

2024-2025学年 第1学期(秋)



数据挖掘

第3章 数据预处理

2024 年 9 月

数据质量：通用AI的关键

AI系统 = 数据 + 模型/算法

前期训练数据准备 + 后期数据飞轮迭代

模型训练

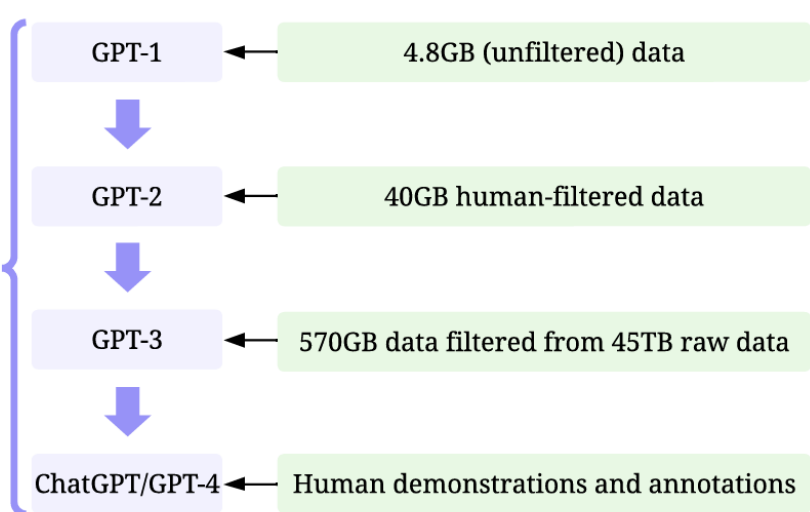
80%

20%

**数据决定了机器学习的上限，
模型和算法只是逼近这个上限。**

- 在OpenAI的GPT3、4模型以及谷歌的PaLM系列模型训练中，大量用到了专有数据，如**高质量书籍数据**和**社交媒体对话数据**等。
- 近期很多研究人员开始转向研究以数据为中心的AI研究，其主要目的就是想办法加强**数据的质量和数量**，而不过多的考虑模型或者说固定模型结构。

Similar model architectures



Data size ↑
Data quality ↑

大量的高质量训练数据是LLM模型成功的关键驱动力

数据质量提升

对于各个大模型来说，在预训练之前，都需要一个精心设计的 Pipeline 来提升预训练数据的质量。

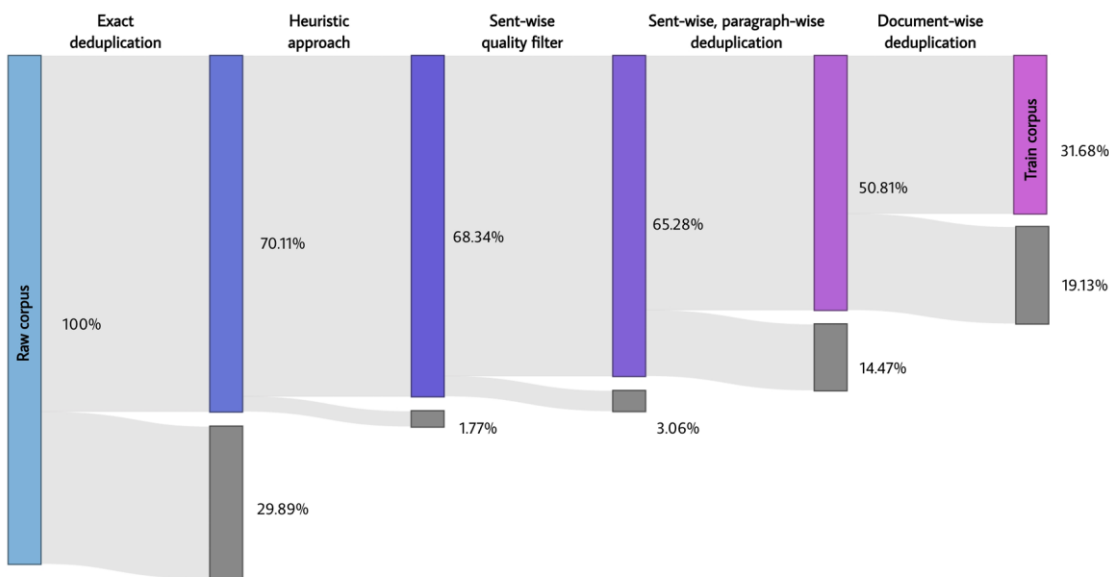


Figure 2: The data processing procedure of Baichuan 2's pre-training data.

