数据科学导论

Introduction to Data Science

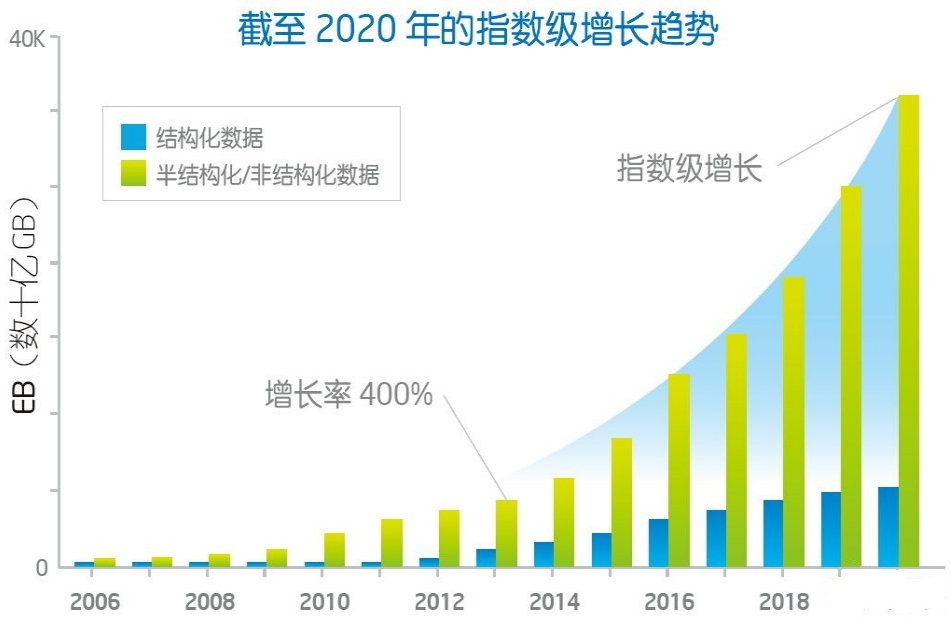
第二章 数据分析基础

黄振亚，陈恩红，刘淇
Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html

回顾：数据分析基础

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



回顾：数据预处理

- 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
  - 数据规约

回顾：数据预处理：数据清理

- 数据清理的目标
  - 解决数据质量问题
  - 让数据更适合分析、建模
- 数据清理基本任务
  - 处理缺失值
  - 清洗噪声数据
  - 纠正不一致数据
  - 根据需求进行清理
  - ……

ID	住址	学历	单位	专业	收入
01	A区	本科	A	CS	C
02	B区	本科	B	EE	C
03	A区	本科	A	CS	C
04	A区	硕士	C	CS	B
05	A区	博士	A	DS	A
…	…	…	…	…	…

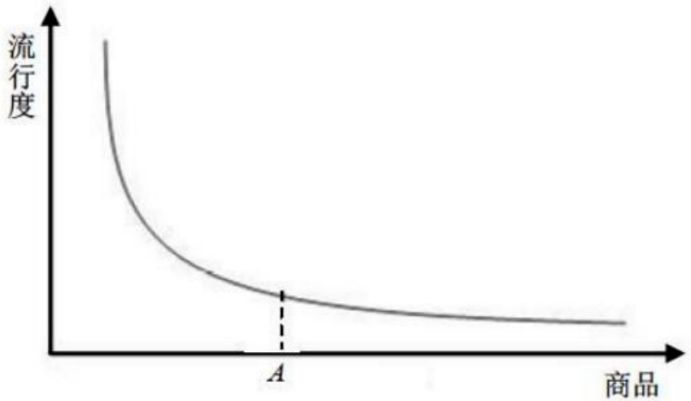
回顾：数据清理-处理缺失值

- 处理缺失数据的方法：首先确认缺失数据的影响
  - 数据删除（可能丢失信息，或改变分布）
    - 删除数据
    - 删除属性
    - 改变权重
  - 数据填充
    - 特殊值填充
      - 空值填充，不同于任何属性值。例，NLP词表补0，DL补mask
      - 样本/属性的均值、中位数、众数填充
    - 使用最可能的数据填充
      - 热卡填充（就近补齐）
      - K最近距离法（KNN）
      - 利用回归等估计方法
      - 期望值最大化方法（EM算法）

模型预测：建立模型预测缺失值

回顾：数据清理-根据需求清理数据

- 在特定的应用任务中，根据目标不同，需要特殊的数据清理方法
  - 推荐系统
    - 通用推荐问题：例，删除记录不足5条的用户
    - 冷启动问题：例，保留没有记录的用户—新用户
  - 教育大数据
  - 社交网络
  - POI任务：Point of interest

