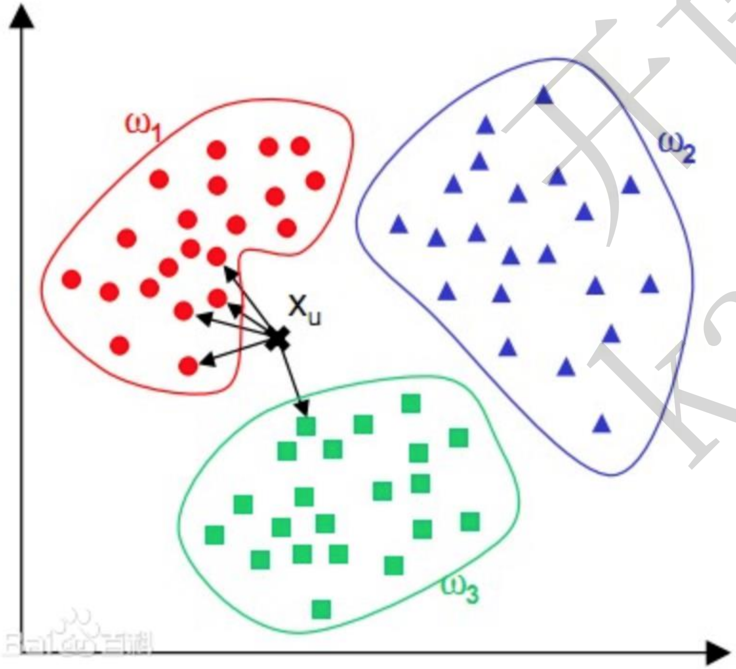
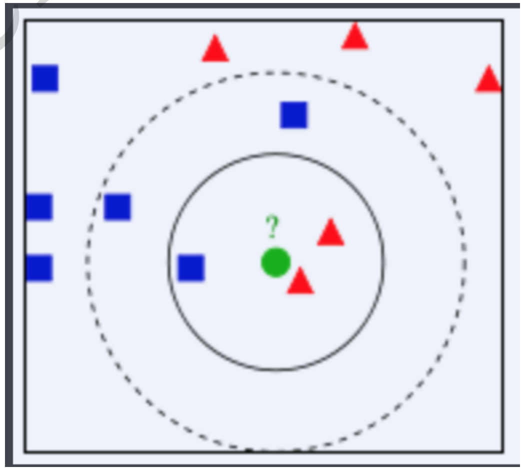
KNN

- K近邻算法简介
- K近邻原理
- 常用距离

开课吧
kaikeba

- KNN作为一种有监督分类算法，是最简单的机器学习算法之一，顾名思义，其算法主体思想就是根据距离相近的邻居类别，来判定自己的所属类别。

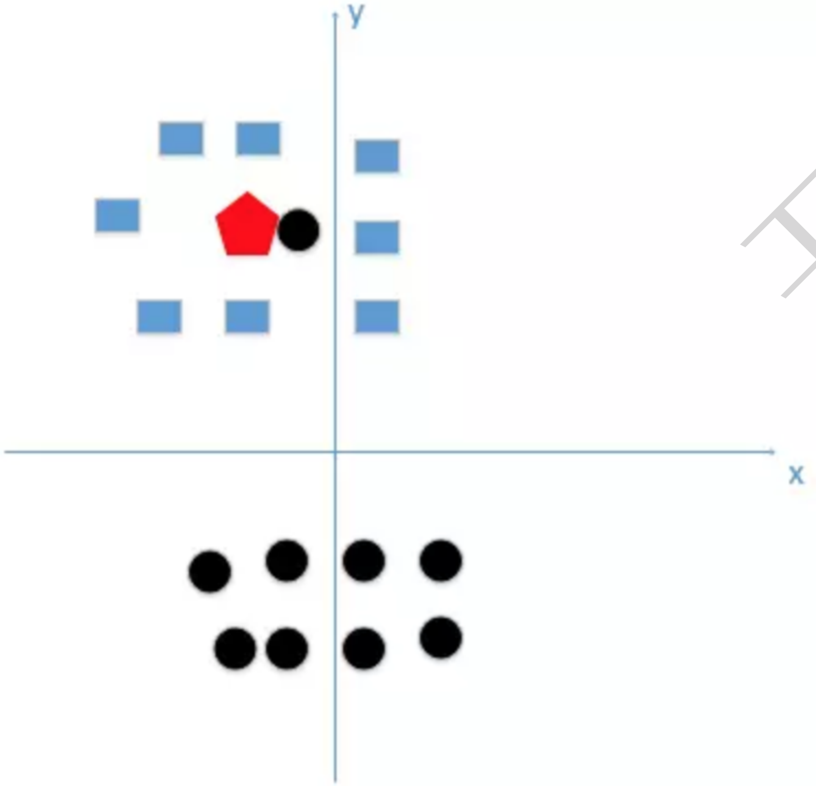


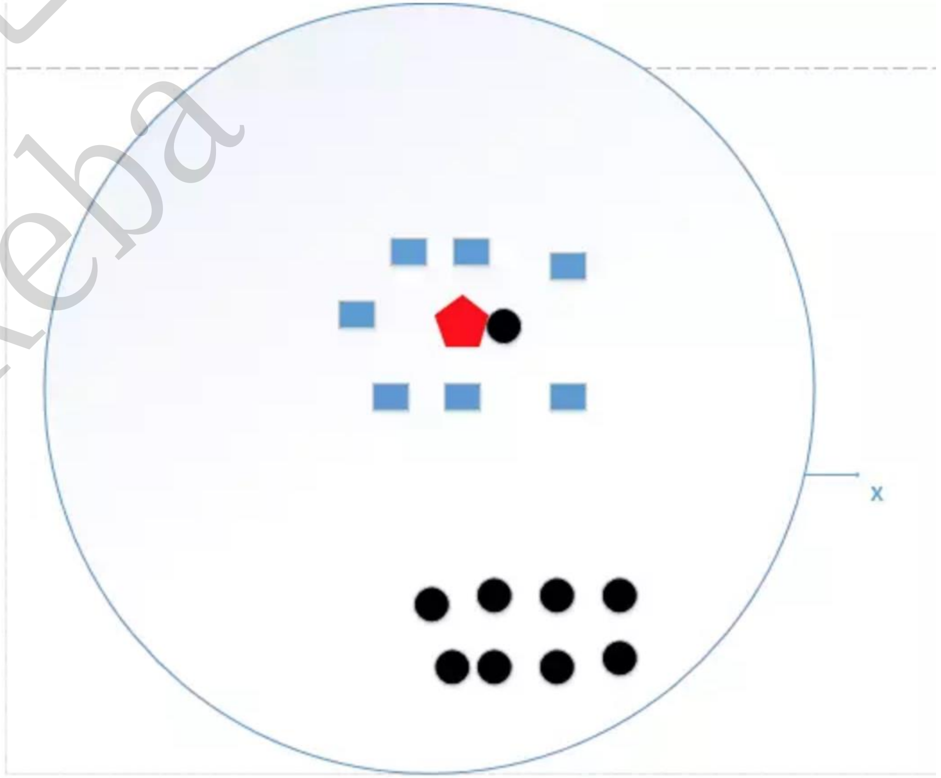


开课吧 kaikeba

- KNN算法步骤：
- （1）收集数据：可以采用公开的数据源
- （2）准备数据：计算距离所需要的数值
- （3）分析数据：剔除垃圾信息
- （4）测试算法：计算错误率
- （5）使用算法：运用在实际中，对实际情况进行预测

• 过拟合与欠拟合（K的选取很重要）





开课吧
kaikeba

我们一般选取一个较小的数值，通常采取 交叉验证法来选取最优的k值。

- 统一量纲

  - A [(179,42),男]、 B [(178,43),男]、 C [(165,36)女]、D [(177,42),男] 、E [(160,35),女]

$$AF = \sqrt{(167-179)^2 + (43-42)^2} = \sqrt{145}$$

$$BF = \sqrt{(167-178)^2 + (43-43)^2} = \sqrt{121}$$

$$CF = \sqrt{(167-165)^2 + (43-36)^2} = \sqrt{53}$$

$$DF = \sqrt{(167-177)^2 + (43-42)^2} = \sqrt{101}$$

$$EF = \sqrt{(167-160)^2 + (43-35)^2} = \sqrt{103}$$

因为由于各个特征量纲的不同，在这里导致了身高的重要性已经远远大于脚码。

开课吧 kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

- 它的的思路是：如果一个样本在特征空间中的k个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。

- 相似如何定义？

丑小鸭定理说明这个问题很不好确定。

丑小鸭定理：上个世纪60年代，模式识别研究的鼻祖之一，美籍日本学者渡边慧证明了“丑小鸭定理”。这个定理说的是“丑小鸭与白天鹅之间的区别和两只白天鹅之间的区别 一样大”。

