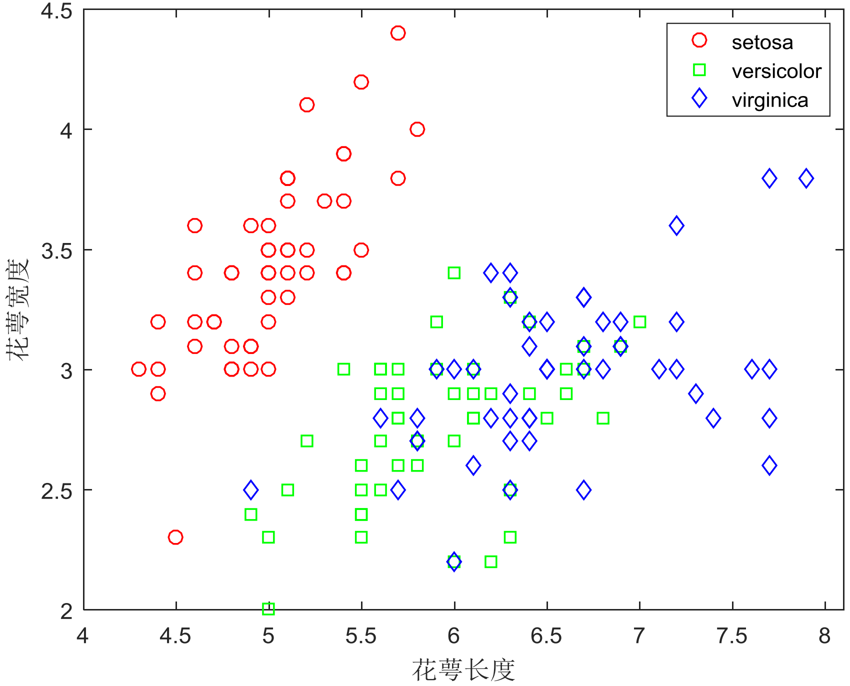
深蓝学院 shenlanxueyuan.com

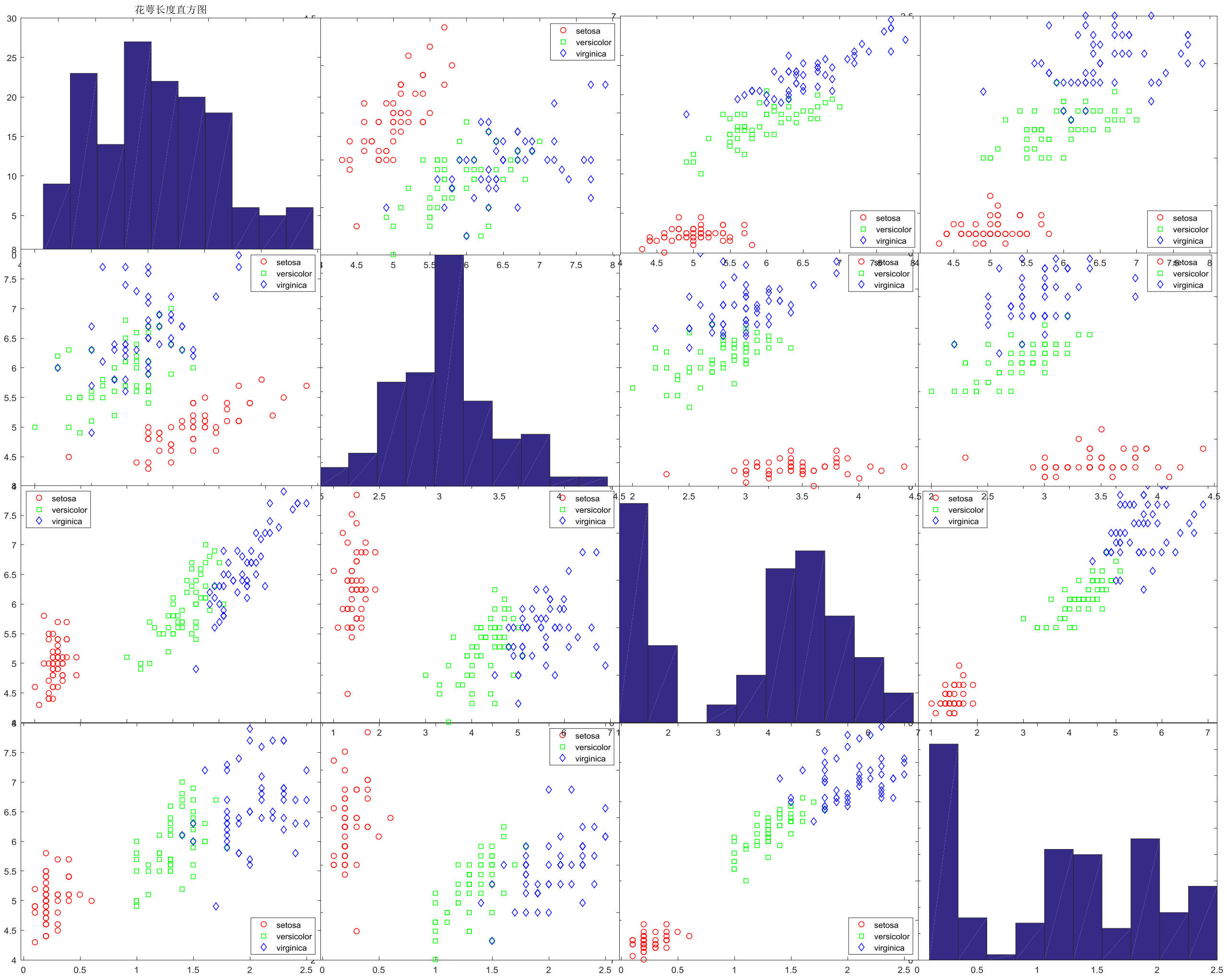
特征相关性分析

```
gscatter(meas(:,1), meas(:,2),species,
        'rgb','osd'); %探索数据
xlabel('Sepal length');
ylabel('Sepal width');
```