Baichuan2 提升数据质量的 Pipeline^[1]

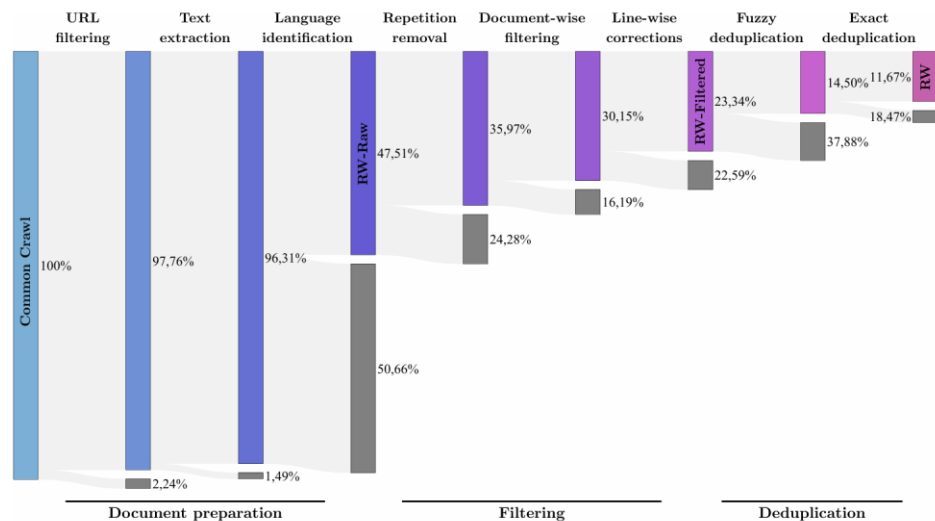


Figure 2. Subsequent stages of Macrodata Refinement remove nearly 90% of the documents originally in CommonCrawl. Notably, filtering and deduplication each result in a halving of the data available: around 50% of documents are discarded for not being English, 24% of remaining for being of insufficient quality, and 12% for being duplicates. We report removal rate (grey) with respect to each previous stage, and kept rate (shade) overall. Rates measured in % of documents in the document preparation phase, then in tokens.

RefinedWeb 的过滤和去重流程^[2]

[1] Yang A, Xiao B, Wang B, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint arXiv:2309.10305, 2023.

[2] Penedo G, Malartic Q, Hesslow D, et al. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only[J]. Advances in Neural Information Processing Systems, 2023.

数据质量的衡量标准



数据预处理 —— 主要任务

- **数据清理**

- 填写缺失值，平滑噪声数据，识别或删除离群，并解决不一致问题

- **数据集成**

- 整合多个数据库，多维数据集或文件

- **数据规约**

- 降维
- 降数据
- 数据压缩

- **数据转换**

- 规范化
- 离散化



南京大學
NANJING UNIVERSITY

目录

01

数据清洗

02

数据集成

03

数据规约

04

数据转换

数据清洗

- 属性值缺失：
 - 例如，职业= “ ” (丢失)
- 噪音，错误或离群
 - 例如，工资= “-10” (错误)
- 不一致的代码或不符的名称
 - 年龄= “42” 生日= “03/07/1997”
 - 曾经评级 “1,2,3” , 现在评级 “A, B, C”

数据清洗 —— 如何处理丢失数据？

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenge	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen	female	26	0	0	STON/O2	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, female		27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstror	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S

数据清洗 —— 如何处理丢失数据？

- **忽略元组**：当类标号缺少时通常这么做（监督式机器学习中训练集缺乏类标签）。当每个属性缺少值比例比较大时，效果比较差
- **手动填写遗漏值**：工作量大
- **自动填写**
 - 使用属性的平均值填充空缺值【代码见下链接，请同学们课后实践】
 - 最有可能的值：基于诸如贝叶斯公式或决策树推理

6.4.2. Univariate feature imputation

The `SimpleImputer` class provides basic strategies for imputing missing values. Missing values can be imputed with a provided constant value, or using the statistics (mean, median or most frequent) of each column in which the missing values are located. This class also allows for different missing values encodings.

