



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

1

新媒体大数据分析

New Media Big Data Analysis

第二章 数据分析

黄振亚，朱孟潇，张凯

课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/NM2025.html>

助教：齐畅，朱家骏

bigdata_2025@163.com

10/21/2025



回顾：数据分析基础

2

- 数据采集 Data Collection
- 数据预处理 Data Preprocessing
- 特征工程 Feature Engineering



数据预处理

3

- 大数据环境下的数据特征
- 为什么需要进行预处理
- 预处理的基本方法
 - 数据清理
 - 数据集成
 - 数据变换



数据预处理：数据变换

4

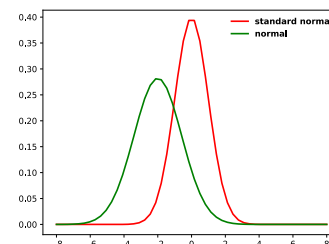
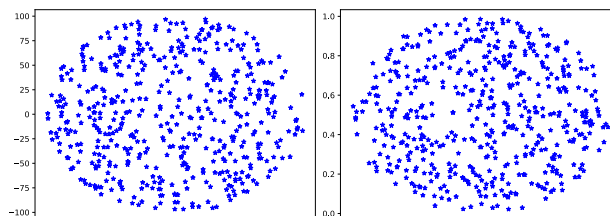
□ 数据变换的目的是将数据转换成适合分析建模的形式

□ 前提条件：尽量不改变原始数据的规律

□ 数据规范化

■ 最小-最大规范化

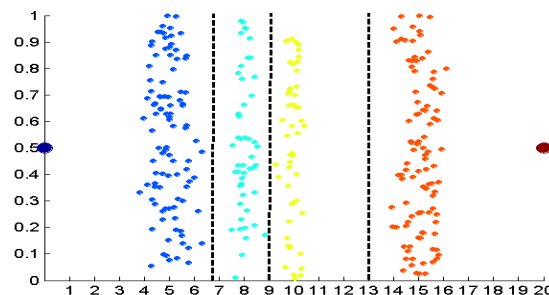
■ z-score规范化



□ 数据离散化

■ 非监督离散化

■ 监督离散化



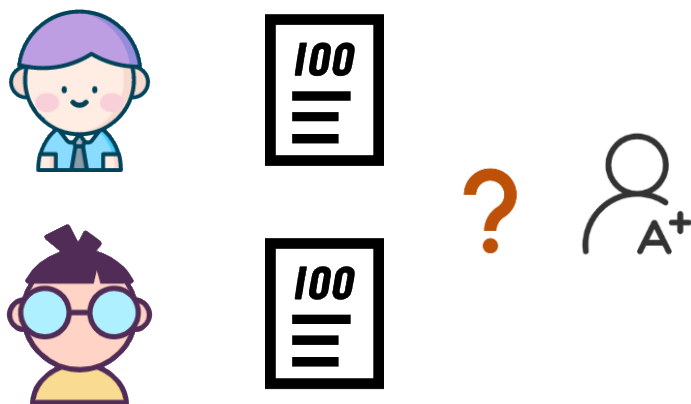


数据预处理：数据变换

5

□ 数据规范化

- 目的：将不同数据（属性）按一定规则进行缩放，使它们具有可比性
- 例如，我们需要考察学生A和学生B的某门课程成绩。A的考试满分是100分（及格60分），B的考试满分是150分（及格90分）。显然，A和B的100分代表着完全不同的含义。



如何用一个同等的标准来比较A与B的成绩数据呢？



数据变换-规范化

6

□ 最小-最大规范化

□ 对原始数据进行线性变换。把数据A的观察值 v 从原始的区间 $[\min_A, \max_A]$ 映射到新区间 $[\text{new_min}_A, \text{new_max}_A]$

■ 0-1规范化又称为归一化

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

□ 数理依据:

$$\frac{v' - \text{new_min}_A}{\text{new_max}_A - \text{new_min}_A} = \frac{v - \min_A}{\max_A - \min_A}$$



