大数据分析与应用

2020春

第1章 数据/数据预处理

- 1.1 数据对象与数据属性
- 1.2 数据的基本统计描述
- 1.3 数据预处理
  - 数据清洗
  - 数据抽样
  - 数据降维
- 1.4 距离与相似度计算

1.1 数据对象与数据属性

- 数据集：多个同类数据对象的集合

  - 初始的数据：图像、声音、文本、数据库记录

  - 最终的形式：矩阵/表格

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- 对数据的第一印象(影响后续方法的选择)

✓Dimensionality ： 维度

  - 维度之间有没有相关性

✓Sparsity：稀疏度

  - 数据出现的次数

✓Distribution：分布

  - 高斯？均匀？

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

1.1 数据对象与数据属性

- 数据对象的基本元素

✓数据行（数据对象）： *samples 样本，examples 样例, instances 用例, data points 数据点, objects 对象, tuples 元组.*

✓数据列（数据属性）： *attribute 属性，dimensions 维度, features 特征, variables 变量*

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

1.1 数据对象与数据属性

- 数据的取值类别
  - 标称属性 （Nominal） ：
    - 有序：成绩 = { A,B,C,D,F }
    - 无序：颜色= {black, blond, brown, grey, red, white }
    - 邮政编码、电话号码
  - 二进制属性 （Binary）
    - 对称的二进制：gender
    - 不对称的：检查结果={positive，negtive}
  - 数值属性（Numeric）：
    - 区间标度：气温 =（-50，50）
    - 比率标度：湿度=（0,100）：潮湿的程度20是10的2倍，所以称之为比率