The following snippet demonstrates how to replace missing values, encoded as `np.nan`, using the mean value of the columns (axis 0) that contain the missing values:

```
import numpy as np
from sklearn.impute import SimpleImputer
imp = SimpleImputer(missing_values=np.nan, strategy='mean')
imp.fit([[1, 2], [np.nan, 3], [7, 6]])

X = [[np.nan, 2], [6, np.nan], [7, 6]]
print(imp.transform(X))
```

<https://scikit-learn.org/stable/modules/impute.html#impute>

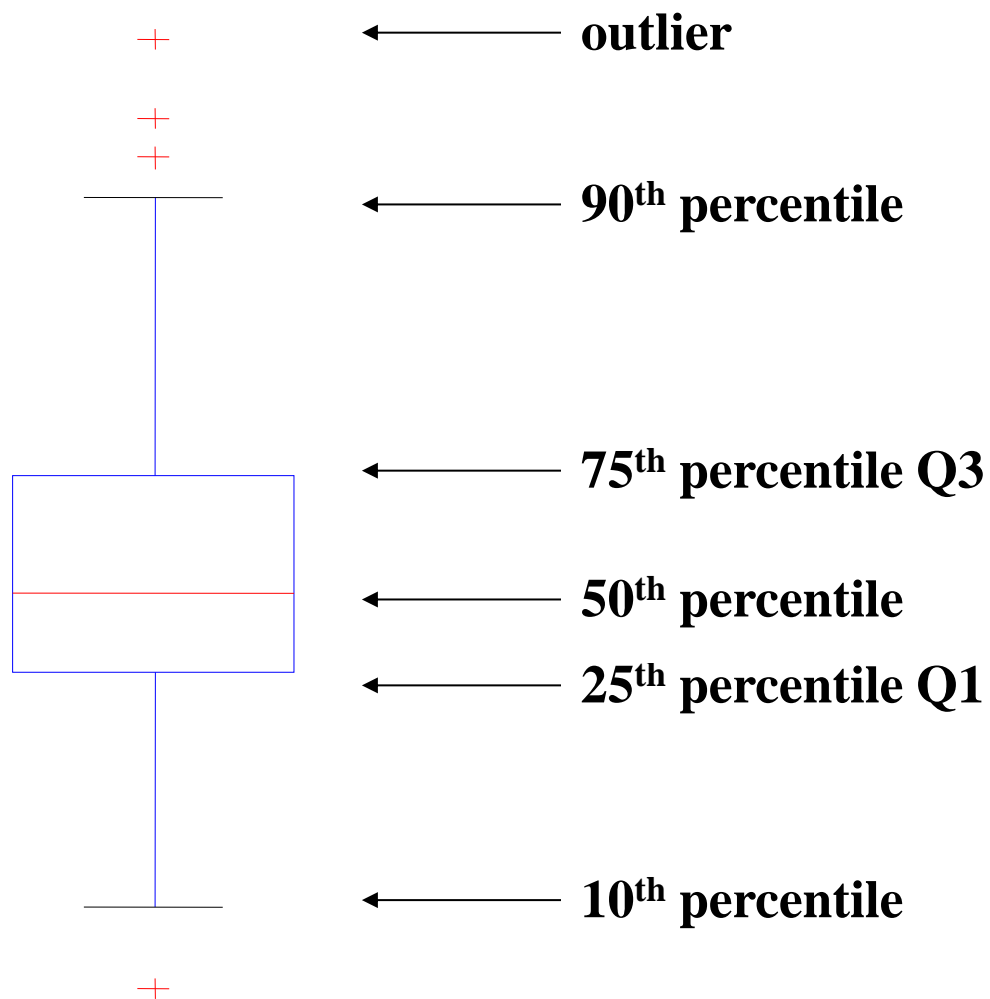
数据清洗 —— 如何处理丢失数据？

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenge	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen	female	26	0	0	STON/O2	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male	28	0	0	330877	8.4583		Q
8	7	0	1	McCarthy	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, female		27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstror	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S

平均值

数据清洗 —— 如何处理噪声数据？

- 箱线图检测离群数据：删除离群点



数据清洗 —— 如何处理不一致数据？

- **不一致的代码或不符的名称**
 - 年龄= “42” 生日= “09/24/1998”
 - 曾经评级 “1,2,3” , 现在评级 “A, B, C”
- **方法**
 - 计算推理、替换
 - 全局替换