数据变换-规范化

7

□ 最小-最大规范化

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- 例：假设某属性规范化前的取值区间为 $[-100, 100]$ ，规范化后的取值区间为 $[0, 1]$ ，采用最小-最大规范化 66，得

$$v' = \frac{66 - (-100)}{100 - (-100)} (1 - 0) + 0 = 0.83$$



数据变换-规范化

8

假设A的课程成绩为70分（0-100分），B的课程成绩为110分（0-150分），采用最小-最大规范化来比较A和B的成绩



70
三三三
✓
00

取值区间为[0,100]，规格后的取值空间为[0,1]，采用最小-最大规范70后为0.7



110
三三三
✓
00

取值区间为[0,150]，规格后的取值空间为[0,1]，采用最小-最大规范110后为0.73



用最小-最大规范化后得出B的成绩更好



数据变换-规范化

9

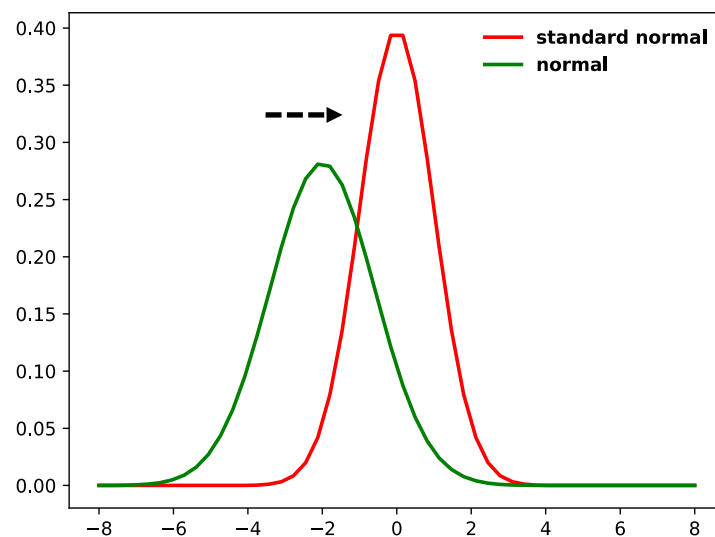
□ z-score规范化

- 最大最小值未知，或者离群点影响较大时，假设数据服从正态分布
 - 某一原始数据 (v) 与原始均值的差再除以标准差，可以衡量某数据在分布中的相对位置

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

- 例：假设某属性的平均值、标准差分别为80、25，用z-score规范化 66

$$v' = \frac{66 - 80}{25} = -0.56$$





数据变换-规范化

10

□ z-score规范化

- 例：假设学生的成绩分布符合正态分布，某素质课考试的平均分为73分，标准差为7分，A得78分；实践课考试的平均分为80分，标准差为6.5分，A得83分。那么A的哪一门考试成绩比较好？



素质课 **78**
三三三三

平均分为73分，标准差为7分，采用z-score规范78后为 $(78-73)/7=0.71$



实践课 **83**
三三三三

平均分为80分，标准差为6.5分，采用z-score规范83后为 $(83-80)/6.5=0.46$

采用z-score规范化得出A的素质课成绩要优于实践课成绩



数据变换-规范化

11

□ 小结

	优点	缺点	适用场景
最小-最大规范化	保留了原始数据中存在的关系，是消除量纲和数据取值范围影响的最简单方法	对最大最小值敏感，新数据加入时，可能改变最大最小值，需重新计算	适用于原始数据不存在很大/很小的一部分数据的时候
z-score 规范化	算法简单方便，结果方便比较，应用于数值型的数据，且不受数据量级的影响	总体平均值和方差不一定可知，在一定程度上要求数据分布，结果没有具体意义，只用于比较	适用于最大最小值未知，或者离群点影响较大的时候



数据预处理：数据变换

12

□ 数据离散化

- 连续数据过于细致，数据之间的关系难以分析
- 划分为离散化的区间，发现数据之间的关联，便于算法处理
 - 同学们成绩：100分制分数使用五分制离散化表示
 - A（大于等于85分），B，C，D，F（小于60分）
 - 人的年龄：离散化为不同的年龄段（引源自世卫组织）
 - 未成年人：0至17岁；
 - 青年人：18岁至45岁；
 - 中年人：46岁至69岁；
 - 老年人：大于70岁。
 - 一年365天：离散化表示为12个月份或四个季节

A+	A	A-
B+	B	B-
C+	C	C-
D+	D	D-
F		





数据变换-离散化

13

□ 数据离散化

- 连续数据过于细致，数据之间的关系难以分析，将其分段为离散化的区间，发现数据之间的关联，便于算法处理
- 非监督离散化（无类别信息）
- 有监督离散化（有类别信息）