开课吧 kaikeba

- 1. 欧氏距离(Euclidean Distance)
  - 欧氏距离是最容易直观理解的距离度量方法，我们小学、初中和高中接触到的两个点在空间中的距离一般都是指欧氏距离。
  - 直线距离
  - 连续数值

开课吧
kaikeba

• 2.曼哈顿距离

- 曼哈顿街区要从一个十字路口开车到另一个十字路口，驾驶距离显然不是两点间的直线距离。这个实际驾驶距离就是“曼哈顿距离”。曼哈顿距离也称为“城市街区距离”

$$d = |X_i - X_j| + |Y_i - Y_j|$$

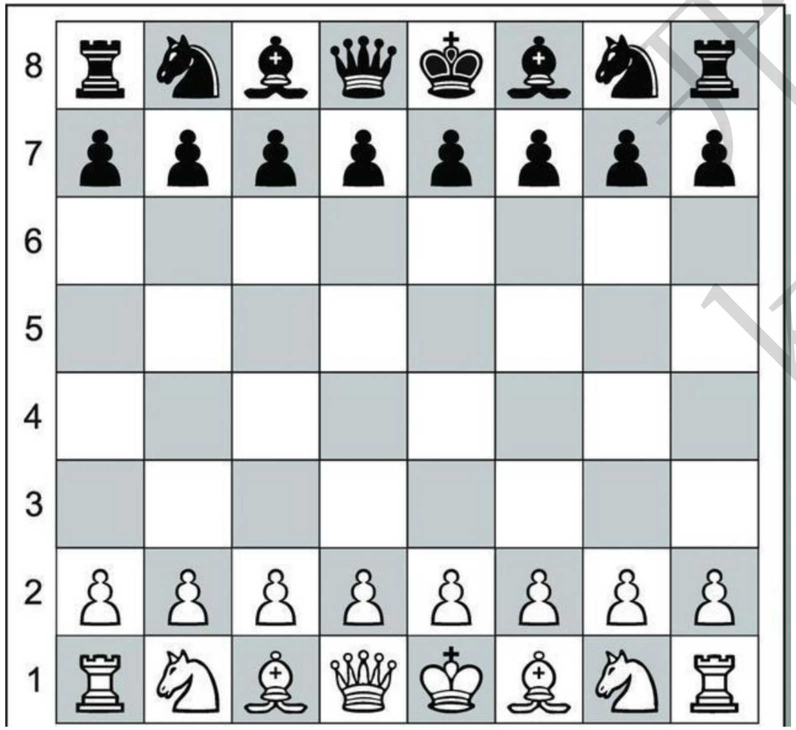
应用在：城市导航，道路规划

开课吧 kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

• 切比雪夫距离

- 国际象棋中，国王可以直行、横行、斜行，所以国王走一步可以移动到相邻8个方格中的任意一个。国王从格子(x1,y1)走到格子(x2,y2)最少需要多少步？这个距离就叫切比雪夫距离。



$$d_{12} = \max(|x_1 - x_2|, |y_1 - y_2|)$$

棋类游戏

开课吧 kaikeba

- 汉明距离
  - 表示两个等长字符串在对应位置上不同字符的数目，我们以d(x, y)表示字符串x和y之间的汉明距离。
  - 1011101 与 1001001 之间的汉明距离是 2。
  - 2143896 与 2233796 之间的汉明距离是 3。
  - "toned" 与 "roses" 之间的汉明距离是 3。

开课吧
kaikeba

- 编辑距离
  - 指两个字串之间，由一个转成另一个所需的最少编辑操作次数，如果它们的距离越大，说明它们越是不同。许可的编辑操作包括将一个字符替换成另一个字符，插入一个字符，删除一个字符。

		b	e	a	u	t	y
	0	1	2	3	4	5	6
b	1	0	1	2	3	4	5
a	2	1	1	1	2	3	4
t	3	2	2	2	2	2	3
y	4	3	3	3	3	3	2
u	5	4	4	4	3	4	3

动态规划算法

d[i,j]=min(d[i-1,j]+1 、d[i,j-1]+1、d[i-1,j-1]+temp)

文本纠错、字符相似等

开课吧 kaikeba

- # 余弦距离
  - 余弦相似度，又称为余弦相似性，是通过计算两个向量的夹角余弦值来评估他们的相似度。余弦相似度将向量根据坐标值，绘制到向量空间中，如最常见的二维空间。

$$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

协同过滤计算相似度、词语相似度等

开课吧
kaikeba

- 杰卡德距离
  - 杰卡德距离(Jaccard Distance) 是用来衡量两个集合差异性的一种指标，它是杰卡德相似系数的补集，被定义为1减去Jaccard相似系数。而杰卡德相似系数(Jaccard similarity coefficient)，也称杰卡德指数(Jaccard Index)，是用来衡量两个集合相似度的一种指标。

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

关联分析中

开课吧 kaikeba

皮尔逊系数

- 相关系数衡量随机变量X与Y相关程度的一种方法，相关系数的取值范围是[-1,1]。相关系数的绝对值越大，则表明X与Y相关度越高。

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-EX)(Y-EY))}{\sqrt{D(X)}\sqrt{D(Y)}}$$

开课吧 kaikeba

• Knn特点：

- 优点
- 简单，易理解，易实现。
- 适合对稀有事件进行分类。
- 适合多分类问题。

- 缺点
- 当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。
- 计算量较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的K个最近邻点。

开课吧
kaikeba

决策树

开课吧 kaikeba

- 1. 什么是决策树

- 2.信息熵回顾

- 3. ID3.0, C4.5, CART

- 4. 过拟合与欠拟合

开课吧
kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

什么是决策树

- 决策树来自决策论, 由多个决策分支和可能的结果（包括资源成本和风险）组成，用来创建到达目标的规划；
- 也可以用来表示算法。

- 分类预测：决策树表示
- 决策树的表示？

开课吧
kaikeba

二分类学习任务
属性
属性值

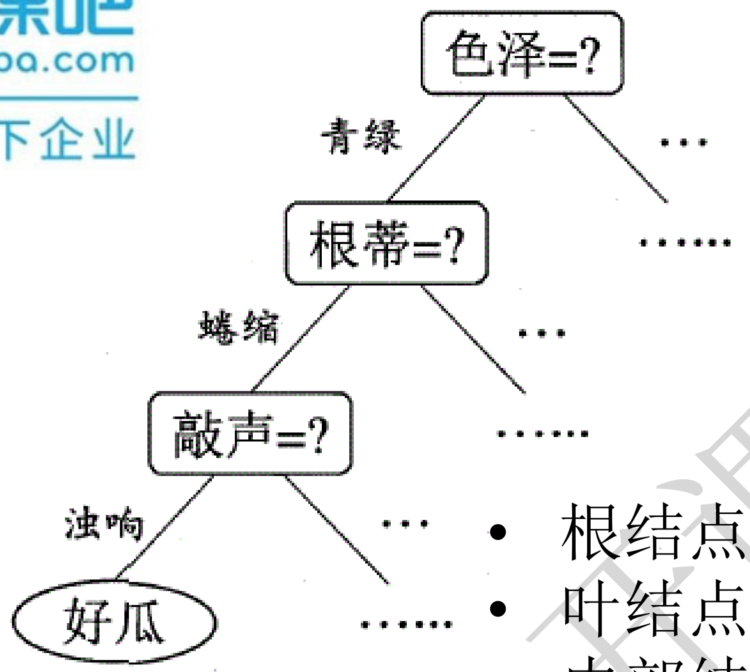


图 4.1 西瓜问题的一棵决策树

- 根结点：包含全部样本
- 叶结点：对应决策结果 “好瓜” “坏瓜”
- 内部结点：对应属性测试

决策树学习的目的：为了产生一颗泛化能力强的决策树，即处理未见示例能力强。

决策树的核心问题

- 决策树的生成→对训练样本进行分组
  - 关键，确定树根节点和分支准则
  - 停止生长时机
- 决策树的修剪→解决过度拟合问题
  - 预先修剪，限值决策树的充分生长，如：限制树的高度
  - 滞后修剪，待决策树充分生长完毕后再进行修剪
    - 当节点和分支数较多时，显然不合适