南京大學
NANJING UNIVERSITY

目录

01

数据清洗

02

数据集成

03

数据规约

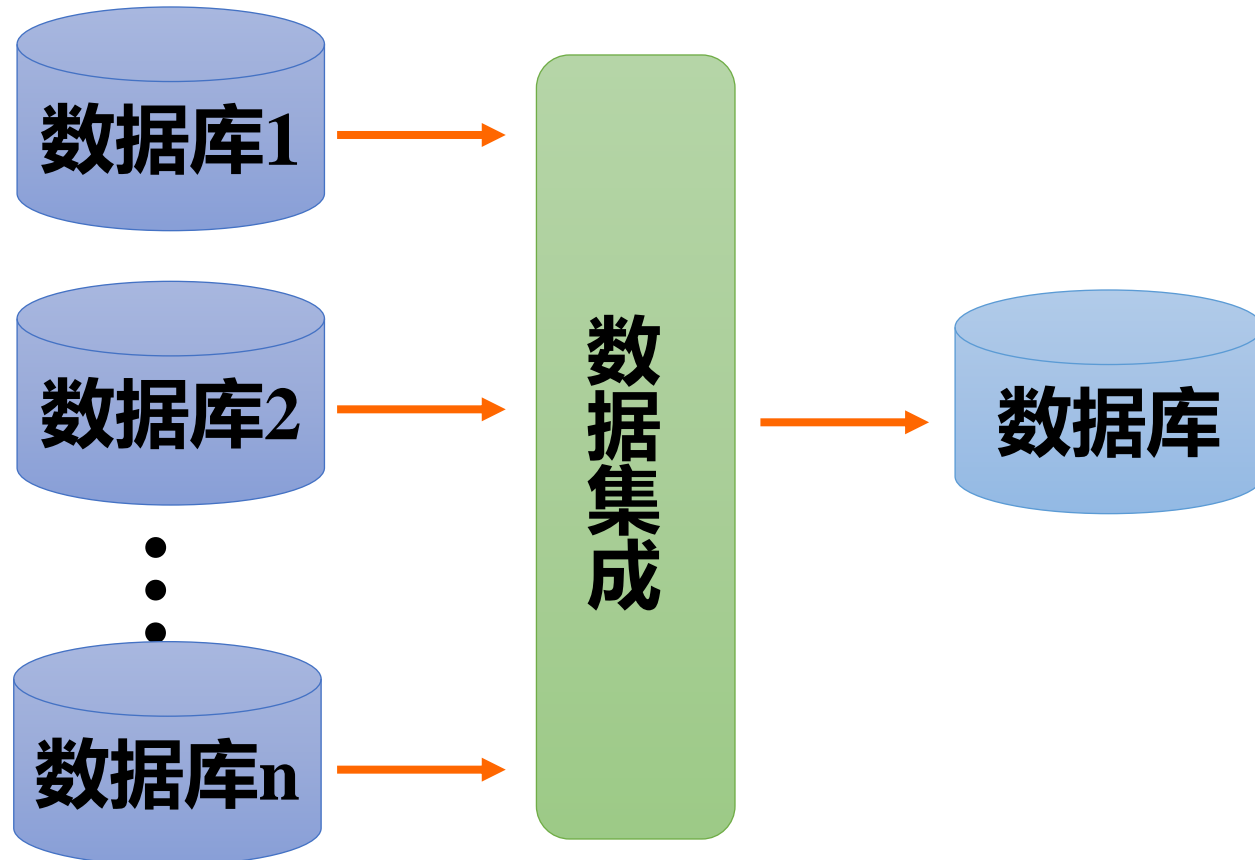
04

数据转换

数据集成

- 数据集成

—— 将来自多个数据源的数据组合成一个连贯的数据源



数据集成 —— 模式集成

- 数据集成：将来自多个数据源的数据组合成一个连贯的数据源
 - 模式集成：整合来自不同来源的元数据

数据库A		
cust-id	name	height
1	张三	1.68
2	李四	1.76

数据库B		
cust-#	name	height
1	Zhang San	5.51
2	Li Si	5.77


数据集成				
id	nameA	heightA	nameB	heightB
1	张三	1.68	ZS	5.51
2	李四	1.76	LS	5.77

数据集成 —— 实体识别问题

- 数据集成: 将来自多个数据源的数据组合成一个连贯的数据源
 - 实体识别问题: 例如, Bill Clinton = William Clinton
 - 识别来自多个数据源的真实世界的实体,