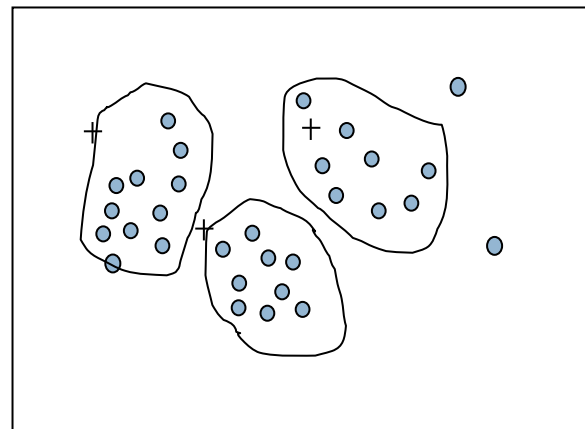
数据变换-离散化

14

□ 非监督离散化 (后续介绍: 聚类算法)

□ 聚类: 监测并且去除噪声数据

- 将类似的数据聚成簇
- 每个簇计算一个值用以将该簇的数据离散化





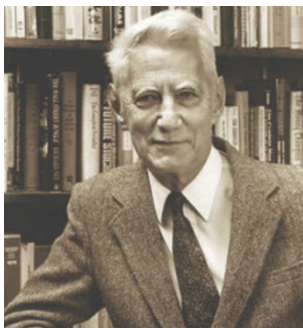
数据变换-离散化

15

□ 有监督离散化—基于熵的离散化

■ 熵用来度量系统的**不确定程度**

- 熵是由 克劳德·艾尔伍德·香农 将热力学的熵，引入到信息论，因此它又被称为香农熵



香农提出了信息熵的概念，为**信息论**和**数字通信**奠定了基础，被誉为“**信息论之父**”



数据变换-离散化

16

□ 信息熵：度量系统的不确定程度

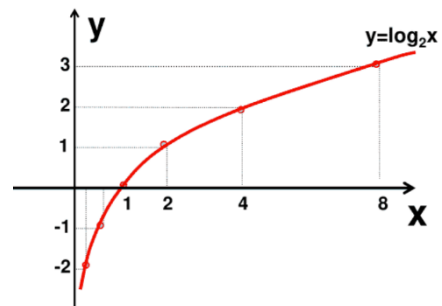
□ 信息量

- 定义一个事件 x 的概率分布为 $P(x)$
- 则事件 x 的自信息量是 $-\log P(x)$, 取值范围: $[0, +\infty]$

□ 信息熵

- 平均而言，发生一个事件得到的自信息量大小
- 即：熵可以表示为自信息量的期望

$$H = - \sum P(x) \log P(x)$$



x	0	1
$P(x)$	0.4	0.6

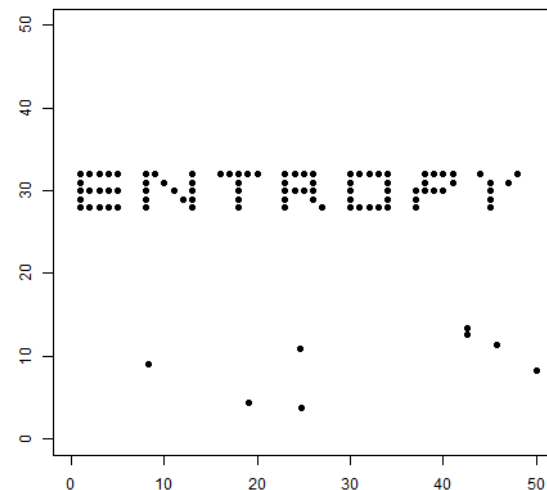
$$\begin{aligned} H(P) &= -P(X=0)\log_2 P(X=0) - P(X=1)\log_2 P(X=1) \\ &= -0.4 * \log_2(0.4) - 0.6 * \log_2(0.6) \\ &\approx 0.97 \end{aligned}$$



数据变换-离散化

17

- 熵与数据离散化有什么关系？——**不确定程度**
 - 数据点单词（ENTROPY）**完整**的时候，容易理解表达的意思，**确定程度较高**，对应的**信息熵也较小**。
 - 数据点被完全打乱的时候，难以理解其意思，造成**不确定性**也就多了，对应的**信息熵也变大了**。
 - 目标：对数据进行离散化后，每个区间的数据的确定性（也称“纯度”）更高因此用熵来对数据进行离散化。