1.1 数据对象与数据属性

●数据类型的相互转换

□标称属性：有序数关系的标称属性的处理

$$r_i \in \{1,\ldots, M_i\} \qquad z_i = \frac{r_i - 1}{M_i - 1}$$

□标称属性：无序数关系，例如：红、黄、蓝、绿

- 方法1: 简单匹配:相等为0，不相等为1 （one hot 编码，独热编码）
- 方法2: 编码为一系列的二进制属性：可以人为的设定不同取值之间的距离

原有属性：颜色	新属性：红	新属性：黄	新属性：绿	新属性：蓝
红	1	0	0	0
绿	0	0	1	0
蓝	0	0	0	1

问题1：冗余
问题2：颜色之间距离是否客观

1.2 数据的基本统计描述

- 基本的数据特征

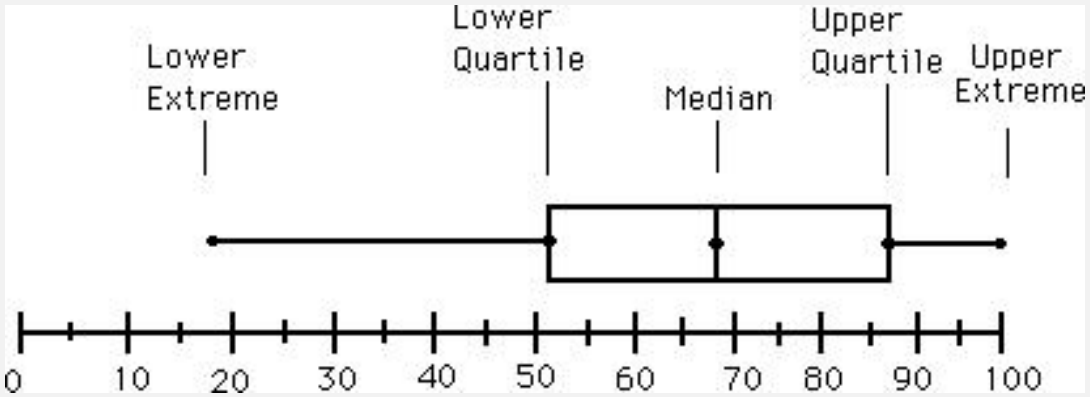
1.中心趋势度描述：

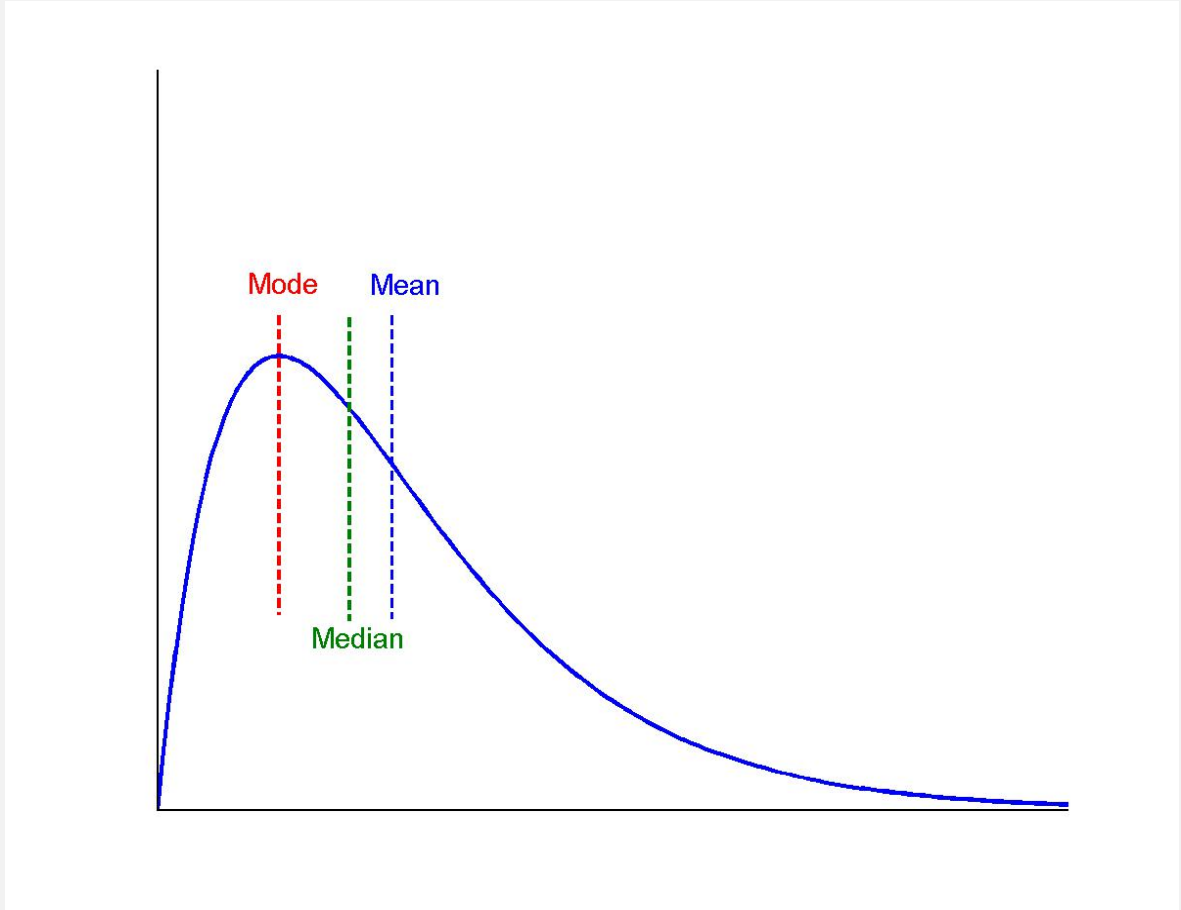
中位数，均值，众数

2.数据的分散度量：

max, min 极大极小值、quantiles 四分位数、

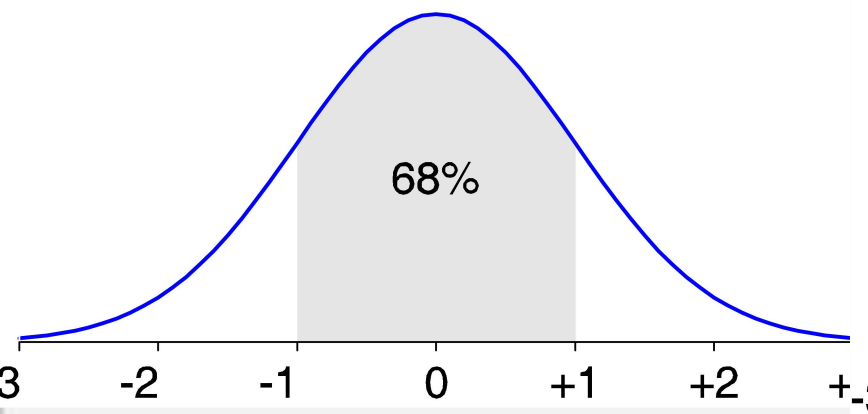
3.数据的离群点

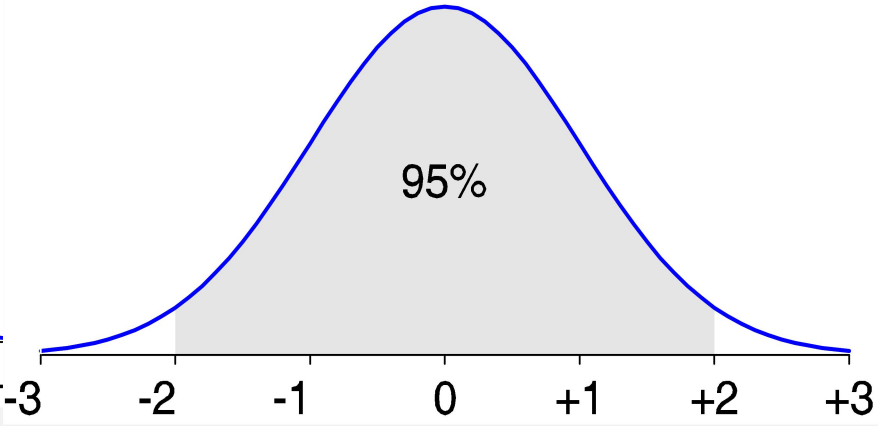


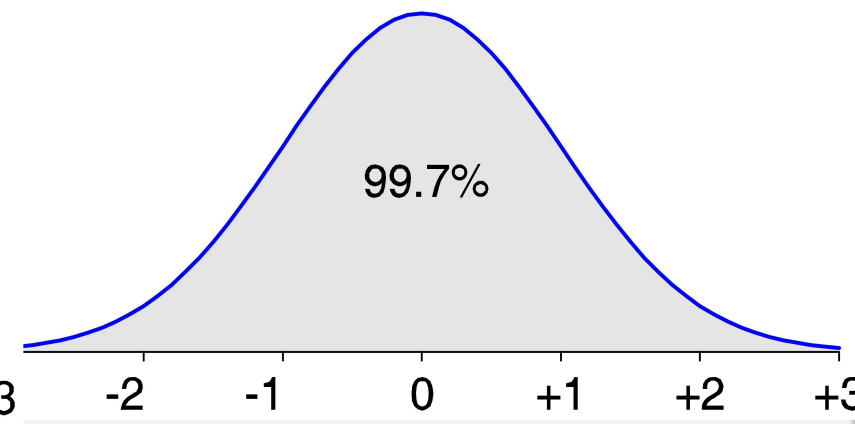


1.2 数据的基本统计描述

- 3. 离群点的选择







1.2 数据的基本统计描述

●基本度量的应用：

□数据归一化：

- 1）特征数值对最后结果的作用不受物理度量数值的影响；2）模型输入的需要

归一化为零均值单位方差

$$z=\frac{x-\mu}{\sigma}$$

归一化到[0,1]区间

$$z=\frac{x-\min}{\max-\min}$$

□数据离散化：应该如何划分数据

- 出于模型（决策树）的需要，使用区间标签或者概念标签替代数值：年龄：20->青年 。

□数据连续化：应该如何生成合理的数据分布

1.3 数据预处理

• 1.3.1数据清洗

- 为什么要数据清洗：