数据库A		
cust-id	name	height
1	张三	1.68
2	李四	1.76

数据库B		
cust-#	name	height
1	Zhang San	5.51
2	Li Si	5.77



数据集成			
id	name	heightA	heightB
1	张三	1.68	5.51
2	李四	1.76	5.77

数据集成 —— 数据冲突检测 and 解决

- 数据集成：将来自多个数据源的数据组合成一个连贯的数据源
 - 数据冲突检测 and 解决
 - 对于同一个真实世界的实体，来自不同源的属性值
 - 可能的原因：不同的表述，不同的尺度，例如，公制与英制单位

数据库A		
cust-id	name	height
1	张三	1.68
2	李四	1.76

数据库B		
cust-#	name	height
1	Zhang San	5.51
2	Li Si	5.77



数据集成		
id	name	height
1	张三	1.68
2	李四	1.76

数据集成

- **数据集成：** 将来自多个数据源的数据组合成一个连贯的数据源
 - 模式集成
 - 实体识别问题
 - 数据冲突检测 and 解决

数据库A		
cust-id	name	height
1	张三	1.68
2	李四	1.76

数据库B		
cust-#	name	height
1	Zhang San	5.51
2	Li Si	5.77



数据集成		
id	name	height
1	张三	1.68
2	李四	1.76

数据集成中的冗余信息的处理

- 整合多个数据库经常发生数据冗余
 - *Object identification*: 相同的属性或对象可能有不同的名字在不同的数据库中
 - *Derivable data*: 一个属性可能是“派生”的另一个表中的属性, 例如, 跑步能力

数据库A		
cust-id	name	3000m
1	张三	13.24
2	李四	11.26

数据库B		
cust-#	name	5000m
1	Zhang San	25.35
2	Li Si	21.27



数据集成		
id	name	run
1	张三	15.24
2	李四	12.14

数据集成中的冗余信息的处理

- 整合多个数据库经常发生数据冗余
 - *Object identification*: 相同的属性或对象可能有不同的名字在不同的数据库中
 - *Derivable data*: 一个属性可能是“派生”的另一个表中的属性, 例如, 跑步能力
- 通过相关性分析和协方差分析可以检测到冗余的属性
- 仔细集成来自多个数据源, 可能有助于减少/避免冗余和不一致的地方, 并提高读取速度和质量

相关分析（离散变量）

姓名	是否下棋
张三	1
王五	0
马六	0
...	...

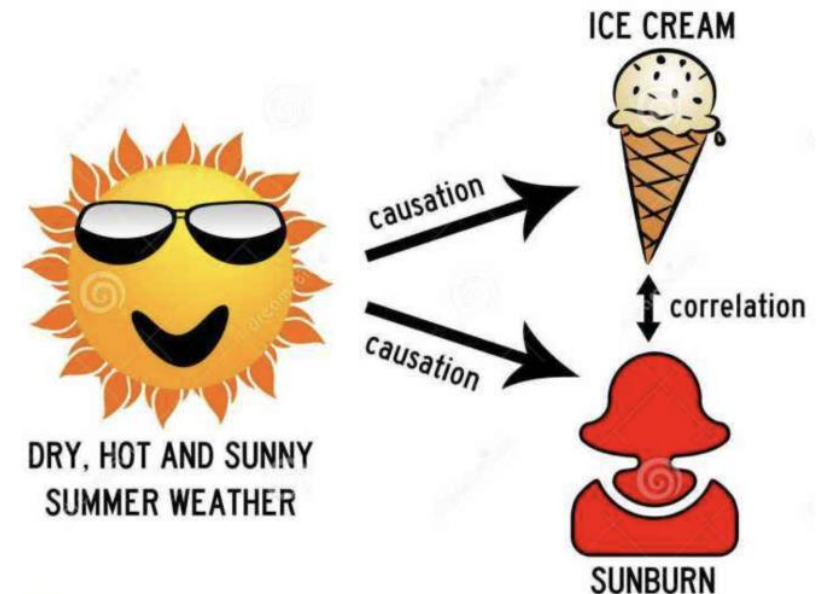
姓名	是否看科幻小说
张三	1
王五	1
马六	0
...	...

相关分析

χ^2 (chi-square) test

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- χ^2 值越大，越有可能变量是相关的
- 相关性并不意味着因果关系
 - # of hospitals and # of car-theft in a city 是相关的
 - 两者都因果联系的第三个变量为人口



相关分析 —— χ^2 (chi-square) test 举例

姓名	是否下棋
张三	1
王五	0
马六	0
...	...

姓名	是否看科幻小说
张三	1
王五	1
马六	0
...	...