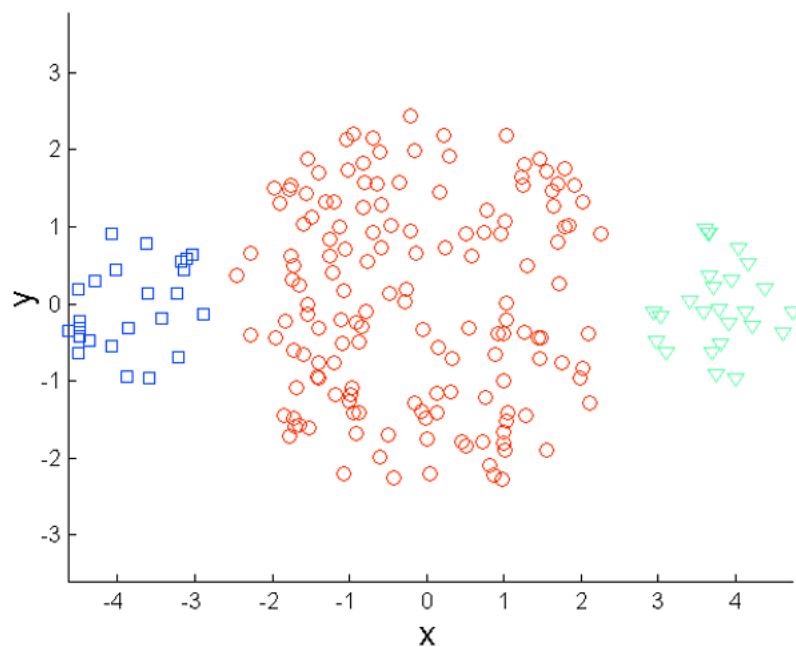




数据变换-离散化

18

- 基于熵的离散化
 - 在x轴上对数据划分



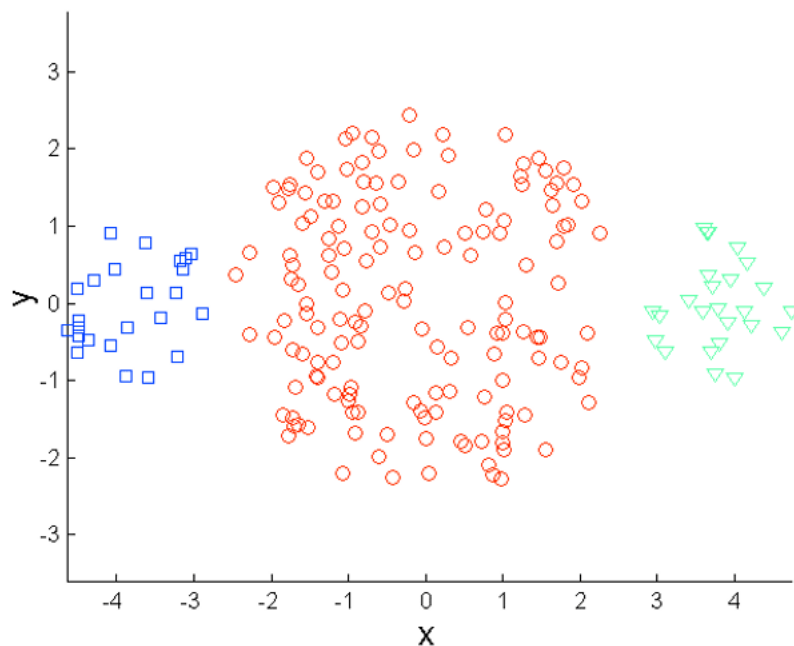
原始数据



数据变换-离散化

19

- 基于熵的离散化
 - 在x轴上对数据划分



原始数据



数据变换-离散化

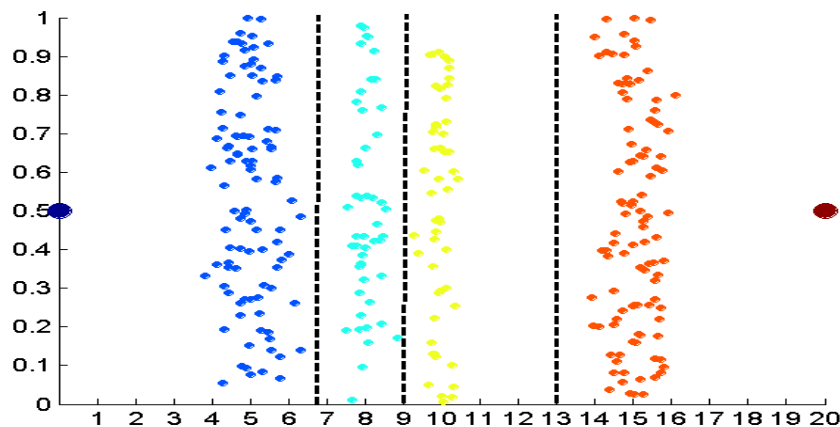
20

□ 熵—计算不确定性以及不纯性

□ 假设数据已经离散，计算离散后的某个区间 t 中的熵：

$$Entropy(t) = -\sum_j p(j | t) \log p(j | t)$$

■ 其中， $p(j | t)$ 表示 第 j 类在区间 t 中的概率；一般对数 $\log$ 以2为底





数据变换-离散化

21

□ 计算 单个区间 的 Entropy

$$Entropy(t) = -\sum_j p(j | t) \log_2 p(j | t)$$

C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$

$$Entropy = -0 \log_2 0 - 1 \log_2 1 = -0 - 0 = 0$$

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$

$$Entropy = -(1/6) \log_2 (1/6) - (5/6) \log_2 (5/6) = 0.65$$

C1	2
C2	4

$$P(C1) = 2/6 \quad P(C2) = 4/6$$

$$Entropy = -(2/6) \log_2 (2/6) - (4/6) \log_2 (4/6) = 0.92$$

- 练习:**
- (1) 假设区里t里面C1和C2的样本数各为3, Entropy是多少? 1
 - (2) 假设区间t里面有4个类, 且样本数一样, Entropy是多少? 2
 - (3) 假设区间t里面有C个类, 且样本数一样, Entropy是多少? $\log C$



数据变换-离散化

22

- 熵—计算不确定性以及不纯性
 - 假设数据已经离散，计算离散后的某个区间 t 中的熵：

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t)$$

- 其中， $p(j|t)$ 表示 第 j 类在区间 t 中的概率；一般对数 $\log$ 以2为底
- 区间里面不同类别的样本均匀分布时，熵值最大（最不确定、最不纯），熵值为： $\log C$
- 区间里面只有一类样本时，熵值最小（最确定、最纯）
- 熵的取值范围： $[0, \log C]$