  - **是否准确** Accuracy: correct or wrong, accurate or not
  - **是否完整** Completeness: not recorded, unavailable, …
  - **是否一致** Consistency: some modified but some not, dangling, …
  - **是否及时** Timeliness: timely update?
  - **是否可信** Believability: how trustable the data are correct?
- 情况：
  - *空值或默认值：Occupation*=“ ” (missing data) “NULL”
  - *不合理值：Salary*=“−10” (an error)
  - *不一致值：Age*=“42”, *Birthday*=“03/07/2018”
  - 一致但使用不同的记录规则：“甲 乙 丙”，“A, B, C”； 2017/9 和 9/2017

1.3 数据预处理——数据清洗

- **手段1：人工修改**
- **手段2：直接删除**：直接删除数据行，删除缺失过多的列
- **手段3：自动填写**的处理方法：
  - 增加一个“未知”类别 e.g.，“unknown”：具有一定的意义
  - 填写本字段的均值、出现概率最大的值
  - 填写本类均值（在更小类别中：例如男生的平均身高，而不是全体学生的）
  - **局部分析（临近点分析）：**
    - **1）计算与缺失数据临近的数据点，查看字段值；**
    - **2）插值：在时间序列中，样本的顺序提供的临近点信息；**

*大多数数据表的数据行和数据列默认是无顺序的

1.3 数据预处理——数据清洗

- 面对具有噪声的数据：
  - 分箱（离散化：不需要特征具有高分辨率）
  - 平滑（零均值噪声）
  - 剔除离群点

1.3 数据预处理——检测何时需要处理

- 数据依赖性检测

1）使用元数据Metadata (e.g., domain, range, dependency, distribution)
- 编码使用的不一致
- 数据表示的不一致 2017/9 和 9/2017