4个特征的协方差矩阵

0.6857	-0.0424	1.2743	0.5163
-0.0424	0.1900	-0.3297	-0.1216
1.2743	-0.3297	3.1163	1.2956
0.5163	-0.1216	1.2956	0.5810






```
for i=1:4
    for j=1:4
        if i==j
            figure;hist(meas(:,i));drawnow;
        else
            figure;
            gscatter(meas(:,i), meas(:,j), …
                        species,'rgb','osd');
        end
    end
end
```

鸢尾花分类实践

深蓝学院 shenlanxueyuan.com

分离训练和测试数据

训练数据各类别数据比例

Value	Count	Percent
setosa	33	33.00%
virginica	29	29.00%
versicolor	38	38.00%

测试数据各类别数据比例

Value	Count	Percent
versicolor	12	24.00%
setosa	17	34.00%
virginica	21	42.00%

$$P(c) = \frac{N_c}{N}$$

```
% 打乱数据排序，并保持标签对应
N = size(meas,1); %全部数据个数
randpN = randperm(N);
randp_meas = meas(randpN,:);
randp_species = species(randpN,:);
% 分离训练2/3和测试1/3数据,
train_datas = randp_meas(1:N/3*2,:);
train_labels = randp_species(1:N/3*2);
test_datas = randp_meas(1+N/3*2:end,:);
test_labels = randp_species(1+N/3*2:end);
disp('训练数据各类别数据比例');
tabulate(train_labels)
disp('测试数据各类别数据比例');
tabulate(test_labels)
```

鸢尾花分类实践

训练模型

$$P(c) = \frac{N_c}{N}$$

$$\hat{P}(c) = \frac{N_c + 1}{N + K}$$

$$P(t_i|c) = \frac{N_{(c,t_i)}}{N_c + |t_i|}$$

$$P(t_i|c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} e^{-\frac{(t_i-\mu)^2}{2\sigma_{c,i}^2}}$$

MLE

鸢尾花分类实践

深蓝学院 shenlanxueyuan.com

Matlab自带Naïve Bayes分类器函数的用法

MODEL=fitcnb (TBL,Y)

MODEL=fitcnb(X,Y,'PARAM1',val1,'PARAM2',val2,...)

```
'DistributionNames'
                  'normal','kernel','mvmn','mn'
'Kernel'        'normal', 'box', 'triangle', or 'epanechnikov'.
'Support'       'unbounded' ,'positive' ,[L,U]
'Width'         scalar,row vector,column vector,matrix
'CategoricalPredictors' - List of categorical predictors.
'ClassNames'    - Array of class names.
'Cost'          - Square matrix, where COST(I,J) is the
                cost of classifying a point into class J if its
                true class is I.
'CrossVal'      'on', 'off'
              - If 'on', performs 10-fold cross-validation.
'CVPartition'   - A partition created with CVPARTITION to use

                the cross-validated tree.
```