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

行合计乘以列合计除以总数

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

相关分析 —— χ^2 (chi-square) test 举例

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

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

这表明，组中的like_science_fiction和play_chess相关

相关分析 —— 相关性数据集成

姓名	是否下棋
张三	1
王五	0
马六	0
...	...

姓名	是否看科幻小说
张三	1
王五	1
马六	0
...	...



数据集成	
姓名	是否兴趣爱好
张三	1
王五	1
马六	0
...	...

相关分析（连续变量）

姓名	病人数目
长沙	30000
武汉	50000
广州	80000
...	...

姓名	小偷数目
长沙	2000
武汉	3500
广州	6000
...	...

相关分析

- 相关系数（也称为皮尔逊相关系数）

$$r_{p,q} = \frac{\sum(p - \bar{p})(q - \bar{q})}{(n - 1)\sigma_p\sigma_q} = \frac{\sum(pq) - n\bar{p}\bar{q}}{(n - 1)\sigma_p\sigma_q}$$

- 其中n是元组的数目，而p和q是各属性的具体值， σ_p 和 σ_q 是各自的标准偏差

姓名	病人数目
长沙	30000
武汉	50000
广州	80000
...	...

姓名	小偷数目
长沙	2000
武汉	3500
广州	6000
...	...

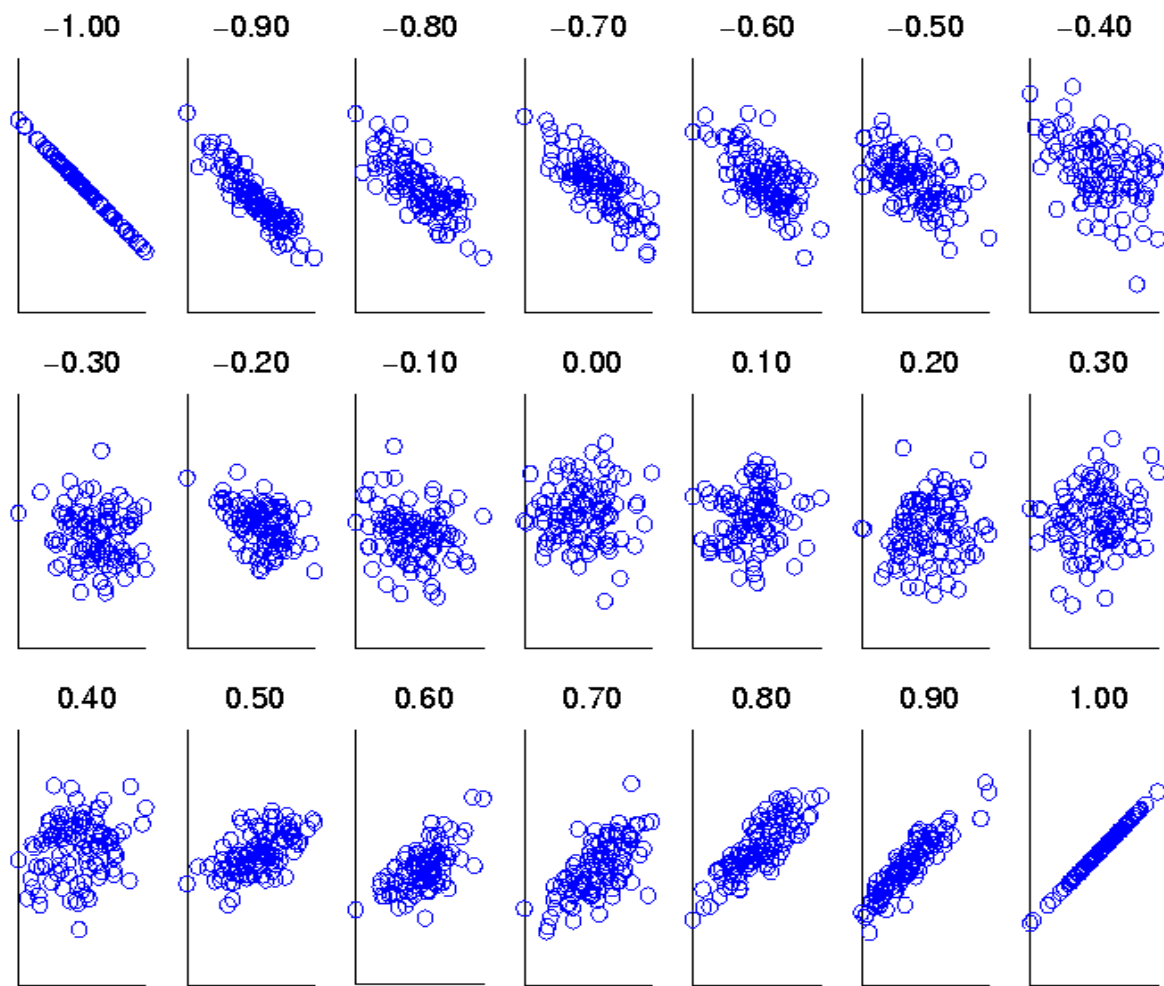
相关分析

$$r_{p,q} = \frac{\sum(p - \bar{p})(q - \bar{q})}{(n - 1)\sigma_p\sigma_q} = \frac{\sum(pq) - n\bar{p}\bar{q}}{(n - 1)\sigma_p\sigma_q}$$

- 当 $r > 0$ 时，表示两变量正相关， $r < 0$ 时，两变量为负相关。
- 当 $|r| = 1$ 时，表示两变量为完全线性相关，即为函数关系。
- 当 $r = 0$ 时，表示两变量间无线性相关关系。
- 当 $0 < |r| < 1$ 时，表示两变量存在一定程度的线性相关。且 $|r|$ 越接近1，两变量间线性关系越密切； $|r|$ 越接近于0，表示两变量的线性相关越弱。
- 一般可按三级划分： $|r| < 0.4$ 为低度线性相关； $0.4 \leq |r| < 0.7$ 为显著性相关； $0.7 \leq |r| < 1$ 为高度线性相关。

相关分析 —— 视觉评估相关

- 散点图显示的相似性，从-1到1



协方差

- 协方差

$$Cov(p, q) = E((p - \bar{p})(q - \bar{q})) = \frac{\sum_{i=1}^n (p_i - \bar{p})(q_i - \bar{q})}{n}$$

$$r_{p,q} = \frac{Cov(p, q)}{\sigma_p \sigma_q}$$

- 其中n是元组的数目， p和q是各自属性的具体值， σ_p 和 σ_q 是各自的标准差。
 - 正相关： $COV(p, q) > 0$
 - 负相关： $COV(p, q) < 0$
 - 不相关： $COVP(p, q) = 0$