2）规则检查 （与数据库课程内容类似）
- 唯一性规则
- 连续性规则
- 空值规则

1.3 数据预处理——数据抽样

- 数据抽样的作用：

  1）数据集太大，只需要计算其中一部分就可以得到较好的模型与结果；（注意事项：抽样后的数据集与原数据集具有一致的数据分布；）

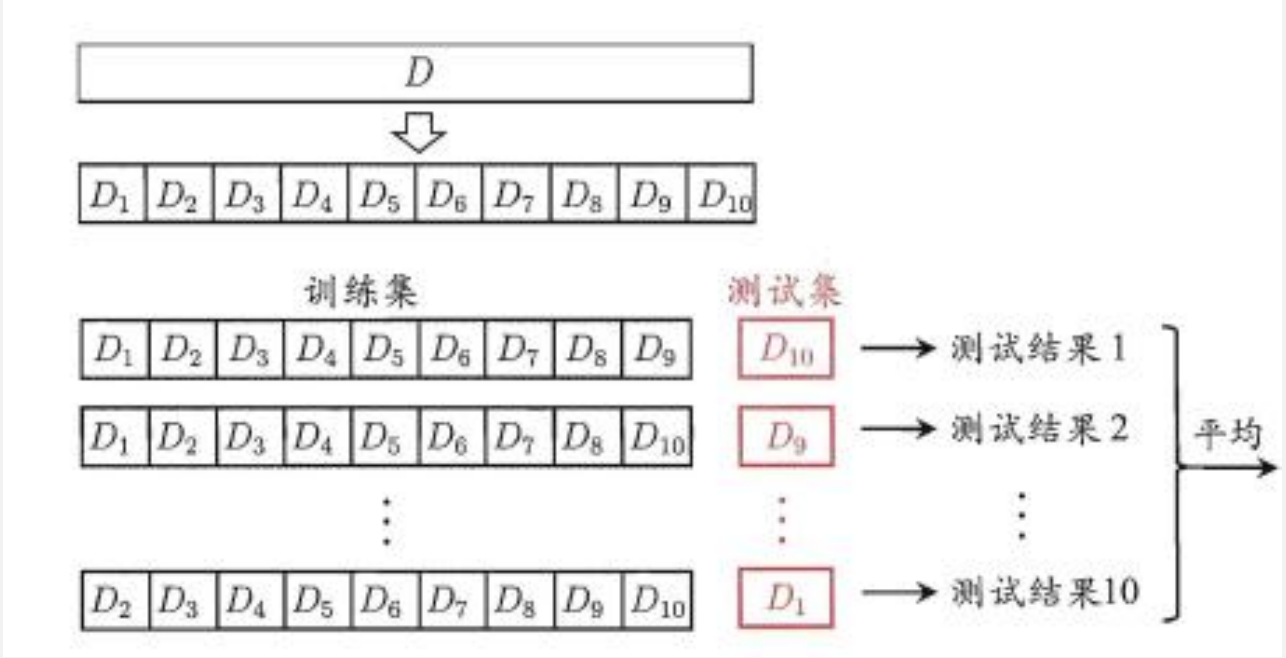
  2）划分训练集与测试集；（用于训练模型的数据与用于测试模型优劣的数据必须不尽相同；分布应该一致）