鸢尾花分类实践

```
'Holdout'        - Holdout validation uses the specified fraction
                   of the data for test, and uses the rest of the
                   data for training. Specify a numeric scalar
                   between 0 and 1.
'KFold'          - Number of folds to use in cross-validated tree,
                   a positive integer. Default: 10
'Leaveout'       - Use leave-one-out cross-validation by setting

                   'on'.
'PredictorNames' - A cell array of names for the predictor
                   variables, in the order in which they appear in


'Prior'          - Prior probabilities for each class.
'ResponseName'   - Name of the response variable Y, a string.
```

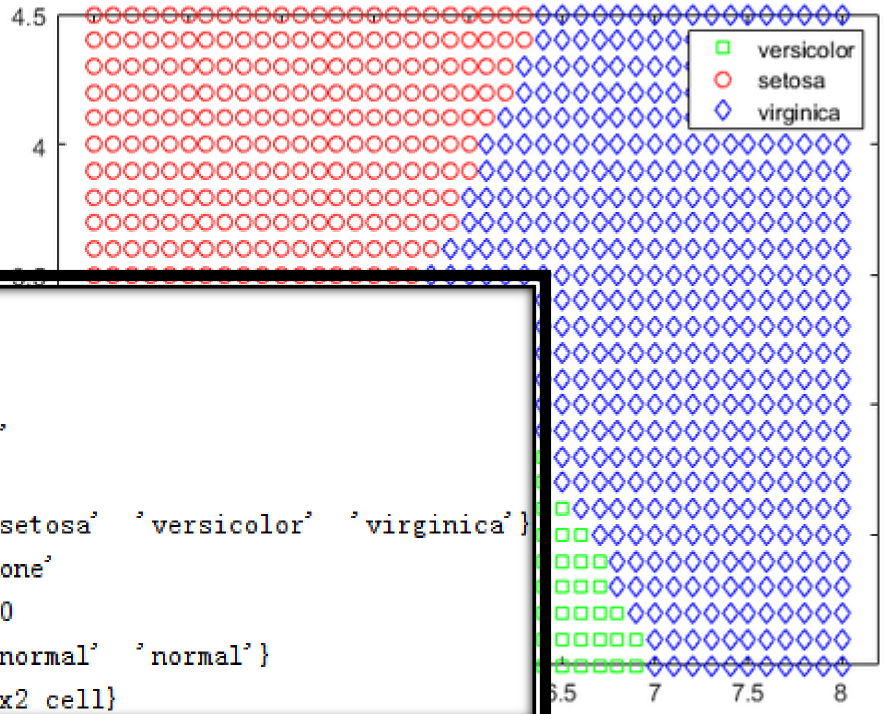
```
'ScoreTransform' - Function handle for transforming scores,

                   string representing a built-in transformation
                   function.
                  'symmetric', 'invlogit', 'ismax',
              'symmetricismax', 'none', 'logit', 'doublelogit',
              'symmetriclogit', and
                   'sign'.
'Weights'        - Vector of observation weights, one weight per
                   observation.
```

鸢尾花分类实践

深蓝学院
shenlanxueyuan.com

```
nbGau = fitcnb(meas(:,1:2), species);
nbGauResubErr = resubLoss(nbGau)
nbGauCV = crossval(nbGau, 'CVPartition',cp);
nbGauCVErr = kfoldLoss(nbGauCV)

labels = predict(nbGa
gscatter(x,y,labels,'gr
```

```
nbGauResubErr =
    0.2200
nbGauCVErr =
    0.2200
```

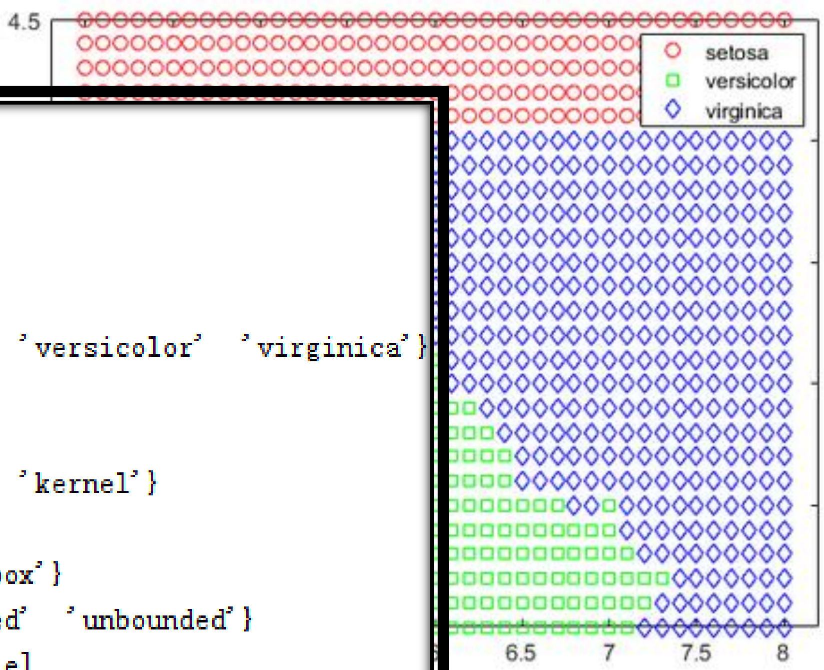



```
nbGau =

  ClassificationNaiveBayes
              ResponseName: 'Y'
     CategoricalPredictors: []
                ClassNames: {'setosa'  'versicolor'  'virginica'}
            ScoreTransform: 'none'
           NumObservations: 150
         DistributionNames: {'normal'  'normal'}
    DistributionParameters: {3x2 cell}
```