- 可具有某些对随机变量的协方差为0， 但不是独立的。一些额外的假设（例如，数据是否服从多元正态分布）做了协方差为0意味着独立。

协方差 —— 举例

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

- 它可以简化计算

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- 假设两只股票A和B具有在1个星期的以下值：

(2, 5) , (3, 8) , (5, 10) , (4, 11) , (6, 14)

- 问题：如果股票都受到同行业的趋势，他们的价格一起上升或下降？
- $E(A) = (2+3+5+4+6) / 5 = 20/5 = 4$
- $E(B) = (5+8+10+11+14) / 5 = 48/5 = 9.6$
- $COV(A, B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$
- 结论：A和B在一起上升，因为 $Cov(A, B) > 0$ 。

目录

01 数据清洗

02 数据集成

03 数据规约

04 数据转换

数据规约策略

- 为什么数据规约 (**data reduction**) ?
 - 由于数据仓库可以存储TB的数据，因此在一个完整的数据集上运行时，复杂的数据分析可能需要一个很长的时间
- 降维
- 降数据
- 数据压缩

数据规约策略

- 为什么数据规约 (data reduction) ?
 - 由于数据仓库可以存储TB的数据，因此在一个完整的数据集上运行时，复杂的数据分析可能需要一个很长的时间
- 降维

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...

数据规约策略

- 为什么数据规约 (data reduction) ?
 - 由于数据仓库可以存储TB的数据，因此在一个完整的数据集上运行时，复杂的数据分析可能需要一个很长的时间
- 降维

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...

数据规约策略 —— 为什么降维

- 原因

- 随着维数的增加，数据变得越来越**稀疏**

- 下例：随着维度增加，数据被绝大数N填充，而实际上我们更加关注生病的数据P或者Y

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

数据规约策略 —— 为什么降维

- 原因
 - 子空间的可能的组合将成倍增长
 - 基于规则的分类方法，建立的规则将组合成倍增长
 - 根据化验测试判定是否咳嗽

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

数据规约策略 —— 为什么降维

- 原因

- 线性（或非线性）机器学习方法，主要需要**学习各个特征的权值参数**。特征越多，需要学习的参数越多，则模型越复杂



- **机器学习训练集原则**：模型越复杂，需要更多的训练集来学习模型参数，否则模型将欠拟合。
- 因此，如果数据集维度很高，而训练集数目很少，在使用复杂的机器学习模型的时候，**首选先降维**。

数据规约策略 —— 为什么降维

- **总结：需要降维的场景**
 - 数据稀疏，维度高
 - 高维数据采用基于规则的分类方法
 - 采用复杂模型，但是训练集数目较少
 - 需要可视化

降维典型方法 —— PCA主成分分析法

- PCA主成分分析法核心idea
 - 数据中很多属性之间可能存在这样或那样的相关性
 - 能不能找到一个方法，将多个相关性的属性组合仅仅形成一个属性？

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...