开课吧 kaikeba

信息熵

香农提出了“信息熵”的概念，解决了对信息的量化度量问题。

香农用“信息熵”的概念来描述信源的不确定性。

“信息熵”(information entropy)是度量样本集合纯度最常用的一种指标. 假定当前样本集合 D 中第 k 类样本所占的比例为 p_k $(k=1,2,\ldots,|\mathcal{Y}|)$, 则 D 的信息熵定义为

$$\operatorname{Ent}(D)=-\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k \,. \tag{4.1}$$

$\operatorname{Ent}(D)$ 的值越小, 则 D 的纯度越高. 对于二分类任务 $|y|=2$

开课吧 kaikeba

假设我们已经知道衡量不确定性大小的这个量已经存在了，不妨就叫做“**信息量**”

- 不会是负数
- 不确定性函数f是概率的p单调递减函数；
- 可加性：两个独立符号所产生的不确定性应等于各自不确定性之和，即

$$f(p_1 \times p_2) = f(p_1) + f(p_2)$$

同时满足这三个条件的函数 f 是负的对数函数，即

$$f(p_i) = \log \frac{1}{p_i} = -\log p_i$$

一个事件的**信息量**就是这个事件发生的概率的负对数。

信息熵是跟所有事件的可能性有关的，是平均而言发生一个事件得到的信息量大小。所以信息熵其实是信息量的期望。

$$E[-\log p_i] = -\sum_{i=1}^{n} p_i \log p_i$$

开课吧
kaikeba

信息熵

- 若一事件有k种结果，对应的概率为P_i。则此事件发生后所得到的信息量I(视为Entropy)为：
  I=-($p_1*log_2(p_1)+p_2*log_2(p_2)+\ldots+p_k*log_2(p_k)$)
- 世界杯四强

- 设 k=4→p_1=0.25,p_2=0.25,p_3=0.25,p_4=0.25
  I=-(.25*log_2(.25)*4)=2
- 设k=4→p_1=0,p_2=0.5,p_3=0,p_4=0.5
  I=-(.5*log_2(.5)*2)=1
- 设 k=4→p_1=1,p_2=0,p_3=0,p_4=0
  I=-(1*log_2(1))=0

开课吧 kaikeba

决策树算法 ID3

ID3算法主要针对属性选择问题。是决策树学习方法中最具影响和最为典型的算法。

该方法使用信息增益度选择测试属性。

当获取信息时，将不确定的内容转为确定的内容，因此信息伴着不确定性。

从直觉上讲，小概率事件比一般概率事件包含的信息量大。如果某件事情是“百年一见”则肯定比“习以为常”的事件包含的信息量大。

如何度量信息量的大小？

开课吧 kaikeba

- 用信息增益度量熵的降低程度
  - 属性A 的信息增益，使用属性A分割样例集合S 而导致的熵的降低程度

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

开课吧 kaikeba

举例：求解划分根结点的最优划分属性

数据集包含17个训练样例：

8个正例（好瓜）占 $p_1 = \frac{8}{17}$

对于二分类任务 $|y| = 2$

9个反例（坏瓜）占 $p_2 = \frac{9}{17}$

以属性“色泽”为例计算其信息增益

根结点的信息熵：

$$\mathrm{Ent}(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{8}{17}\log_2\frac{8}{17} + \frac{9}{17}\log_2\frac{9}{17}\right) = 0.998\ .$$

开课吧 kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

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

开课吧
kaikeba

用“色泽”将根结点划分后获得3个分支结点的信息熵分别为：

$$\mathrm{Ent}(D^1) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.000 ,$$

$$\mathrm{Ent}(D^2) = -\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right) = 0.918 ,$$

$$\mathrm{Ent}(D^3) = -\left(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right) = 0.722 ,$$

属性“色泽”的信息增益为：

$$\begin{aligned}\mathrm{Gain}(D,\text{色泽}) &= \mathrm{Ent}(D) - \sum_{v=1}^{3}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)\\ &= 0.998 - \left(\frac{6}{17}\times 1.000 + \frac{6}{17}\times 0.918 + \frac{5}{17}\times 0.722\right)\\ &= 0.109 .\end{aligned}$$

开课吧 kaikeba

• 连续数值离散化

样本集 D

连续属性 a ，有n个不同的取值，将n个取值从小到大排序：

$$\{a^1, a^2, ..., a^n\}$$

划分点t（**数值**）将 D 划分为两个子集 D_t^- 和 D_t^+

$$\underbrace{\{a^1, a^2, ..., a^i}_{D_t^-}, \underbrace{a^{i+1}, ..., a^n\}}_{D_t^+}$$

t

a^i a^{i+1}

$$[a^i, a^{i+1})$$

与离散属性不同，若当前结点划分属性为连续属性，该连续属性还可被再次选作后代结点的最优划分属性。

密度	好瓜
0.243	否
0.245	否
0.343	否
0.360	否
0.403	是1
0.437	是2
0.481	是3
0.556	是4
0.593	否
0.680	是5
0.634	是6
0.639	否
0.657	否
0.666	否
0.697	是7
0.719	否
0.774	是8

根结点包含17个训练样本，密度有17个不同取值
候选划分点集合包含16个候选值
每一个划分点能得到一个对应的信息增益

$$\mathrm{Gain}(D,a) = \max_{t\in T_a} \mathrm{Gain}(D,a,t)$$

$$= \max_{t\in T_a} \mathrm{Ent}(D) - \sum_{\lambda\in\{-,+\}} \frac{|D_t^\lambda|}{|D|}\mathrm{Ent}(D_t^\lambda),$$

根结点的信息熵仍为：$Ent(D)=0.998$

$$Ent(D_t^-) = -(\frac{0}{4}\log_2\frac{0}{4}+\frac{4}{4}\log_2\frac{4}{4}) = 0$$

$$Ent(D_t^+) = -(\frac{8}{13}\log_2\frac{8}{13}+\frac{5}{13}\log_2\frac{5}{13}) = 0.961$$

$$Gain(D,\text{密度},0.381)$$

$$= Ent(D) - [\frac{4}{17}\times Ent(D_t^-) + \frac{13}{17}\times Ent(D_t^+)]$$

$$= 0.263$$

决策树停止生长条件

节点达到完全纯度

树的深度达到用户所要的深度

节点中样本个数少于用户指定个数

指标下降的最大幅度小于用户指定的幅度

开课吧
kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

C4.5决策树

$$\mathrm{Gain\_ratio}(D,a)=\frac{\mathrm{Gain}(D,a)}{\mathrm{IV}(a)}\,, \tag{4.3}$$

其中

$$\mathrm{IV}(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|} \tag{4.4}$$

称为属性 a 的"固有值"(intrinsic value) [Quinlan, 1993]. 属性 a 的可能取值数目越多(即 V 越大), 则 $\mathrm{IV}(a)$ 的值通常会越大. 例如, 对表 4.1 的西

开课吧
kaikeba

增益率准则对可取值数目较少的属性有所偏好

著名的C4.5决策树算法综合了**信息增益准则**和**信息率准则**的特点：先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益比最高的。

开课吧
kaikeba.com
慧科集团旗下企业

天气	温度	湿度	风速	活动
晴	炎热	高	弱	取消
晴	炎热	高	强	取消
阴	炎热	高	弱	进行
雨	适中	高	弱	进行
雨	寒冷	正常	弱	进行
雨	寒冷	正常	强	取消
阴	寒冷	正常	强	进行
晴	适中	高	弱	取消
晴	寒冷	正常	弱	进行
雨	适中	正常	弱	进行
晴	适中	正常	强	进行
阴	适中	高	强	进行
阴	炎热	正常	弱	进行
雨	适中	高	强	取消

Info(D) = -9/14 * log2(9/14) - 5/14 * log2(5/14) = 0.940

Info(风速) = 6/14 * [- 3/6 * log2(3/6) – 3/6 * log2(3/6)] +
8/14 * [- 6/8 * log2(6/8) - 2/8 * log2(2/8)]
= 0.892

Gain(风速) = Info(D) - Info(WINDY) = 0.940 - 0.892 = 0.048

H(风速) = - 6/14 * log2(6/14) - 8/14 * log2(8/14) =
0.9852281360342516

IGR(风速) = Info(风速) / H(风速) = 0.048/0.9852281360342516
= 0.048719680492692784

开课吧 kaikeba

CART分类树算法

基尼值

$$\mathrm{Gini}(D) = \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'}$$

$$= 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2 \,. \tag{4.5}$$

直观来说, Gini(D) 反映了从数据集 D 中随机抽取两个样本, 其类别标记不一致的概率. 因此, Gini(D) 越小, 则数据集 D 的纯度越高.