本页代码和图片来源Matlab R2016b 帮助文档

鸢尾花分类实践

nbKD = fitcnb(meas(:,1:2), species,
'DistributionNames','kernel', 'Kernel','box');
nbKDResubErr = resubLoss(nbKD) nbKDCV =
crossval(nbKD, 'CV
kfoldLoss(nbKDCV)
gscatter(x,y,labels,



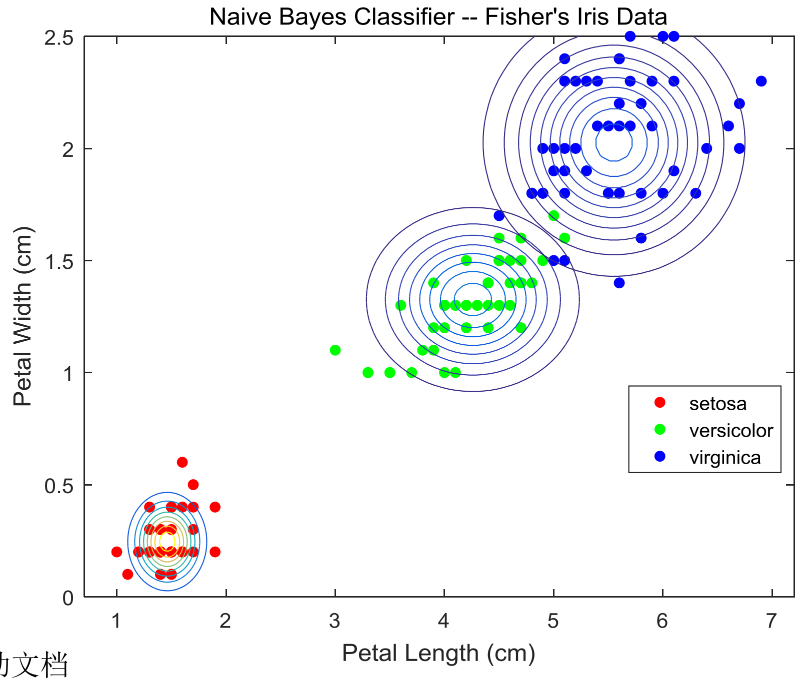

```
nbKD =

  ClassificationNaiveBayes
            ResponseName: 'Y'
   CategoricalPredictors: []
              ClassNames: {'setosa'  'versicolor'  'virginica'}
          ScoreTransform: 'none'
         NumObservations: 150
       DistributionNames: {'kernel'  'kernel'}
  DistributionParameters: {3x2 cell}
                  Kernel: {'box'  'box'}
                 Support: {'unbounded'  'unbounded'}
                   Width: [3x2 double]
```

nbKDResubErr =
0.2067
nbKDCVErr =
0.2133

本页代码和图片来源Matlab R2016b 帮助文档

鸢尾花分类实践

\Documents\MATLAB\Examples\TrainANaiveBayesClassifierFitcnbExample\
TrainANaiveBayesClassifierFitcnbExample.m

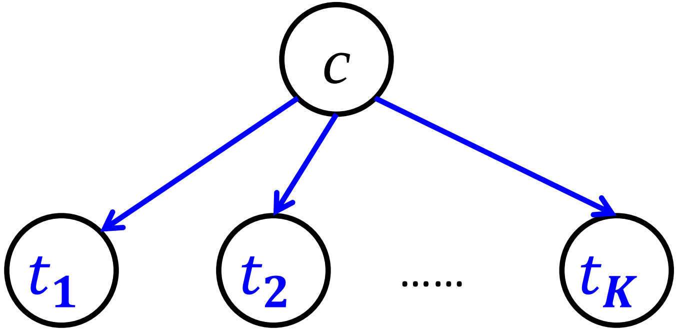


本页代码和图片来源Matlab R2016b 帮助文档

小结

朴素贝叶斯分类器
(Naïve Bayesian Classifier)

$$f_{nbc}(\boldsymbol{x}) = \underset{c \in Y}{\operatorname{argmax}} \left\{ P(c) \prod_{i=1}^{K} P(t_i|c) \right\}$$