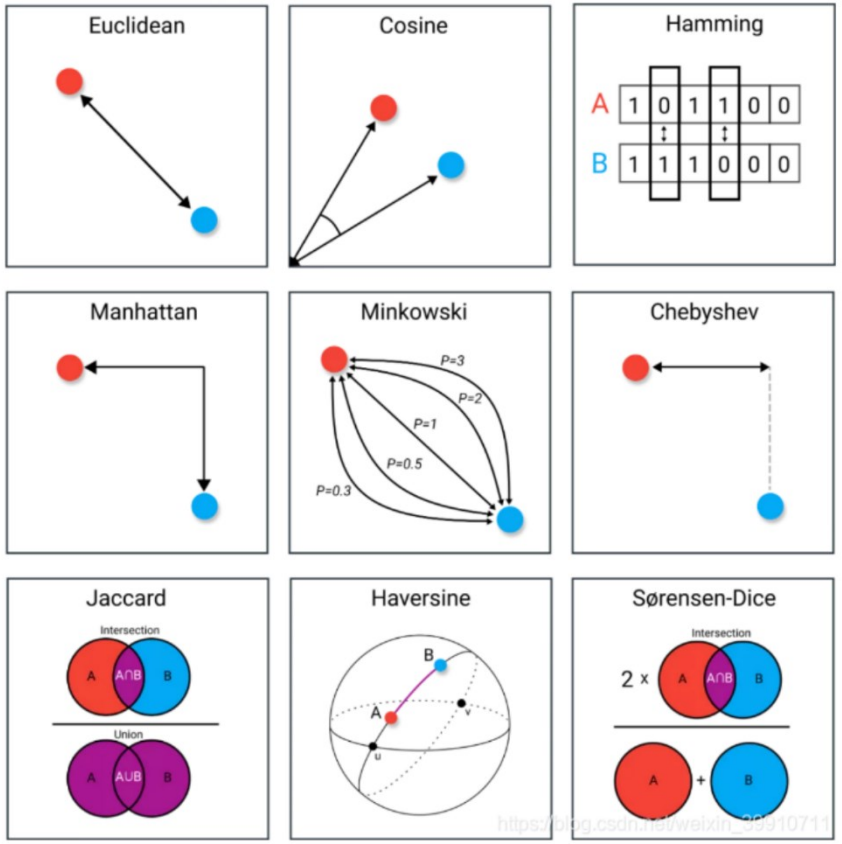


回顾：数据预处理：数据集成

- 数据集成
  - 将多个数据源的数据整合到一个**一致的**数据存储中
  - 集成数据（库）时，经常出现冗余数据
    - 冗余的属性
    - 冗余的样本
  - 冗余数据带来的问题: 浪费存储、重复计算
  - 数据集成的目标
    - 获得更多的数据
    - 获得更完整的数据
    - 获得更全面的数据画像，如用户画像

回顾：数据预处理：数据集成

- 检测冗余样本
  - 思想：数据样本之间的相关性，数据融合、去除冗余
  - 方法：距离度量
    - 欧几里得距离
    - 曼哈顿距离
    - 汉明距离
    - 明氏距离
    - ……
  - 方法：相似度计算
    - 余弦相似度
    - Jaccard相似度
    - ……



数据预处理：数据集成

数据的相关性分析

- **无序数据：每个数据样本的不同维度是没有顺序关系的**
  - 余弦相似度、相关度、欧几里得距离、Jaccard
- **有序数据：对应的不同维度(如特征)是有顺序(rank)要求的**
  - 在信息检索中，如何判断不同检索方法返回的页面序列的优劣
  - 在推荐系统中，如何判断不同推荐序列的好坏
    - **Spearman Rank(斯皮尔曼等级)相关系数**
    - **标准化的折损累计增益(NDCG)**
    - 肯德尔相关性系数
      - kendall correlation coefficient
- 课外阅读：PageRank算法

i	相关度
1	3
2	3
3	2
4	0
5	1
6	2

方法返回结果

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

真实结果

数据预处理：数据集成

数据的相关性分析—举例

- 已知：6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。
- 如果我们设计了两个检索算法，返回结果分别是(b)和(c)，请问哪个方法的结果与真实结果更相似？

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

(c)方法2返回结果

数据预处理：数据集成

- **有序数据的距离度量(信息检索、推荐系统等)**
  - **Spearman Rank(斯皮尔曼等级)相关系数**
    - 比较两组变量的相关程度
    - 当关系是非线性时，它是两个变量之间关系评价的更好指标

$$\rho_S = \frac{\sum_{i=1}^{N}(R_i-\bar{R})(S_i-\bar{S})}{[\sum_{i=1}^{N}(R_i-\bar{R})^2 \sum_{i=1}^{N}(S_i-\bar{S})^2]^{\frac{1}{2}}} \Rightarrow \rho_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

    - ρ_S：表示斯皮尔曼相关系数
    - d_i^2：表示每一对样本之间等级的差
    - n：表示样本容量
  - ρ_S的范围：-1 to 1 (正相关:$\rho_S > 0$，负相关:$\rho_S < 0$，不相关:$\rho_S = 0$)