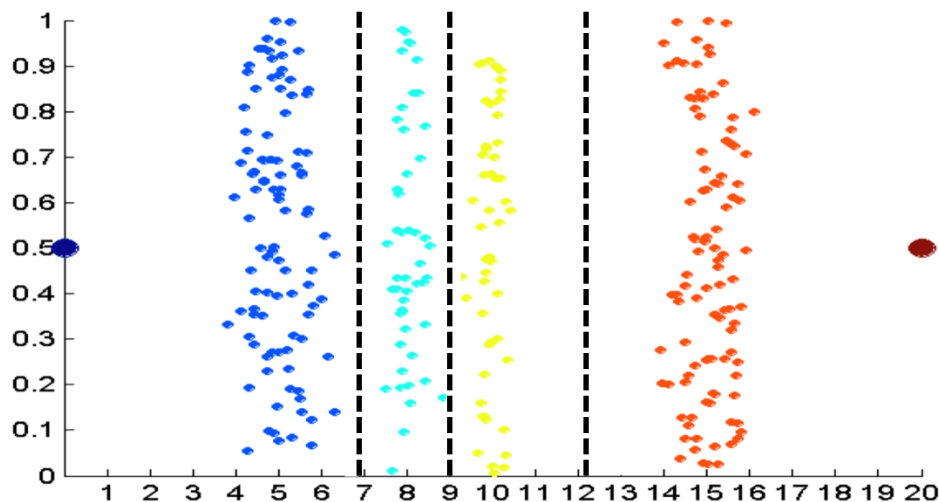
结
论



数据变换-离散化

23

- 根据Entropy进行二分离散化
 - 先找到一个分隔点（属性值），把所有数据分到两个区间
 - 分别对两个子区间的数据进行二分
 - 重复以上步骤





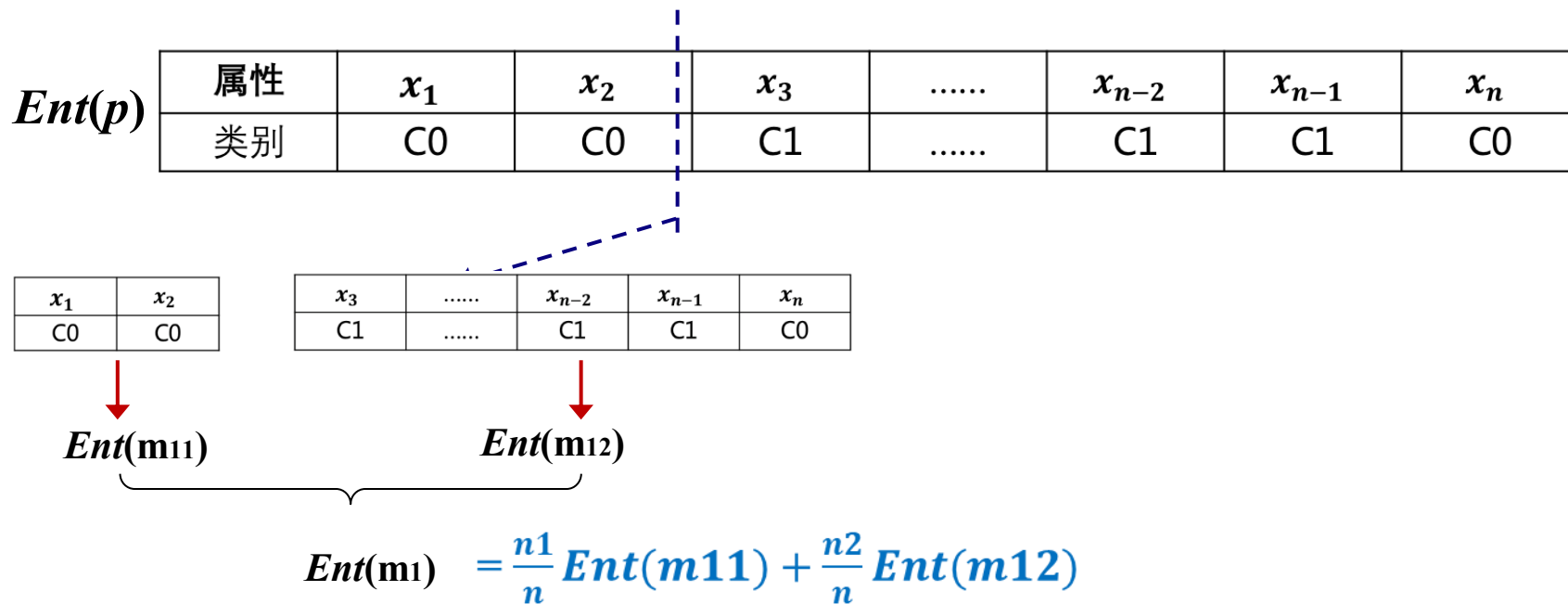
数据变换-离散化

24