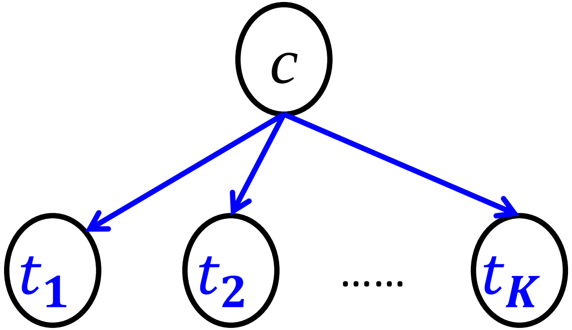
优点：

1. 只需要计算组合概率，所需估计的参数较少
2. 对数据较少/缺失数据的鲁棒性好
3. 能够充分利用领域知识和样本数据
4. 能够学习变量间的因果关系
5. 具有自我纠正能力

缺点：

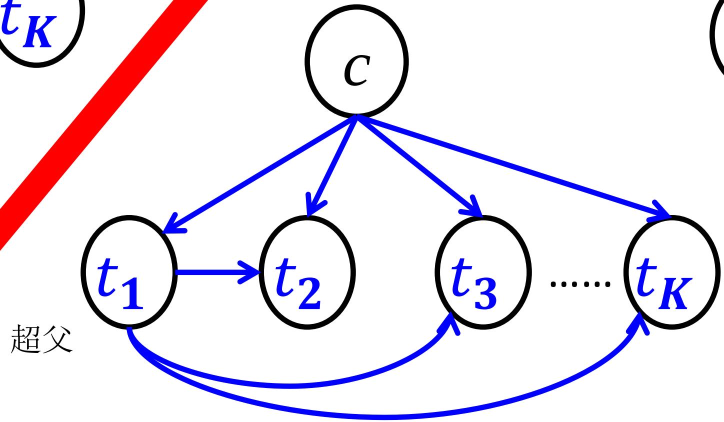
① 对于输入数据的准备方式较为敏感
② 独立假设条件在实际中可能不成立
③ 不能学习特征间的交互关系

扩展

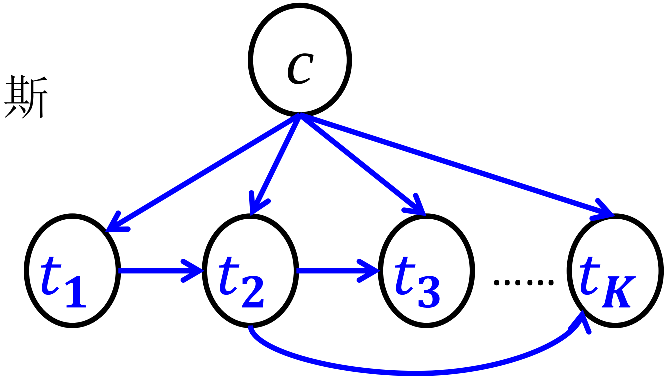


朴素贝叶斯

半朴素贝叶斯



SPODE



强相关属性的依赖

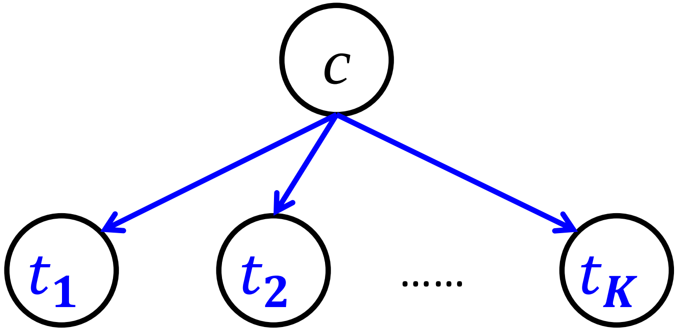
TAN

在线问答

Q&A

朴素贝叶斯分类器
(Naïve Bayesian Classifier)

$$f_{nbc}(\boldsymbol{x}) = \underset{c \in Y}{\operatorname{argmax}} \left\{ P(c) \prod_{i=1}^{K} P(t_i|c) \right\}$$



深蓝学院
shenlanxueyuan.com

深蓝学院
官网链接

感谢各位聆听！

Thanks for Listening

附录：
二维高斯分布示意图