数据预处理：数据集成

- **有序数据的距离度量(信息检索、推荐系统等)**
  - **Spearman Rank(斯皮尔曼等级)相关系数**
  - X = (a, b, c, d, e, f)
  - Y = (c, a, e, d, f, b)

$$d_i = Y_i - X_i$$

- $d_i^2 = (4, 1, 4, 0, 1, 16)$
- $\rho = 1 - \frac{6(26)}{6(36-1)} \approx 1 - 0.743 = 0.257$

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

数据预处理：数据集成

数据的相关性分析—课后思考

- Spearman Rank相关度与Pearson相关度之间的联系与区别？

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad \rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$$\rho_s = \frac{\sum_{i=1}^{N}(R_i-\bar{R})(S_i-\bar{S})}{[\sum_{i=1}^{N}(R_i-\bar{R})^2 \sum_{i=1}^{N}(S_i-\bar{S})^2]^{\frac{1}{2}}} \qquad p_{X,Y} = \frac{\sum_{i=1}^{n}(X_i-\tilde{X})(Y_i-\tilde{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\tilde{X})^2 \sum_{i=1}^{n}(Y_i-\tilde{Y})^2}}$$

斯皮尔曼相关系数被定义成等级数据变量(rank/order variables)之间的皮尔逊相关系数

数据预处理：数据集成

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

数据的相关性分析——练习题2 （计算Spearman）

- 已知6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。如果我们设计了两个检索算法，返回结果分别是(b)和(c)，
- 请问：哪个方法的结果与真实结果更相似？

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

(c)方法2返回结果

只考虑了每个位置的数据与真实数据的顺序差异，但是没有考虑到不同位置(position)的重要性差异

数据预处理：数据集成

- **有序数据的距离度量(信息检索、推荐系统等)**
  - **NDCG(Normalized Discounted cumulative gain)**
    - **CG(累计增益)**：只考虑到了相关性的关联程度，没有考虑每个推荐结果处于**不同位置**对整个推荐效果的影响

      rel_i表示处于位置 ii 的推荐结果的相关性
    - **DCG(折损累计增益)**：就是在每一个CG的结果上处以一个折损值，目的就是为了让排名越靠前的结果越能影响最后的结果

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$$

    - i表示推荐结果的位置， i越大，则推荐结果在推荐列表中排名越靠后推荐效果越差，DCG越小

Qi Liu, Yong Ge, Enhong Chen, and Hui Xiong. Personalized Travel Package Recommendation. ICDM'2011, **(Best Research Paper Award)**

数据预处理：数据集成

- **有序数据的距离度量(信息检索、推荐系统等)**
  - **NDCG(** Normalized Discounted cumulative gain **)**
    - **NDCG：**由于搜索结果随着检索词的不同，返回的数量不一致，而DCG是一个累加的值，没法针对两个不同的搜索结果进行比较，因此需要标准化处理，这里是除以IDCG:

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

IDCG为理想（ideal）情况下最大的DCG值，指推荐系统为某一用户返回的最好推荐结果列表(或者，真实的数据序列)

数据预处理：数据集成

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

- 例，假设搜索返回的6个物品，其相关性分别是 3、2、3、0、1、2
  - CG@6 = 3+2+3+0+1+2
  - DCG@6 = 7+1. 89+3.5+0+0. 39+1.07 = 13.85
- 假如我们实际召回了8个物品，除了上面的6个，还有两个物品，第7个相关性为3，第8个相关性为0。那么在理想情况下的相关性分数排序应该是
  - 3、3、3、2、2、1、0、0。
- 计算IDCG@6：
  - IDCG = 7+4.42+3.5+1.29+1.16+0.36 = 17.73
- 可以计算NDCG@6：
  - NDCG@6 = 13.85/17.73 = 0.78

$$CG_k = \sum_{i=1}^{k} rel_i \qquad DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$$

i	rel
1	3
2	2
3	3
4	0
5	1
6	2

方法返回结果

i	rel
1	3
2	3
3	3
4	2
5	2
6	1

真实结果

课堂练习：数据集成

$$CG_k = \sum_{i=1}^{k} rel_i$$

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$$

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

数据相关性分析——练习题3

- 已知6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。如果我们设计了两个检索算法，它们的返回结果分别是(b)和(c)，请问哪个方法的结果与真实结果更相似（根据**NDCG的计算结果**）。