□ 如何确定分隔点？——计算分隔后的信息增益

□ 信息增益 (Information Gain)

■ 表示在某个条件下，信息不确定性减少的程度



$$Gain1 = Ent(p) - Ent(m_1) > 0$$



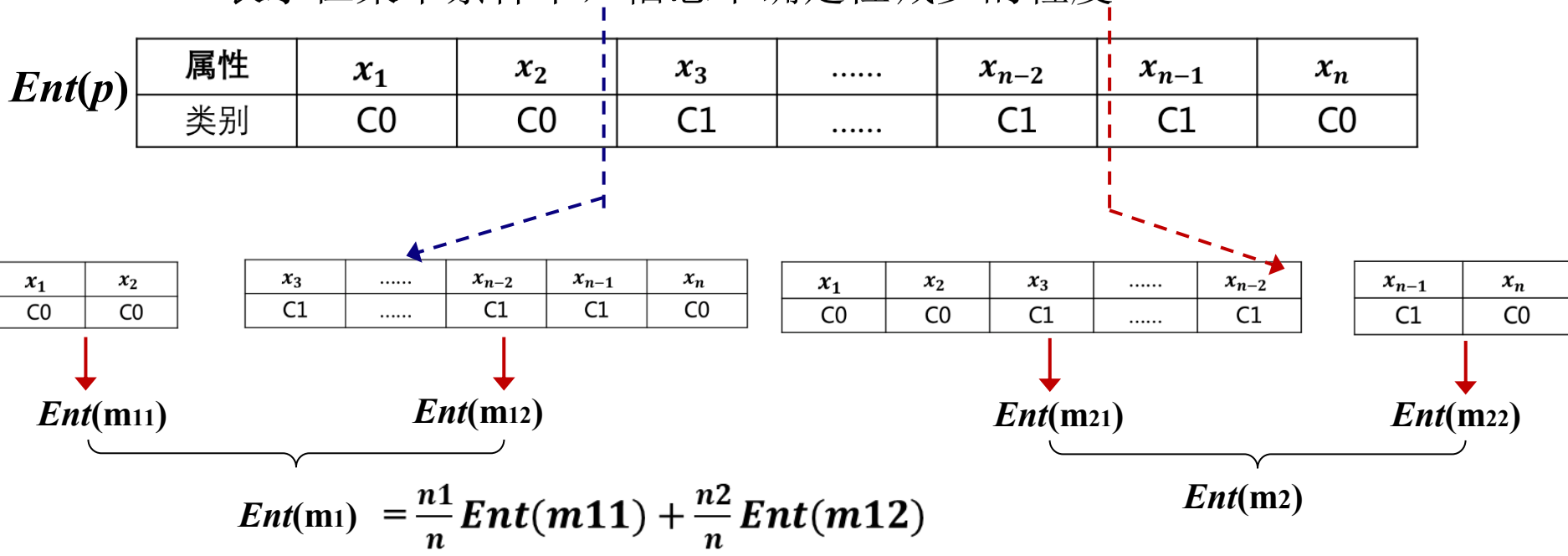
数据变换-离散化

25

□ 如何确定分隔点？——计算分隔后的信息增益

□ 信息增益 (Information Gain)