基尼指数

$$\mathrm{Gini\_index}(D, a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} \mathrm{Gini}(D^v) \,. \tag{4.6}$$

于是, 我们在候选属性集合 A 中, 选择那个使得划分后基尼指数最小的属性作为最优划分属性, 即 $a_* = \underset{a \in A}{\arg\min}\ \mathrm{Gini\_index}(D, a)$.

tid	有房者	婚姻状况	年收入	拖欠贷款者
1	是	单身	125K	否
2	否	已婚	100K	否
3	否	单身	70K	否
4	是	已婚	120K	否
5	否	离异	95K	是
6	否	已婚	60K	否
7	是	离异	220K	否
8	否	单身	85K	是
9	否	已婚	75K	否
10	否	单身	90K	是

拖欠贷款与否为标签

有房者、婚姻状况、年收入为特征

开课吧
kaikeba

	有房	无房
否	3	4
是	0	3

Gini(有房)=1-(3/3)²-(0/3)²=0

Gini(无房)=1-(4/7)²-(3/7)²=0.4849

Gini=0.3×0+0.7×0.4898=0.343

开课吧
kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

	单身或已婚	离异
否	6	1
是	2	1

$Gini(t_1)=1-(6/8)^2-(2/8)^2=0.375$

$Gini(t_2)=1-(1/2)^2-(1/2)^2=0.5$

$Gini=8/10\times0.375+2/10\times0.5=0.4$

	单身或离异	已婚
否	3	4
是	3	0

$Gini(t_1)=1-(3/6)^2-(3/6)^2=0.5$

$Gini(t_2)=1-(4/4)^2-(0/4)^2=0$

$Gini=6/10\times0.5+4/10\times0=0.3$

	离异或已婚	单身
否	5	2
是	1	2

$Gini(t_1)=1-(5/6)^2-(1/6)^2=0.2778$

$Gini(t_2)=1-(2/4)^2-(2/4)^2=0.5$

$Gini=6/10\times0.2778+4/10\times0.5=0.3667$

开课吧 kaikeba

	60	70	75	85	90	95	100	120	125	220								
	65		72		80		87		92		97		110		122		172	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
是	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0
否	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1
Gini	0.400		0.375		0.343		0.417		0.400		0.300		0.343		0.375		0.400	

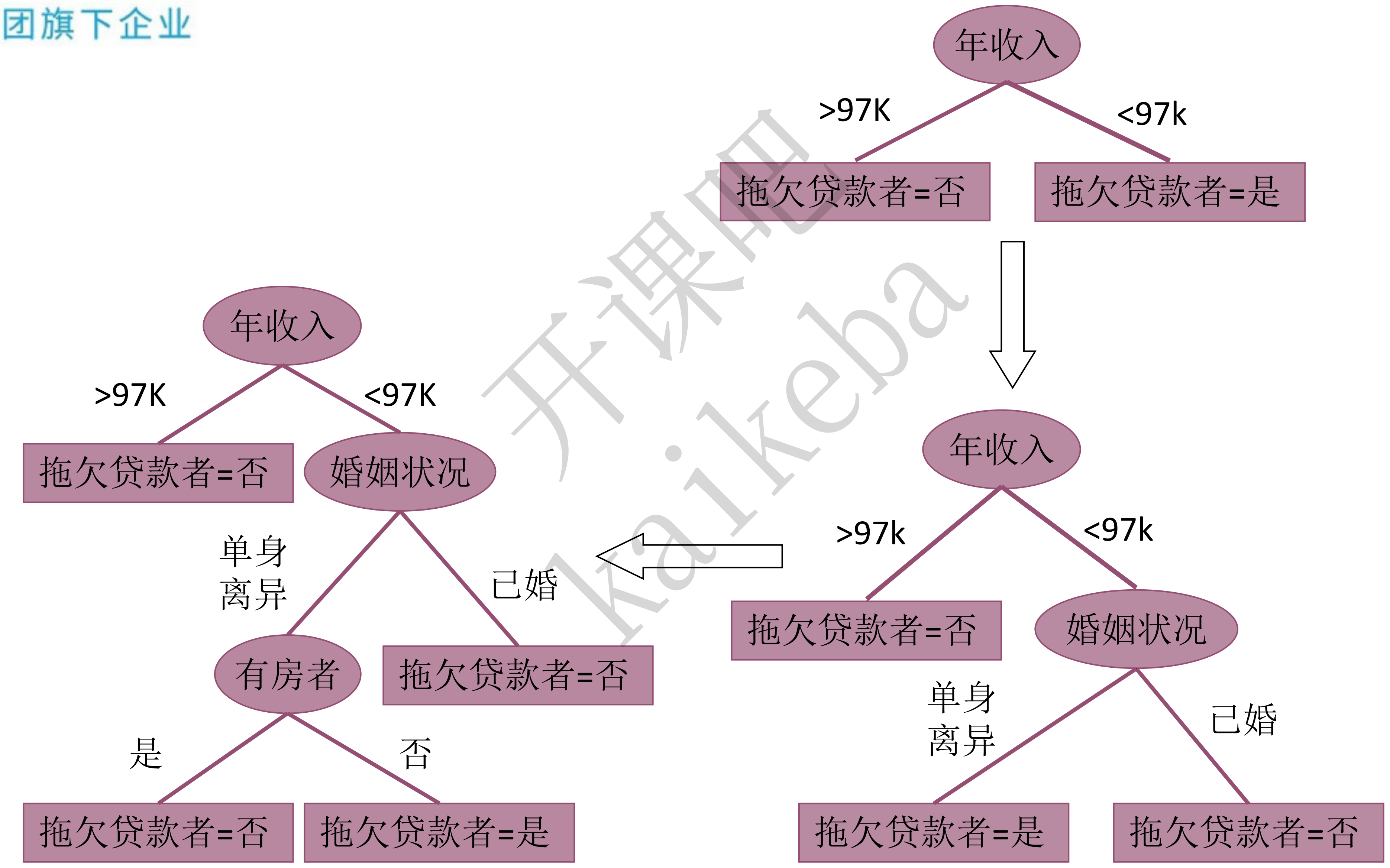
	<=65	>65
是	0	3
否	1	6

Gini(有房)=1-(1/1)²-(0/1)²=0

Gini(无房)=1-(3/9)²-(6/9)²=0.4444

Gini=0.1×0+0.9×0.4444=0.4000

开课吧
kaikeba.com
慧科集团旗下企业



开课吧
kaikeba

开课吧 kaikeba.com 慧科集团旗下企业

• CART回归树

决策树的生成就是递归地构建二叉决策树的过程. 对回归树用平方误差最小化准则，对分类树用基尼指数（Gini index）最小化准则，进行特征选择，生成二叉树.