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

(c)方法2返回结果

可以只列出计算公式，不用给出计算结果

- **0.9746**
- **0.9889**

数据预处理：数据集成

课后阅读

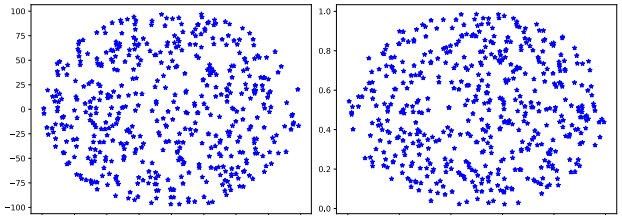
- Defu Lian, Haoyu Wang, Enhong Chen, Xing Xie. LightRec: a Memory and Search-Efficient Recommender System. WWW 2020.
- Qi Liu, Zhenya Huang, Enhong Chen,, EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction, TKDE
- Zhenya Huang, Qi Liu, Enhong Chen, et al, Question Difficulty Prediction for READING Problems in Standard Tests, AAAI'2017
- Qi Liu, Yong Ge, Enhong Chen, and Hui Xiong. Personalized Travel Package Recommendation. ICDM'2011, (Best Research Paper Award)
- PageRank算法

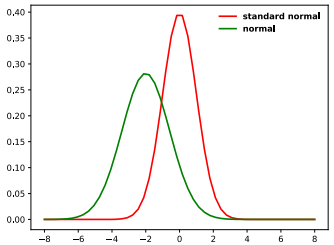
数据预处理

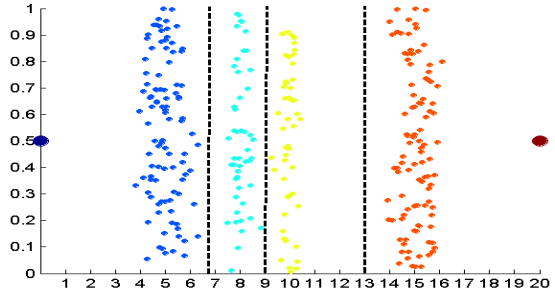
- 大数据环境下的数据特征
- 为什么需要进行预处理
- 预处理的基本方法
  - 数据清理
  - 数据集成
  - 数据变换
  - 数据规约

数据预处理：数据变换

- 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
    - 小数定标规范化
  - 数据离散化
    - 非监督离散化
    - 监督离散化

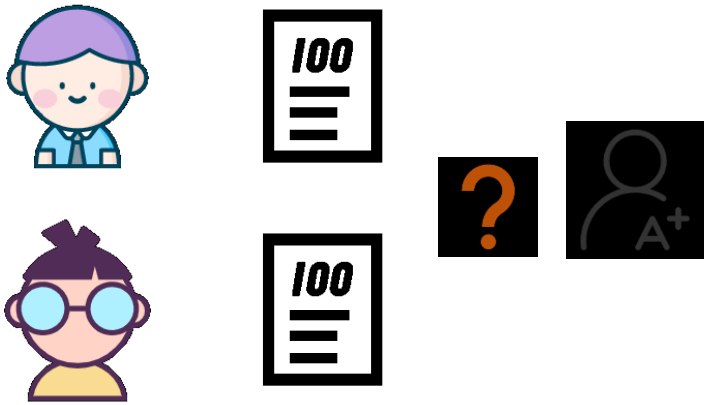






数据预处理：数据变换

- **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

- 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 [$new\_\min_A$, $new\_\max_A$]
    - 0-1规范化又称为归一化

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - 数理依据：

$$\frac{v' - new\_min_A}{new\_max_A - new\_min_A} = \frac{v - min_A}{max_A - min_A}$$

数据变换-规范化

- 最小-最大规范化

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - 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

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

快速练习：采用最小-最大规范化 -80？

数据变换-规范化

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

A

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

B

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

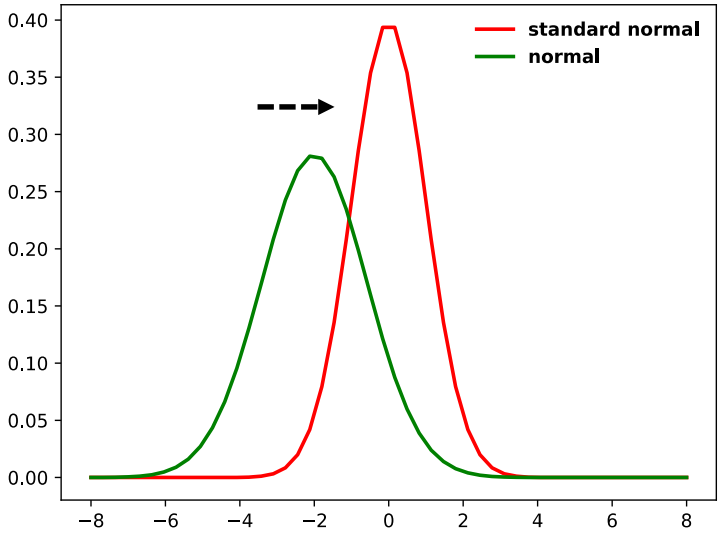
数据变换-规范化

- 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

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

  - 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

- 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？

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

- 小数定标规范化
  - 通过移动小数点的位置来进行规范化。小数点移动多少位取决于属性A的取值中的最大绝对值。

$$v' = \frac{v}{10^j}$$

其中，j 是使 Max(|v'|)<1的最小整数

  - 比如属性A的取值范围是-999到88，那么最大绝对值为999，小数点就会移动3位，即新数值=原数值/1000。那么A的取值范围就被规范为-0.999到0.088。