  几个名词：

  ◆验证集（在模型训练的过程中使用）

  ◆交叉验证

- 方法：有放回抽样、无放回抽样

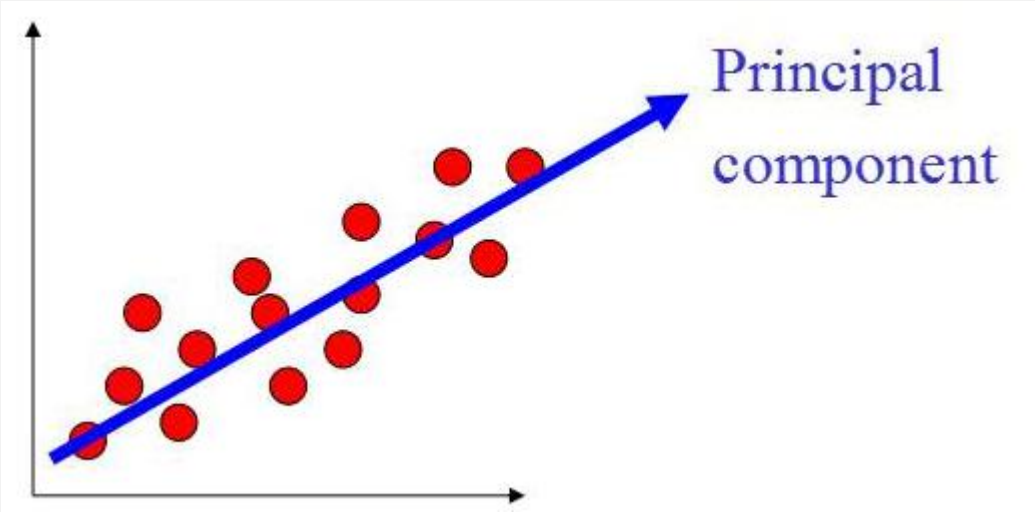


1.3 数据预处理——数据降维

□选择出对后续数据分析有用的数据列（数据列的组合与变换）

●极大地影响模型的训练效果

- 主成分分析（PCA）
- 奇异值分解（SVD）
- 人工



1.4 数据的相似性与距离

相似性与相异性计算是最本质的数据计算

■数据行的相似性

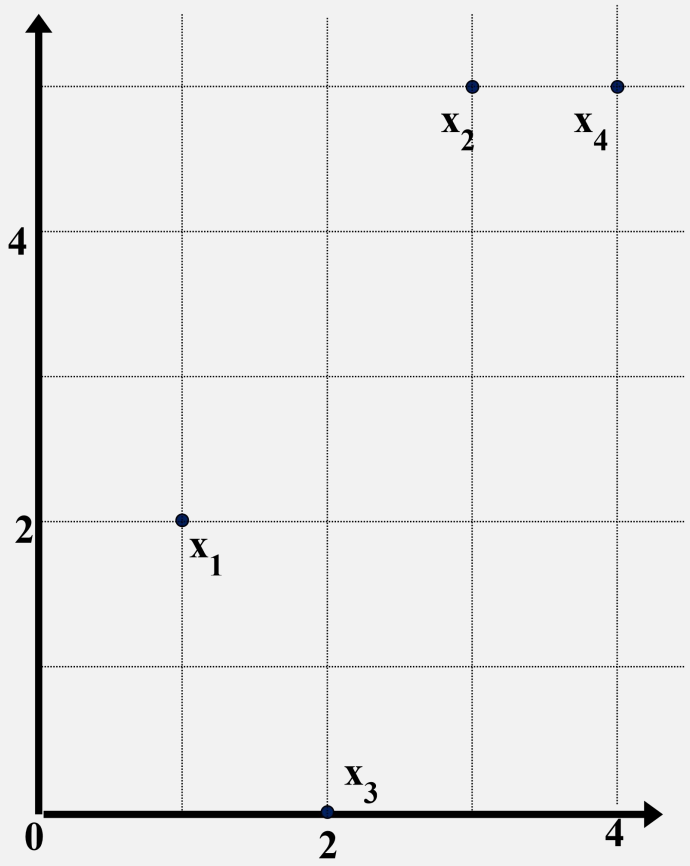
- ●相似性、相异性、临近性、距离：表达了同一个内涵
- ●标称属性、二进制属性怎么计算?

■数据列的相似性

- ●正相关/负相关/不相关
- ●意义：减少存储/避免计算出错/特征提取的需要

1.4 数据的相似性与距离

数据结构：相似矩阵



数据矩阵

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

距离矩阵

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

1.4 数据的相似性与距离

- 广义的距离
- 例如闵可夫斯基距离：L-*h* norm

$$d(i,j)=\sqrt[h]{|x_{i1}-x_{j1}|^h+|x_{i2}-x_{j2}|^h+\cdots+|x_{ip}-x_{jp}|^h}$$

- 只要符合下列条件，都可以作为距离的度量
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (正定性Positive definiteness)
  - d(i, j) = d(j, i) (对称性Symmetry)
  - d(i, j) ≤ d(i, k) + d(k, j) (三角不等式Triangle Inequality)