1. 回归树的生成

假设 X 与 Y 分别为输入和输出变量，并且 Y 是连续变量，给定训练数据集

$$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right]$$

为了便于理解，下面举一个简单实例。训练数据见下表，目标是得到一棵最小二乘回归树。

x	1	2	3	4	5	6	7	8	9	10
y	5.56	5.7	5.91	6.4	6.8	7.05	8.9	8.7	9	9.05

$$\min_{js}[\min_{c_1} Loss(y_i, c_1) + \min_{c_2} Loss(y_i, c_2)]$$

例如，取 $s = 1.5$。此时 $R_1 = \{1\}, R_2 = \{2,3,4,5,6,7,8,9,10\}$，这两个区域的输出值分别为：
$c_1 = 5.56, c_2 = \frac{1}{9}(5.7 + 5.91 + 6.4 + 6.8 + 7.05 + 8.9 + 8.7 + 9 + 9.05) = 7.50$。得到下表：

s	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
c_1	5.56	5.63	5.72	5.89	6.07	6.24	6.62	6.88	7.11
c_2	7.5	7.73	7.99	8.25	8.54	8.91	8.92	9.03	9.05

开课吧
kaikeba

把c_1, c_2的值代入到上式，如：$m(1.5) = 0 + 15.72 = 15.72$。同理，可获得下表：

s	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5
m(s)	15.72	12.07	8.36	5.78	3.91	1.93	8.01	11.73	15.74

显然取 $s = 6.5$时，$m(s)$最小。因此，第一个划分变量$j = x, s = 6.5$

对R_1继续进行划分:

x	1	2	3	4	5	6
y	5.56	5.7	5.91	6.4	6.8	7.05

取切分点$[1.5, 2.5, 3.5, 4.5, 5.5]$，则各区域的输出值$c$如下表

s	1.5	2.5	3.5	4.5	5.5
c_1	5.56	5.63	5.72	5.89	6.07
c_2	6.37	6.54	6.75	6.93	7.05

开课吧 kaikeba

计算m(s)：

s	1.5	2.5	3.5	4.5	5.5
m(s)	1.3087	0.754	0.2771	0.4368	1.0644

s=3.5时m(s)最小。

假设在生成**3个区域**之后停止划分，那么最终生成的回归树形式如下：

$$T = \begin{cases} 5.72 & x \leq 3.5 \\ 6.75 & 3.5 \leqslant x \leq 6.5 \\ 8.91 & 6.5 < x \end{cases}$$

开课吧 kaikeba

剪枝，即通过主动去掉一些分支来降低过拟合的风险。

决策树的剪枝策略 $\begin{cases} \textbf{预剪枝} \\ \textbf{后剪枝} \end{cases}$

预剪枝：在决策树生成过程中，对每个结点在划分前先进行估计，若当前结点的划分不能带来决策树泛化性能提升，则停止划分并将当前结点标记为叶结点

后剪枝：先从训练集生成一棵完整的决策树，然后自底向上地对非叶结点进行考察，若将该结点对应的子树替换为叶结点能带来决策树泛化性能提升，则将该子树替换为叶结点。

训练集S和测试集T

$$D = S \cup T \text{ 且 } S \cap T = \varnothing$$

开课吧 kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

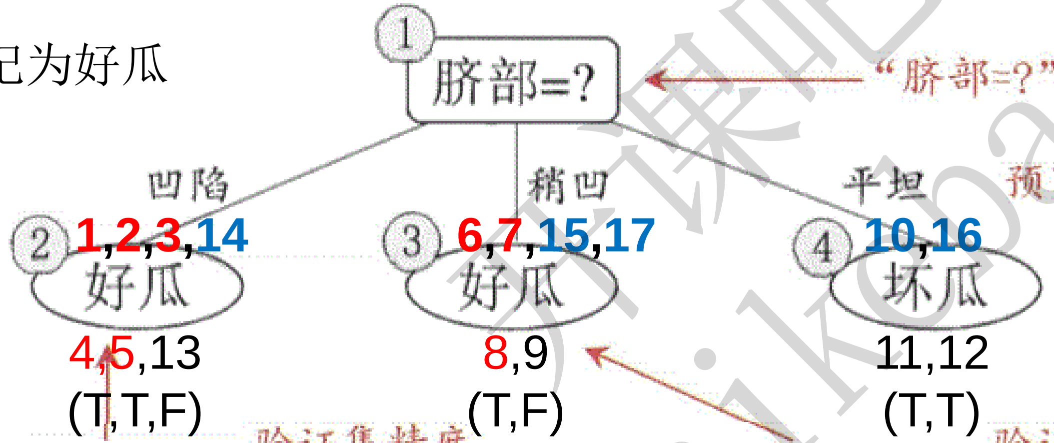
预剪枝

精度：正确分类的样本占所有样本的比例

训练集：好瓜 坏瓜
1,2,3,6,7,10,14,15,16,17

验证集：4,5,8,9,11,12,13

该结点假设标记为好瓜



验证集精度
"脐部=?" 划分前：42.9% $=\frac{3}{7}$
划分后：71.4% $=\frac{5}{7}$
预剪枝决策：划分

验证集精度
"色泽=?" 划分前：71.4%
划分后：57.1%
预剪枝决策：禁止划分

验证集精度
"根蒂=?" 划分前：71.4%
划分后：71.4%
预剪枝决策：禁止划分

开课吧 kaikeba

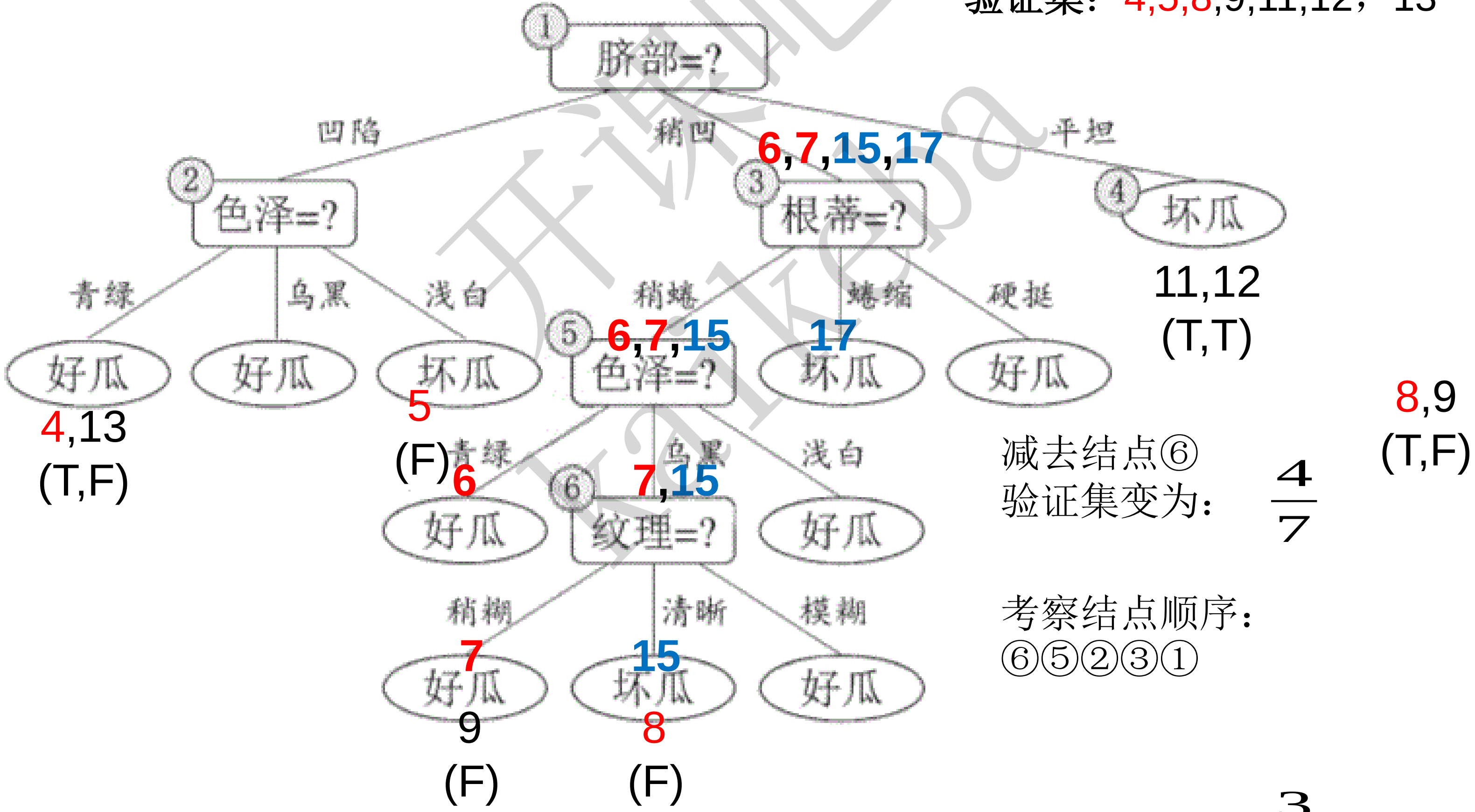
先从训练集生成一棵完整的决策树，然后自底向上地对非叶结点进行考察，若将该结点对应的子树替换为叶结点能带来决策树泛化性能提升，则将该子树替换为叶结点。