数据变换-规范化

- 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
小数定标规范化	算法实现简单	不适用于不同含义数据的比较，无实际意义	使用含义相同的数据，且最大最小相差较大

数据预处理：数据变换

- 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

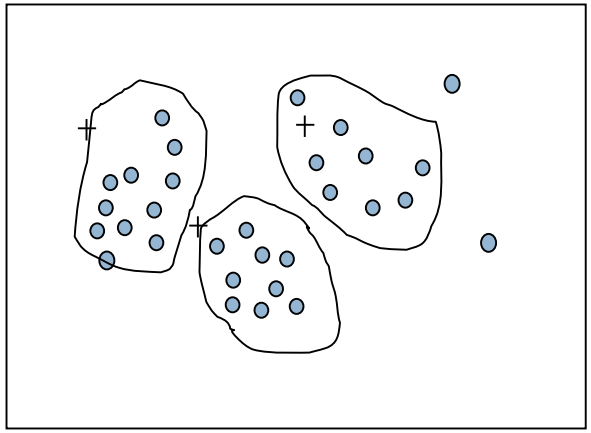
A+	A	A-
B+	B	B-
C+	C	C-
D+	D	D-
	F	

数据变换-离散化

- **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）**

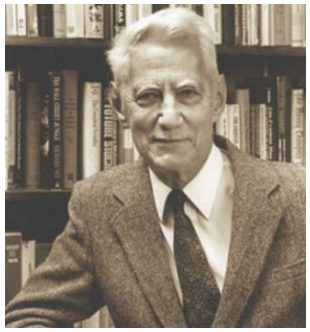
数据变换-离散化

- 非监督离散化 (参考上一节内容：**数据清理-噪声数据**)
  - 分箱
    - 1. 排序数据，并将他们分到等深的箱中
    - 2. 按箱平均值平滑、按箱中值平滑、按箱边界平滑等
  - **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

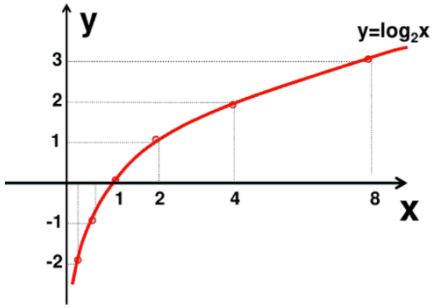
- 有监督离散化—基于熵的离散化
  - 熵用来度量系统的**不确定程度**
    - 熵是由 克劳德·艾尔伍德·香农 将热力学的熵，引入到信息论，因此它又被称为香农熵



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

- 信息熵：度量系统的不确定程度
  - 信息量
    - 定义一个时间x的概率分布为P(x)
    - 则事件x的自信息量是-logP(x)，取值范围：[0, +∞]
  - 信息熵
    - 平均而言，发生一个事件我们得到的自信息量大小
    - 即：熵可以表示为自信息量的期望