1.4 数据的相似性与距离

- h = 1: L-1 距离 （曼哈顿距离 ）
  - 在每个特征上的距离的绝对值之和

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$$

- h = 0: L-0距离
  - 数值不同的特征的数量

$$d(i,j)=(|x_{i1}-x_{j1}|^0+|x_{i2}-x_{j2}|^0+...+|x_{ip}-x_{jp}|^0)$$

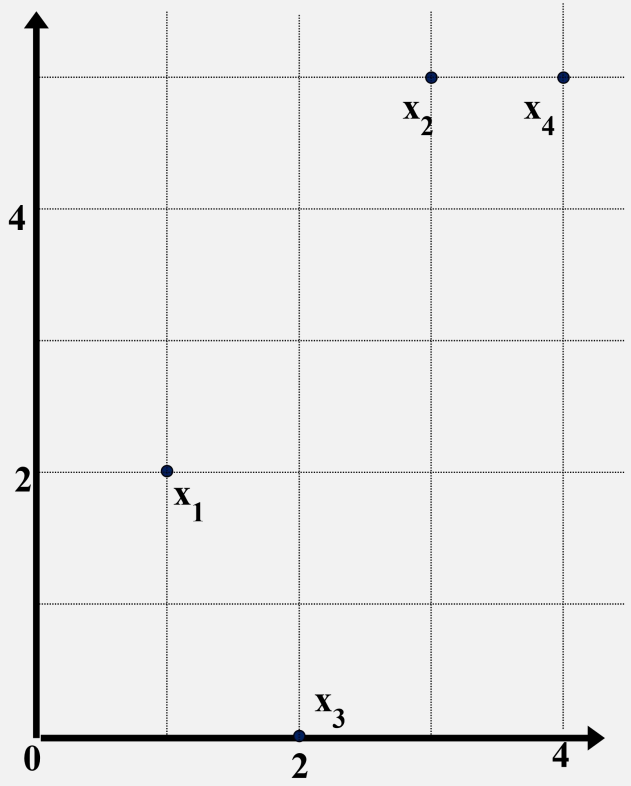
- $h \to \infty$.
  - 距离最大的特征之间的距离，作为数据点的距离

$$d(i,j)=\lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if}-x_{jf}|^h\right)^{\frac{1}{h}}=\max_{f}^{p}|x_{if}-x_{jf}|$$

1.4 数据的相似性与距离

举例：

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

1.4 数据的相似性与距离

余弦相似度：不考虑绝对的数值所造成的差异

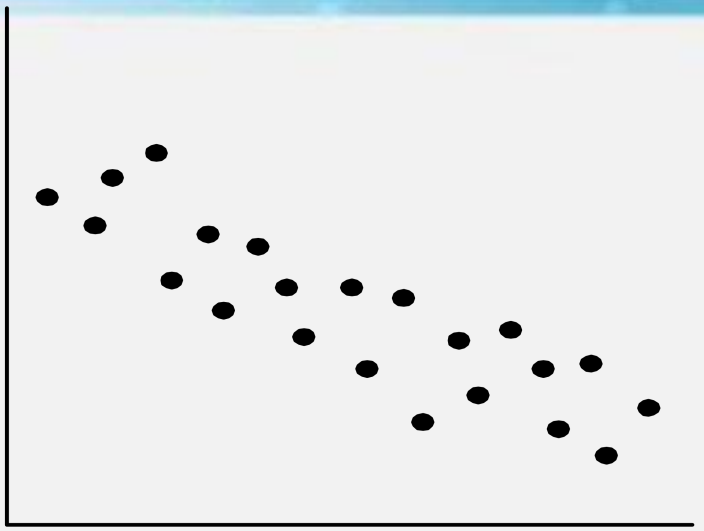
Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

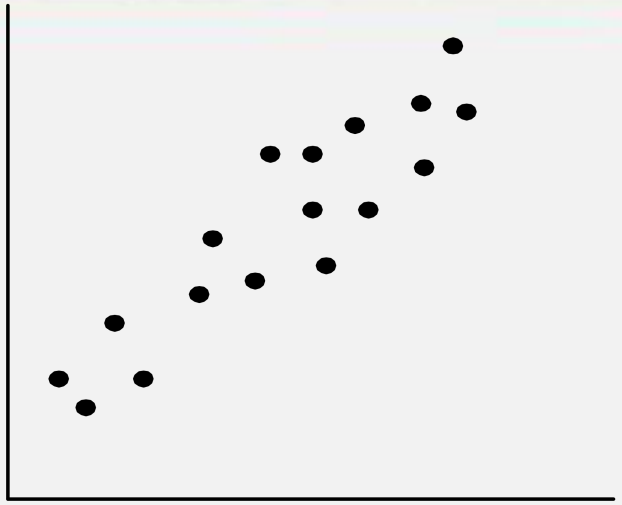
- $\cos(d_1, d_2) = (d_1 \bullet d_2) / \|\|d_1\|\| \, \|\|d_2\|\|$,