后剪枝

训练集：好瓜 坏瓜

1,2,3,6,7,10,14,15,16,17

验证集：4,5,8,9,11,12，13



减去结点⑥
验证集变为： $\frac{4}{7}$

考察结点顺序：
⑥⑤②③①

验证集精度： $\frac{3}{7}$

CART采用CCP（代价复杂度）剪枝方法。代价复杂度选择节点表面误差率增益值最小的非叶子节点，删除该非叶子节点的左右子节点，若有多个非叶子节点的表面误差率增益值相同小，则选择非叶子节点中子节点数最多的非叶子节点进行剪枝。

可描述如下：

令决策树的非叶子节点为$\{T_1,T_2,T_3,\ldots\ldots,T_n\}$。

a）计算所有非叶子节点的表面误差率增益值$\alpha=\{\alpha_1,\alpha_1,\alpha_1,\ldots\ldots,\alpha_n\}$

b）选择表面误差率增益值α_i最小的非叶子节点T_i（若多个非叶子节点具有相同小的表面误差率增益值，选择节点数最多的非叶子节点）。

c）对T_i进行剪枝

表面误差率增益值的计算公式：

$$\alpha=\frac{R(t)-R(T)}{N(T)-1}$$

其中：

$R(t)$表示叶子节点的误差代价，$R(t)=r(t)\cdot p(t)$，$r(t)$为节点的错误率，$p(t)$为节点数据量的占比；

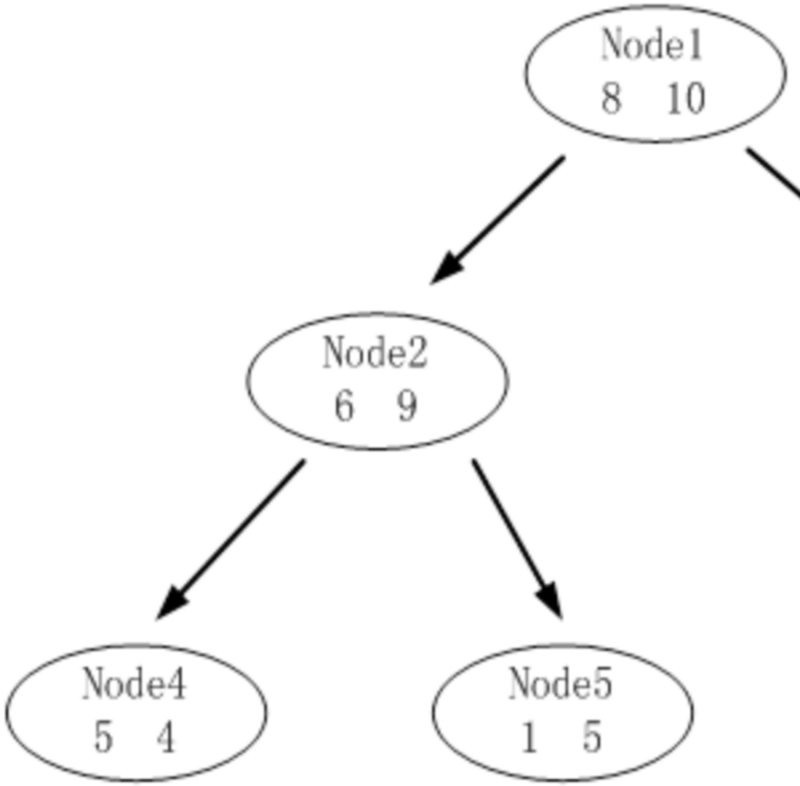
$R(T)$表示子树的误差代价，$R(T)=\sum_i^m r_i(t)\cdot p_i(t)$，$r_i(t)$为子节点i的错误率，$p_i(t)$表示节点i的数据节点占比；

$N(T)$表示子树节点个数。

开课吧
kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

- 总样本为40



$$R(t)=\frac{8}{18}\cdot\frac{18}{40}=\frac{1}{5}$$

$$R(T)=\sum_{i}^{m} r_i(t)\cdot p_i(t)=\frac{1}{3}\cdot\frac{3}{40}+\frac{4}{9}\cdot\frac{9}{40}+\frac{1}{6}\cdot\frac{6}{40}=\frac{6}{40}$$

$$\alpha=\frac{R(t)-R(T)}{N(T)-1}=\frac{\frac{1}{5}-\frac{6}{40}}{3-1}=\frac{1}{40}$$

采用交叉验证方式确定剪枝

开课吧
kaikeba

机器学习中的熵

- 香农熵：
  - 不确定性的度量

$$H(P) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$$

H(p) = -(50%*log(50%)+50%*log(50%))
= 1

H(p) = -(75%*log(75%)+25%*log(25%))
= 0.81

开课吧 kaikeba

机器学习中的熵

- 交叉熵：
  - 考虑两个分布，真实分布q 和 预测分布p

- X0=[0.25,0.25,0.25,0.25]
- X1=[0.125,0.125,0.25,0.5]
- X2=[1/16,1/16,1/8,3/4]

X2和x1 谁更能准确的表示x0

- H(X0,X0)=2
- H(X0,X1)=2.25
- H(X0,X2)=2.85

$$H_p(q) = \sum_x q(x) \log_2 \left(\frac{1}{p(x)} \right)$$

开课吧 kaikeba

机器学习中的熵

- LR中交叉熵损失函数：

$$Cost(h_\theta(x), y) = -ylog(h_\theta(x)) - (1-y)log(1-h_\theta(x))$$

类比：y 作为1类别的真实概率
1-y 为0类别的真实概率
h(θ) 预测为1类别的概率
1- h(θ) 预测为0类别的概率

开课吧 kaikeba

机器学习中的熵

- 相对熵：
  - 又称Kullback熵，Kullback-Leible散度,Kl距离等

- X0=[0.25,0.25,0.25,0.25]
- X1=[0.125,0.125,0.25,0.5]
- X2=[1/16,1/16,1/8,3/4]

定义为：

- H(X0,X0)=2
- H(X0,X1)=2.25
- H(X0,X2)=2.85

$$D_q(p) = H_q(p) - H(p) = \sum_x q(x) \log_2 \left(\frac{q(x)}{p(x)} \right)$$

- D(x0||x1)=0.25
- D(x0||x1)=0.85

衡量两个分布的距离

开课吧 kaikeba

机器学习中的熵

- 互信息：
  - I(X,Y)=D(P(X,Y) || P(X)P(Y))
  - 公式：

$$-\sum_{x,y} p(x,y)(log(\frac{p(x)p(y)}{p(x,y)}))$$

互信息指的是两个随机变量之间的关联程度

开课吧
kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

朴素贝叶斯

开课吧
kaikeba

1. 贝叶斯公式

2. 贝叶斯算法原理

3. 高斯朴素贝叶斯

4. 垃圾邮件识别

开课吧 kaikeba

(贝叶斯公式) 设 $A_1, A_2, \cdots, A_n$ 为一个完备事件组,

$P(A_i) > 0$, $i = 1, \cdots, n$, 对任一事件 B, 若 $P(B) > 0$, 有

$$P(A_k \mid B) = \frac{P(A_k B)}{P(B)}$$

$$= \frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)}, \qquad (k = 1, 2, \cdots, n)$$

$$P(A_k \mid B)=\frac{P(A_k B)}{P(B)}=\frac{P(A_k)P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i)P(B \mid A_i)},$$

$\mathrm{P}(\boldsymbol{A_k})$先验概率

$\mathrm{P}(\boldsymbol{A_k}|\mathrm{B})\boldsymbol{A_k}$后验概率

$\mathrm{P}(\boldsymbol{B}|\boldsymbol{A_k})$似然函数

2.贝叶斯算法原理

开课吧
kaikeba

特征1-高	特征2-富	特征3-帅	见面
高	富	帅	见
高	富	锉	见
高	穷	锉	不见
矮	富	锉	不见
矮	穷	帅	见
矮	穷	锉	不见

矮	富	帅	**见？不见？**

开课吧 kaikeba

朴素贝叶斯算法简介

- 在分类（classification）问题中，常常需要把一个事物分到某个类别。一个事物具有很多属性，把它的众多属性看做一个向量，即$x=(x_1,x_2,x_3,\ldots,x_n)$，用x这个向量来代表这个事物。
- 有类别集合 $y=（y_1,y_2,y_3,\ldots.y_n）$
- 分别计算 $p(y_1|x)$ $p(y_2|x)$ $p(y_3|x)\ldots.$ $p(y_n|x)$，如果$p(y_k|x)$