■ 表示在某个条件下，信息不确定性减少的程度





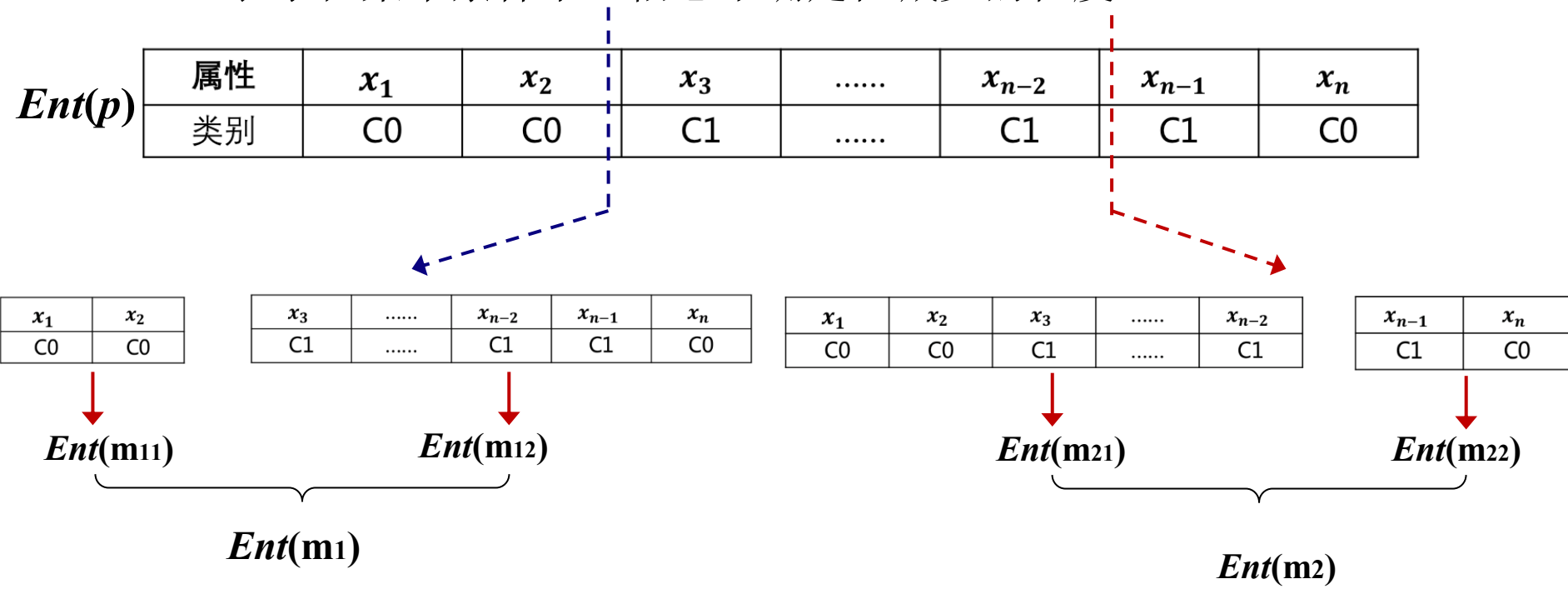
数据变换-离散化

26

□ 如何确定分隔点？——计算分隔后的信息增益

□ 信息增益 (Information Gain)

■ 表示在某个条件下，信息不确定性减少的程度



$$Gain1 = Ent(p) - Ent(m_1)$$

Vs

$$Gain2 = Ent(p) - Ent(m_2)$$



数据变换-离散化

27

□ 如何确定分隔点？——计算分隔后的信息增益

□ 信息增益 (Information Gain) :

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^k \frac{n_i}{n} Entropy(i) \right)$$

- 信息增益：表示在某个条件下，信息不确定性减少的程度。
- 父节点 P 被分隔为 K 个区间
- n 表示总记录数，n_i表示区间 i 中的记录数

□ 确定分隔点 j :