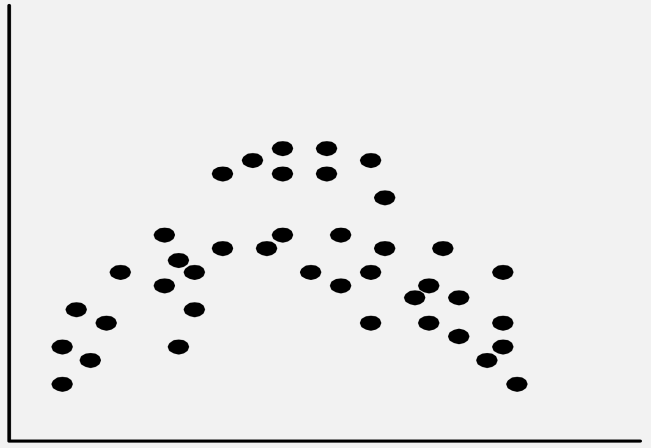
d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

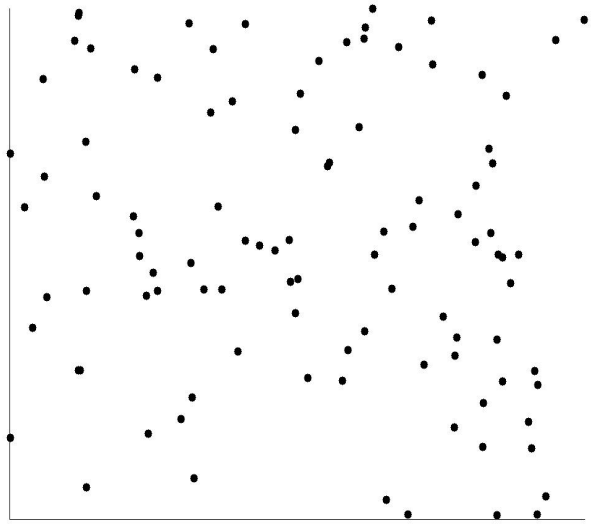
$d_1 \bullet d_2$ = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25
$\|\|d_1\|\|$= $(5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}=(42)^{0.5}$ = 6.481
$\|\|d_2\|\|$= $(3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}=(17)^{0.5}$ = 4.12
$\cos(d_1, d_2)$ = 0.94