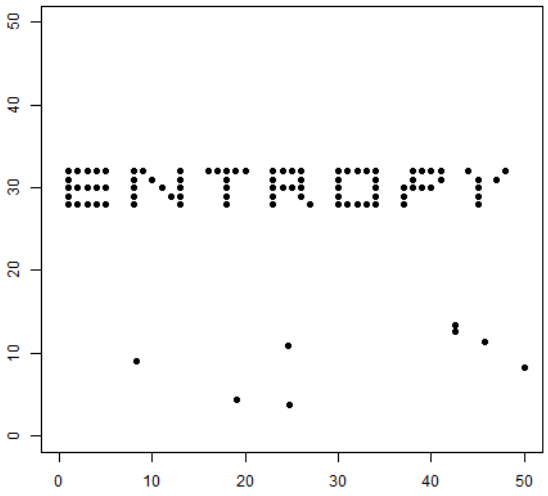
$$H = -\sum P(x)\log P(x)$$

X	0	1
P(x)	0.4	0.6

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

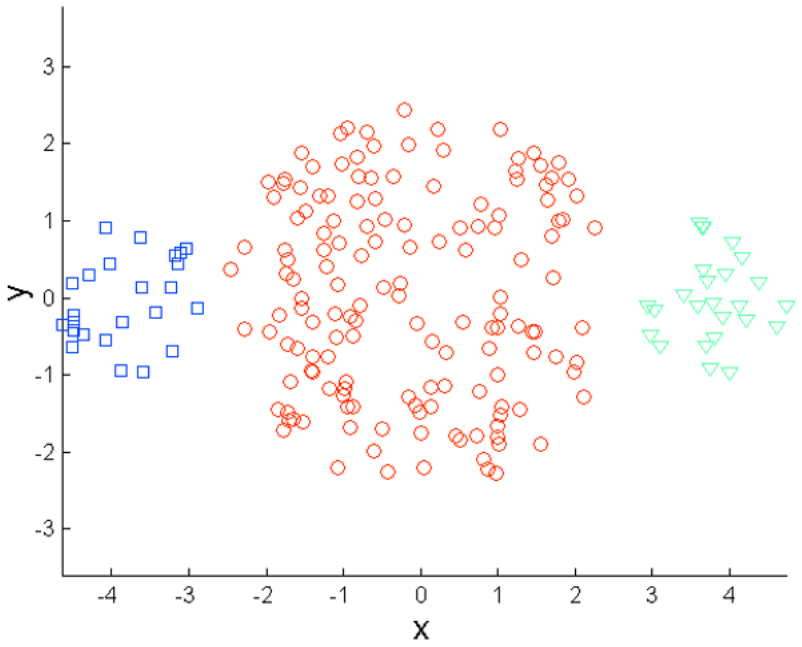
数据变换-离散化

- 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（又称“纯度”）更高因此用熵来对数据进行离散化。



数据变换-离散化

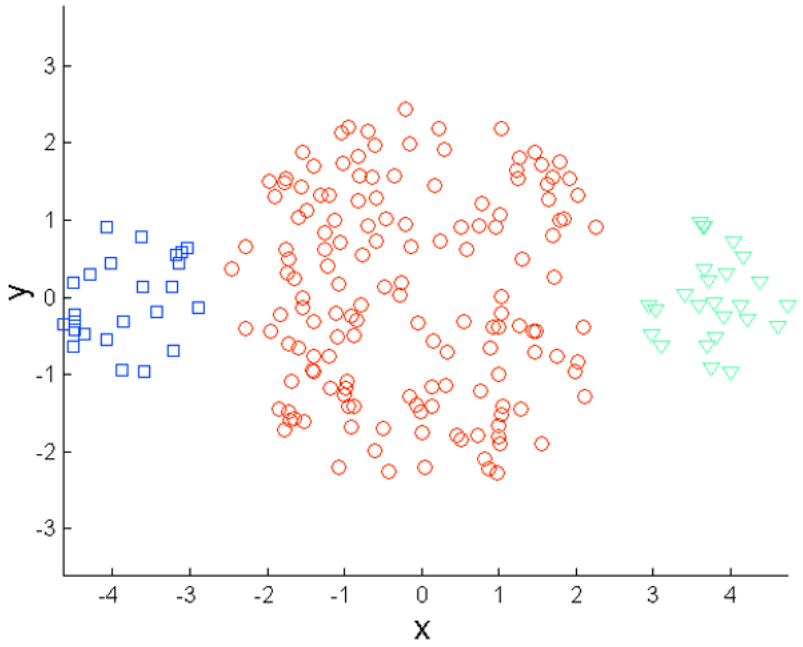
- 基于熵的离散化
  - 在x轴上对数据划分



原始数据

数据变换-离散化

- 基于熵的离散化
  - 在x轴上对数据划分



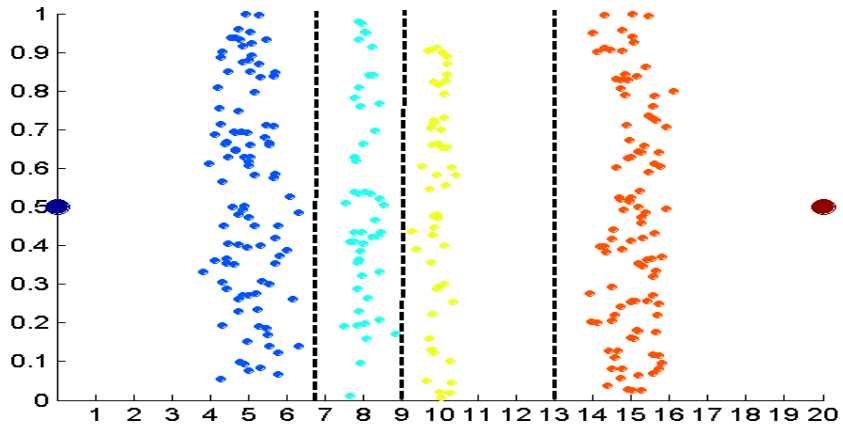
原始数据

数据变换-离散化

- 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:

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

    - 其中，$p(j \mid t)$ 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

- 计算 单个区间 的 Entropy

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

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? logC

数据变换-离散化

- 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:

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

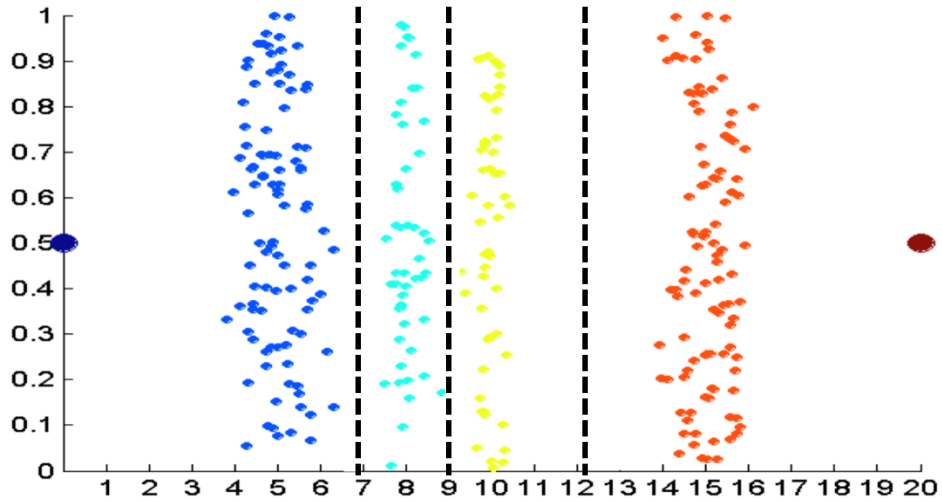
  - 其中，$p(j \mid t)$ 表示 第 j 类在区间 t 中的概率；一般对数log以2为底

结论

- **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：logC**
- **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, logC]**

数据变换-离散化

- 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隔
  - 重复以上步骤



数据变换-离散化

- **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

$Ent(p)$

属性	x_1	x_2	x_3	……	x_{n-2}	x_{n-1}	x_n
类别	C0	C0	C1	……	C1	C1	C0

x_1	x_2
C0	C0

x_3	……	x_{n-2}	x_{n-1}	x_n
C1	……	C1	C1	C0

$Ent(m_{11})$ $Ent(m_{12})$

$$Ent(m_1) = \frac{n1}{n} Ent(m11) + \frac{n2}{n} Ent(m12)$$

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

数据变换-离散化

- **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

$Ent(p)$

属性	x_1	x_2	x_3	……	x_{n-2}	x_{n-1}	x_n
类别	C0	C0	C1	……	C1	C1	C0

x_1	x_2
C0	C0

x_3	……	x_{n-2}	x_{n-1}	x_n
C1	……	C1	C1	C0

x_1	x_2	x_3	……	x_{n-2}
C0	C0	C1	……	C1

x_{n-1}	x_n
C1	C0

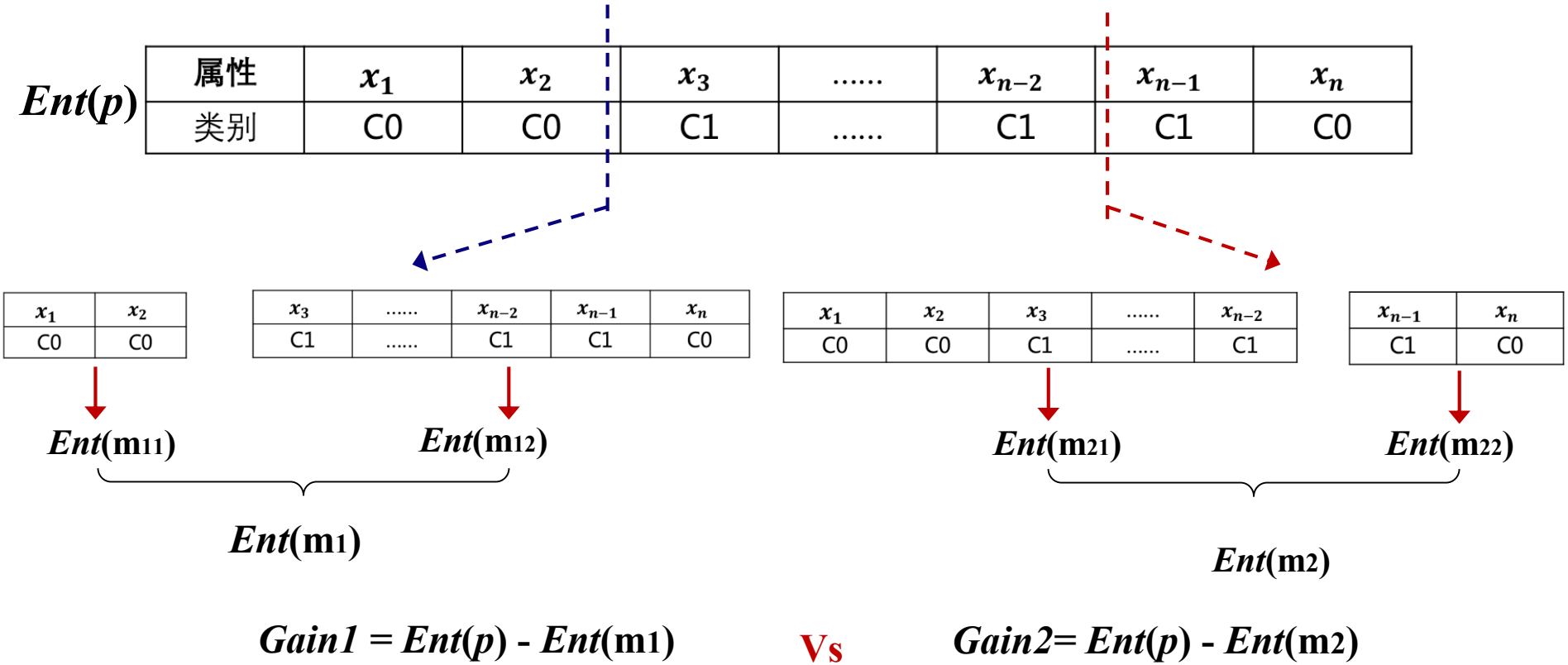
$Ent(m_{11})$ $Ent(m_{12})$ $Ent(m_{21})$ $Ent(m_{22})$