- 选择信息增益最大的分隔点，即

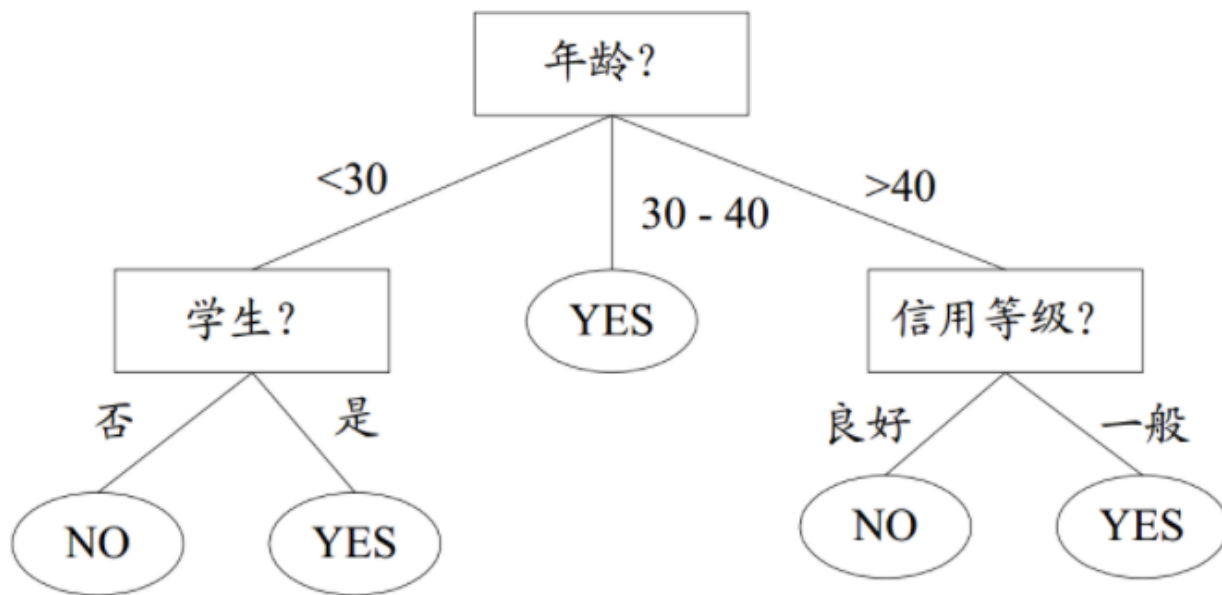
$$j = \max(GAIN_{split})$$



数据变换-离散化

28

- 十大经典机器学习算法
 - 决策树 (第四章：数据挖掘)





数据变换-离散化

29

熵 (*Entropy*) 的应用举例

- 使用熵进行旅游季节 (Travel Season) 的划分
- 假设不同的景点适合不同的季节进行旅游



- 根据景点的类别分布, 计算区间 (季节) 中的熵:

$$WAE(i; S^P) = \frac{|S_1^P(i)|}{|S^P|} Ent(S_1^P(i)) + \frac{|S_2^P(i)|}{|S^P|} Ent(S_2^P(i))$$



课后学习

30

□ 前沿文献调研：“熵在数据科学中的应用”

□ 推荐1：基于技术分布的熵值预测公司发展前景

- 技术的发展一般处于5个阶段(萌芽期、过热期、低谷期、复苏期和成熟期)，如果公司的技术发展在以上阶段分布越均衡，可能它的发展前景就越好

- Bo Jin, Yong Ge, Hengshu Zhu, Li Guo, Hui Xiong and Chao Zhang. Technology Prospecting for High Tech Companies through Patent Mining ICDM'2014

□ 推荐2：基于交叉熵的机器学习目标函数设计

- 信息熵、交叉熵和相对熵：<https://charlesliuyx.github.io/2017/09/11/>



参考资料

31

- Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. Personalized Travel Package Recommendation. ICDM'2011
- Bo Jin, Yong Ge, Hengshu Zhu, Li Guo, Hui Xiong and Chao Zhang. Technology Prospecting for High Tech Companies through Patent Mining ICDM'2014
- 数据规范化的几种方法: <https://www.jianshu.com/p/55aee18b3fbc>
- Z-score (Z值) 的意义: http://blog.sina.com.cn/s/blog_72208a6a0101cdt1.html
- 信息熵是什么: <https://www.zhihu.com/question/22178202>
- 交叉熵损失函数的优点: https://blog.csdn.net/qq_41853758/article/details/82826820
- 信息熵、交叉熵和相对熵: <https://charlesliuyx.github.io/2017/09/11/>
- 常见的三种数据规范化方法及其python实现: <https://joshuaqyh.github.io/2019/02/24/>
- 一种基于信息熵的离散化方法 (MDLP) python实现: <https://zhuanlan.zhihu.com/p/74839156>