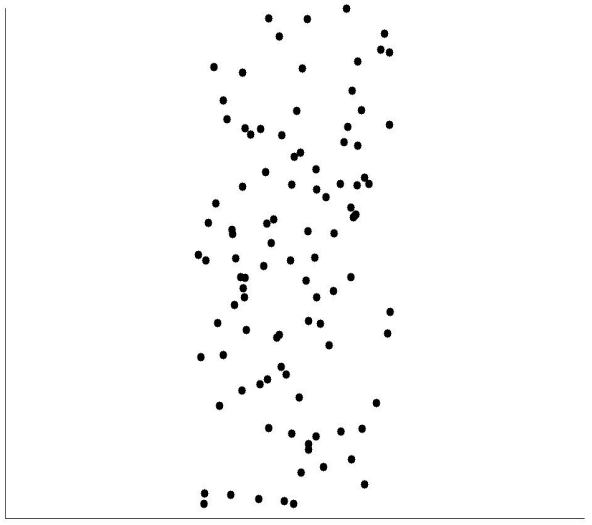
1.4 数据的相似性与距离——列的相似性

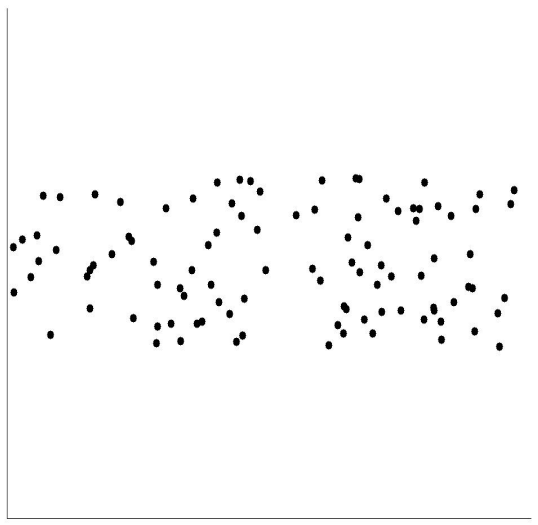






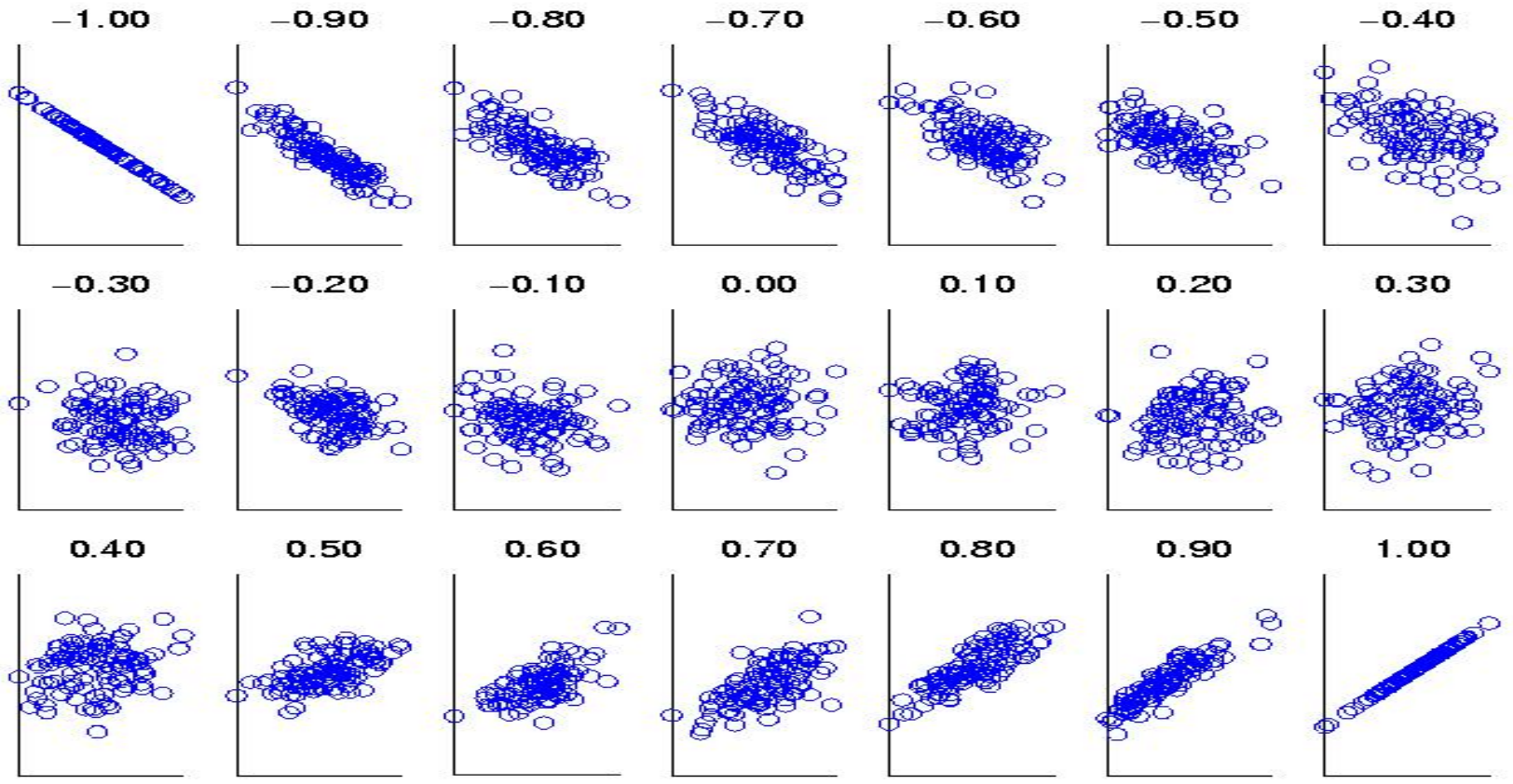






1.4 数据的相似性与距离

- 数值属性



$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

协方差与相关系数

1.4 数据的相似性与距离

二进制属性：注意非对称情况下的处理

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

$$d(i,j)=\frac{r+s}{q+r+s+t}$$

$$d(i,j)=\frac{r+s}{q+r+s}$$

$$sim_{Jaccard}(i,j)=\frac{q}{q+r+s}$$

- Jaccard 系数等同于相关性

$$coherence(i,j)=\frac{sup(i,j)}{sup(i)+sup(j)-sup(i,j)}=\frac{q}{(q+r)+(q+s)-q}$$

1.4 数据的相似性与距离

- 二进制属性：卡方检验
- 例子

姓名	会下棋	喜欢科幻
小明	1	1
小红	0	0
小刚	1	0
。。。		

	会下棋	不会下棋	Sum (行)
喜欢科幻小说	250(90)	200(360)	450
不喜欢科幻小说	50(210)	1000(840)	1050
Sum(列.)	300	1200	1500

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$