  $=\max\{p(y_1|x)\ p(y_2|x)\ p(y_3|x)\ldots.\ p(y_n|x)\}$, x就属于$y_k$类。

- 如何计算 $p(y_k|x)$

方法：运用贝叶斯公式 $p(y_k|x)=p(x|y_K)*p(y_k)/p(x)$

在之前已介绍$x=(x_1, x_2, x_3, \cdots, x_n)$，根据朴素贝叶斯假设

$x_1, x_2, x_3, \cdots, x_n$是相互独立的

则有 $p(x | y_k) = p(x_1, x_2, x_3, \cdots, x_n|y_k) = p(x_1 | y_k) *p(x_2 | y_k)$

$\ldots\ldots *p(x_n | y_k)$ (1)

开课吧 kaikeba

特征1-高	特征2-富	特征3-帅	见面
高	富	帅	见
高	富	锉	见
高	穷	锉	不见
矮	富	锉	不见
矮	穷	帅	见
矮	穷	锉	不见

矮	富	帅

max(P(见|[矮，富，帅])，P(不见|[矮，富，帅]))

P(见|[矮，富，帅])=P（[矮，富，帅] | 见）P(见)/P(矮，富，帅)
=P（矮|见）*P（富|见）*P（帅|见）*P（见）/(P（矮）*P（富）*P（帅）)
=(1/3*2/3*2/3*1/2)/(1/2*1/2*1/3)
=

P(不见|[矮，富，帅])=P（[矮，富，帅] | 不见）P(不见)/P(矮，富，帅)
=P（矮|不见）*P（富|不见）*P（帅|不见）*P（不见）/(P（矮）*P（富）*P（帅）)
=(2/3*1/3*1/3*1/2)/(1/2*1/2*1/3)
=

开课吧 kaikeba

特征1-高	特征2-富	特征3-帅	见面
高	富	帅	见
高	富	锉	见
高	穷	锉	不见
矮	富	锉	不见
高	穷	帅	见
矮	穷	锉	不见

矮	富	帅

max(P(见|[矮， 富， 帅])， P(不见|[矮， 富， 帅]))

P(见|[矮， 富， 帅])=P（[矮， 富， 帅] | 见）P(见)/P(矮， 富， 帅)
=P（矮|见）*P（富|见）*P（帅|见）*P（见）/(P（矮）*P（富）*P（帅）)
=(0/3*2/3*2/3*3/6)/P0
=0

P(不见|[矮， 富， 帅])=P（[矮， 富， 帅] | 不见）P(不见)/P(矮， 富， 帅)
=P（矮|不见）*P（富|不见）*P（帅|不见）*P（不见）/(P（矮）*P（富）*P（帅）)
=(2/3*1/3*1/3*1/2)/P0
=0.385

开课吧 kaikeba

特征1-高	特征2-富	特征3-帅	见面
高	富	帅	见
高	富	锉	见
高	穷	锉	不见
矮	富	锉	不见
矮	穷	帅	见
矮	穷	锉	不见

矮	富	帅

$$P_{new} = \frac{N_x + 1}{N_M + k}$$

max(P(见|[矮，富，帅])， P(不见|[矮，富，帅]))

P(见|[矮，富，帅])=P（[矮，富，帅] | 见）P(见)/P0
=P（矮|见）*P（富|见）*P（帅|见）*P（见）
=(1/5*3/5*3/5*4/8)
=0.04

P(不见|[矮，富，帅])=P（[矮，富，帅] | 不见）P(不见)/P0
=P（矮|不见）*P（富|不见）*P（帅|不见）*P（不见）
=(3/5*2/5*2/5*4/8)
=0.05

开课吧 kaikeba

3.高斯朴素贝叶斯

开课吧
kaikeba

特征1-高	特征2-富	特征3-帅	见面
180	50K	90	见
190	40K	30	见
175	5K	80	不见
160	20K	40	不见
170	6K	70	见
165	7K	59	不见

特征1-高	特征2-富	特征3-帅	见面
高	富	帅	见
高	富	锉	见
高	穷	帅	不见
矮	富	锉	不见
高	穷	帅	见
矮	穷	锉	不见

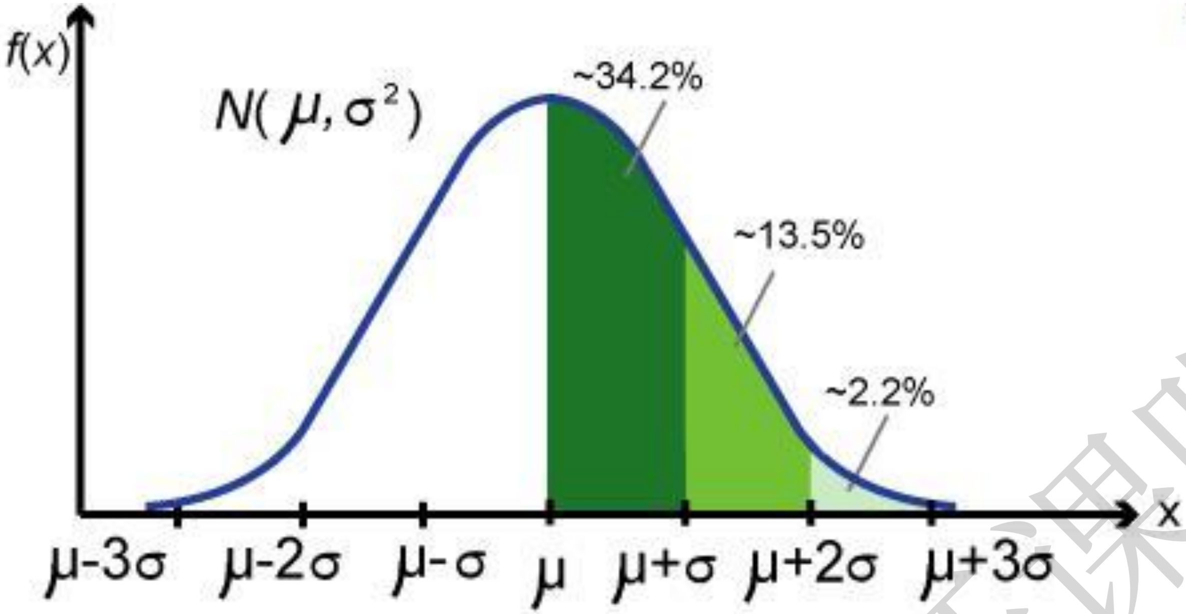
分箱法:将连续数据分段后成为离散数据的一种预处理办法。

>=170:高 <170:矮

>=20K:富 <20K:穷

>=60:帅 <60:锉

开课吧 kaikeba



$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

均值为μ:

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}$$

方差为σ:

$$s^2=\frac{\sum_{i=1}^{n}(x-u)^2}{n-1}$$

开课吧 kaikeba

特征1-高	特征2-富	特征3-帅	见面
180	50K	90	见
190	40K	30	见
175	5K	80	不见
160	20K	40	不见
170	6K	70	见
165	7K	59	不见

170	30K	70

以特征1为例：
见面的均值为 180，标准差为：8.24，所以P（170|见）概率为

$$P(Yi|Xi) = \frac{1}{\sqrt{2\pi\delta}} e^{-\frac{(x-\mu)^2}{2\delta^2}} = \frac{1}{\sqrt{2\pi\delta}} e^{-\frac{(x-\mu)^2}{2\delta^2}} =0.028$$