$$Ent(m_1) = \frac{n1}{n} Ent(m11) + \frac{n2}{n} Ent(m12)$$

$Ent(m_2)$

数据变换-离散化

- **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

- **如何确定分隔点？**--计算分隔后的信息增益
  - 信息增益（Information Gain）：

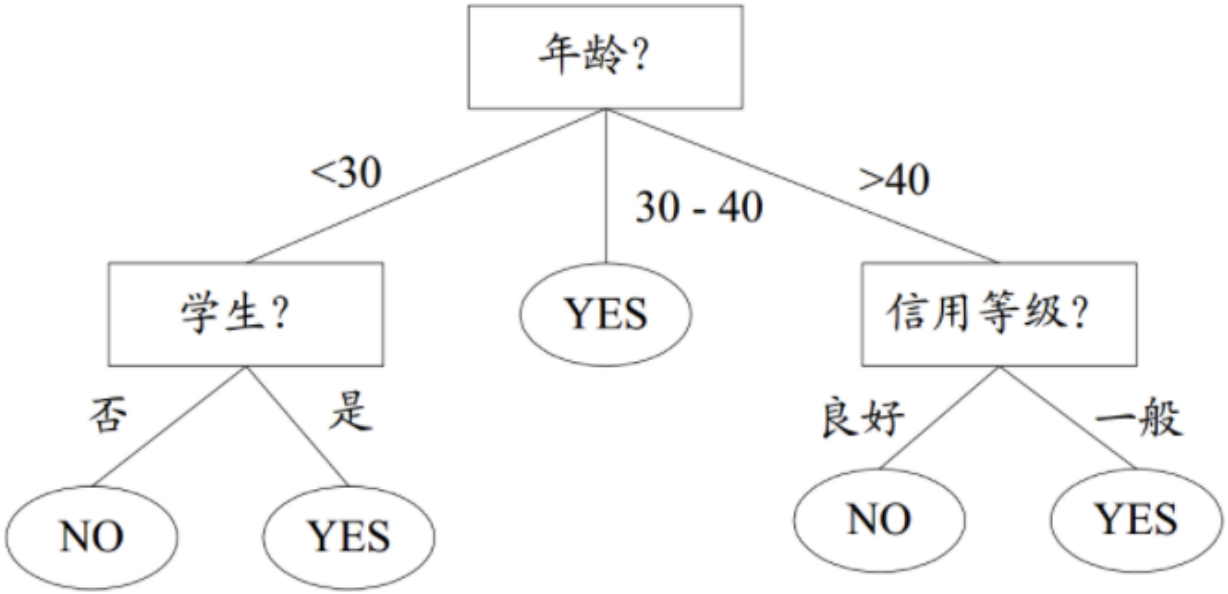
$$GAIN_{\text{split}} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

    - 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

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

数据变换-离散化

- 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

- 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




1月 2月 3月 4月 5月 6月 7月 8月 9月 10月 11月 12月

  - 根据景点的类别分布，计算区间（季节）中的熵:

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

Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. Personalized Travel Package Recommendation. ICDM'2011. Best Research Paper

课后学习

- 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

- 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

课堂练习：计算信息增益

- **问题描述**：假设一组连续值及其所属类别如下表所示，利用信息增益求第一次划分的分隔点。

属性	0.243	0.245	0.437	0.481	0.608	0.666
类别	C0	C0	C1	C1	C1	C0

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

$\log_2(1/3) = -1.585$, $\log_2(2/3) = -0.585$
$\log_2(3/5) = -0.737$, $\log_2(2/5) = -1.322$
$\log_2(3/4) = -0.415$, $\log_2(1/4) = -2.0$