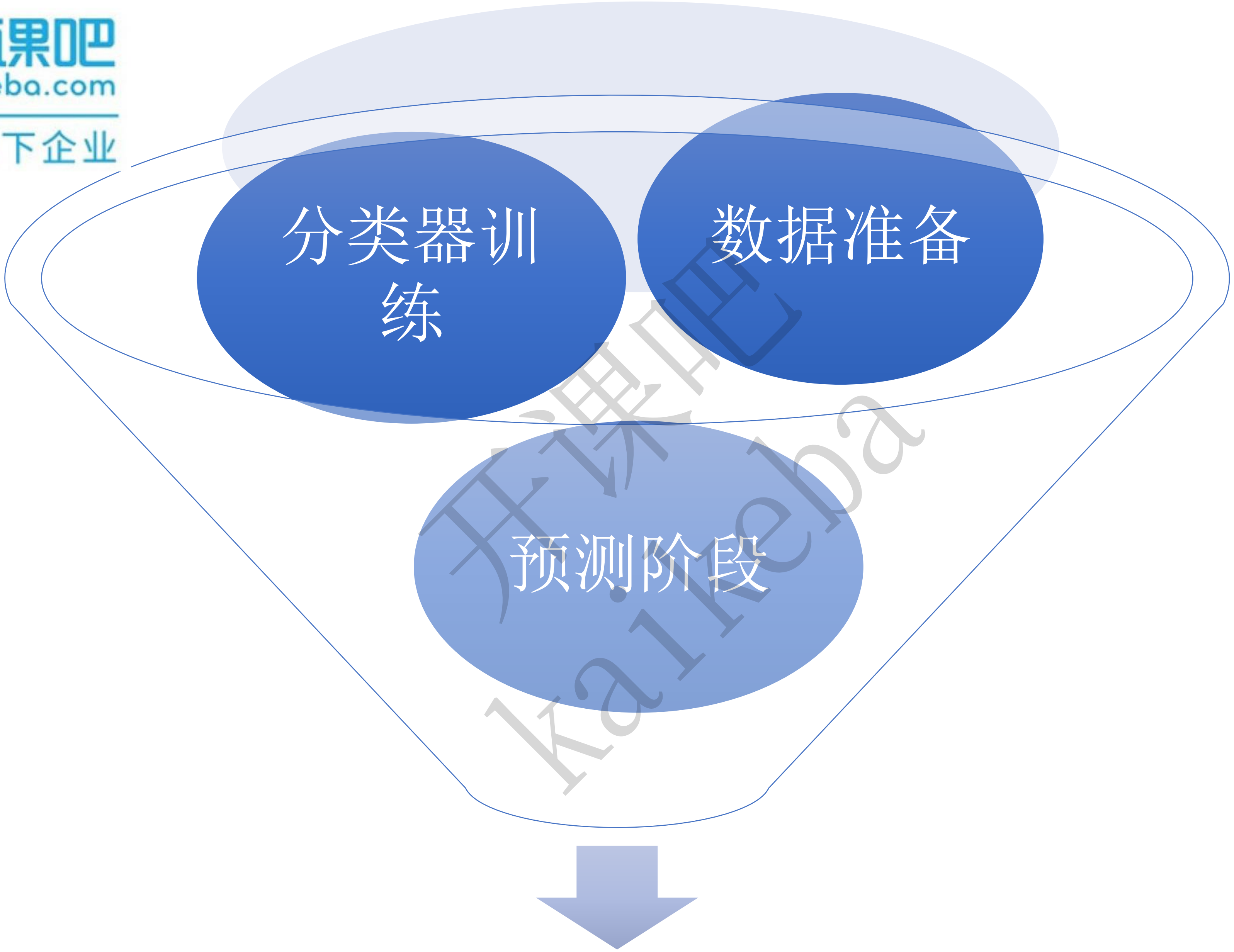
同样我们可以计算出其余部分然后用贝叶斯定理计算即可。

开课吧 kaikeba

4.垃圾邮件识别

开课吧 kaikeba

开课吧
kaikeba.com
慧科集团旗下企业



开课吧
kaikeba

- 如何将贝叶斯分类器应用到垃圾邮件识别中来。

在垃圾邮件识别过程中，假设我们有一个文档d{t1,t2,t3...}和一个固定的标签集合C={c1,c2}

比如：

需要|为|企业|开具|发票，请|联系|我。（垃圾邮件）

附件|是|我的|报表，如果|可行，请|批示。（非垃圾邮件）

开课吧 kaikeba

计算条件概率

$P(X1 = t1, X2 = t2, \ldots, Xn = tn \mid yi)$ 和 $P(C = yi)$

对比后验概率

$P(C = yi \mid X1 = t1, X2 = t2, \ldots, Xn = tn), (i = 1, 2)$

哪个标签大就将其归为某个类别。

开课吧 kaikeba

样本	Email内容	分词	标签
N01	有偿开发票	[’有偿’，’开发票’]	垃圾邮件
N02	夜店酒吧	[’夜店’，’酒吧’]	垃圾邮件
N03	收到请回复	[’收到’，’请’，’回复’]	非垃圾邮件
N04	请填写基本信息	[’请’，’填写’，’基本’，’信息’]	非垃圾邮件
N05	下午4点开会	[’下午’，’4’，’点’，’开会’]	非垃圾邮件

样本出现14个集合为:
['下午', '酒吧', '开会', '开发票', '有偿', '回复', '4', '信息', '请', '基本', '填写', '收到', '夜店', '点']

样本	Email内容	分词	向量	标签
N01	有偿开发票	['有偿', '开发票']	[0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]	1
N02	夜店酒吧	['夜店', '酒吧']	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0]	1
N03	收到请回复	['收到', '请', '回复']	[0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0]	0
N04	请填写基本信息	['请', '填写', '基本', '信息']	[0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0]	0
N05	下午4点开会	['下午', '4', '点', '开会']	[1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1]	0

开课吧
kaikeba

将文本转化为了计算机可以处理的0，1变量，这种方法叫做词袋模型

标签	概率
1	3/7
0	4/7

计算标签概率

注意：已经做了拉普拉斯平滑

计算条件概率

(0+1)/(2+2)

2:垃圾邮件条数

2:标签的类别

	垃圾邮件中出现	P(x|1)	非垃圾邮件中出现	P（x|0）
下午	0	0.25	1	0.4
酒吧	1	0.5	0	0.2
开会	0	0.25	1	0.4
开发票	1	0.5	0	0.2
有偿	1	0.5	0	0.2
回复	0	0.25	1	0.4
4	0	0.25	1	0.4
信息	0	0.25	1	0.4
请	0	0.25	1	0.4
基本	0	0.25	1	0.4
填写	0	0.25	1	0.4
收到	0	0.25	1	0.4
夜店	1	0.5	0	0.2
点	0	0.25	1	0.4

开课吧 kaikeba

- 请填写开发票信息 -> 请|填写|开发票|信息

  ->[0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0]

- P(X|1)=log(（1-p(x1|1)）*（1-p(x2|1)）*... p(x4|1) ...)

- P(X|0)= log(（1-p(x1|0)）*（1-p(x2|0)）*... p(x4|0) ...)

- 对比P(1|X)和P(0|X)

开课吧 kaikeba

开课吧
kaikeba.com
慧科集团旗下企业

多项式分布

在多项式模型中， 设某文档d=(t1,t2,…,tk)，tk是该文档中出现过的单词，允许重复，则

先验概率P(c)= 类c下单词总数/整个训练样本的单词总数
类条件概率P(tk|c)=(类c下单词tk在各个文档中出现过的次数之和+1)/(类c下单词总数+|V|)

V是训练样本的单词表（即抽取单词，单词出现多次，只算一个），|V|则表示训练样本包含多少种单词。 P(tk|c)可以看作是单词tk在证明d属于类c上提供了多大的证据，而P(c)则可以认为是类别c在整体上占多大比例(有多大可能性)。

开课吧 kaikeba

句子			类别
开发票	有偿	开发票	垃圾邮件
开发票	开发票	提成	垃圾邮件
需要	开发票		垃圾邮件
开发票	提供	纳税	非垃圾邮件

有偿	开发票	纳税	提供	需要	提成
1	2	0	0	0	0
0	2	0	0	0	1
0	1	0	0	1	0
0	1	1	1	0	0

P(1)=8/11

P(0)=3/11

$P(w0|1) = (1+1) \;/\; (8+6)$

$P(w1|1) = (5+1) \;/\; (8+6)$

$P(w2|1) = (0+1) \;/\; (8+6)$

$P(w3|1) = (0+1) \;/\; (8+6)$

$P(w4|1) = (1+1) \;/\; (8+6)$

$P(w5|1) = (1+1) \;/\; (8+6)$

$P(w0|0) = (0+1) \;/\; (3+6)$

$P(w1|0) = (1+1) \;/\; (3+6)$

$P(w2|0) = (1+1) \;/\; (3+6)$

$P(w2|0) = (1+1) \;/\; (3+6)$

$P(w4|0) = (0+1) \;/\; (3+6)$

$P(w5|0) = (0+1) \;/\; (3+6)$

开课吧 kaikeba

数据预处理

条件概率计算

高斯朴素贝叶斯：（Gaussian Naive Bayes）

伯努利朴素贝叶斯（ Bernoulli Naive Bayes）

多项式朴素贝叶斯（Multinomial Naive Bayes）

后验概率计算

开课吧 kaikeba