降维典型方法 —— PCA主成分分析法

- PCA主成分分析法核心idea

- 数据中很多属性之间可能存在这样或那样的相关性
- 能不能找到一个方法，将多个相关性的属性组合仅仅形成一个属性？
- 主成分分析就是设法将原来众多具有一定相关性的属性（比如 p 个属性），重新组合成一组相互无关的综合属性来代替原来属性。通常数学上的处理就是将原来 p 个属性作线性组合，作为新的综合属性

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...

降维典型方法 —— PCA主成分分析法

- PCA主成分分析法核心idea
 - 数据中很多属性之间可能存在这样或那样的相关性
 - 能不能找到一个方法，将多个相关性的属性组合仅仅形成一个属性？
 - 主成分分析就是设法将原来众多具有一定相关性的属性（比如 p 个属性），重新组合成一组相互无关的综合属性来代替原来属性。通常数学上的处理就是将原来 p 个属性作线性组合，作为新的综合属性

$$z_1 = 0.7x_1 + 0.76x_2 + 0.68x_3$$

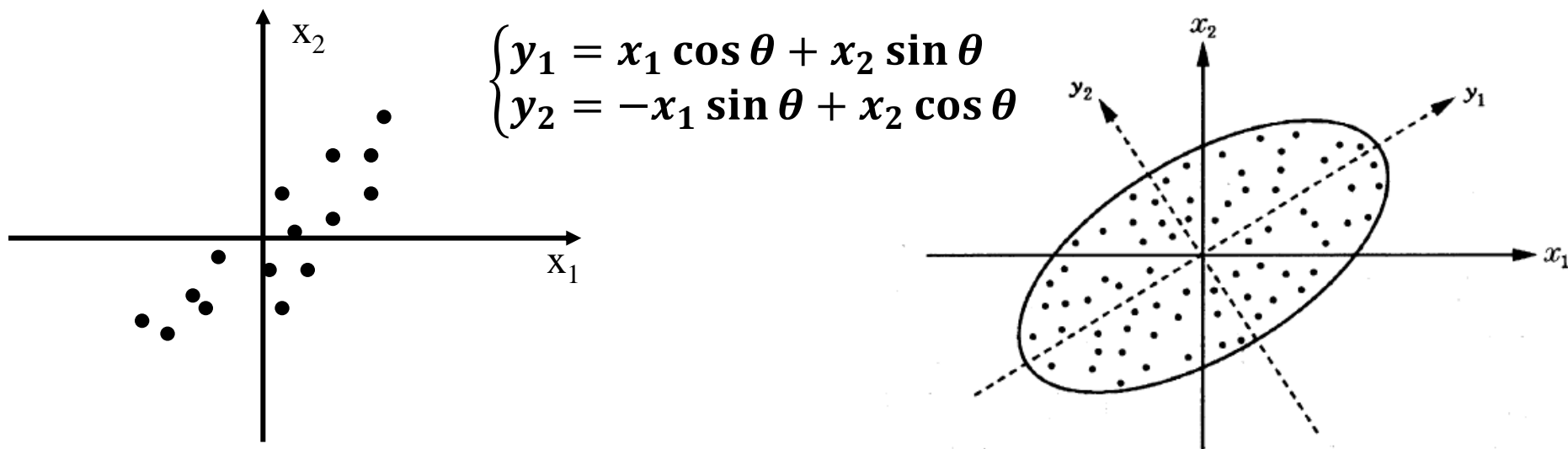
学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...

降维典型方法 —— 主成分计算

定义：记 x_1, x_2, \dots, x_p 为原变量指标, z_1, z_2, \dots, z_m ($m \leq p$)
为新变量指标

$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \dots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \dots + l_{2p}x_p \\ \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \dots + l_{mp}x_p \end{cases}$$

降维典型方法 —— 主成分几何意义



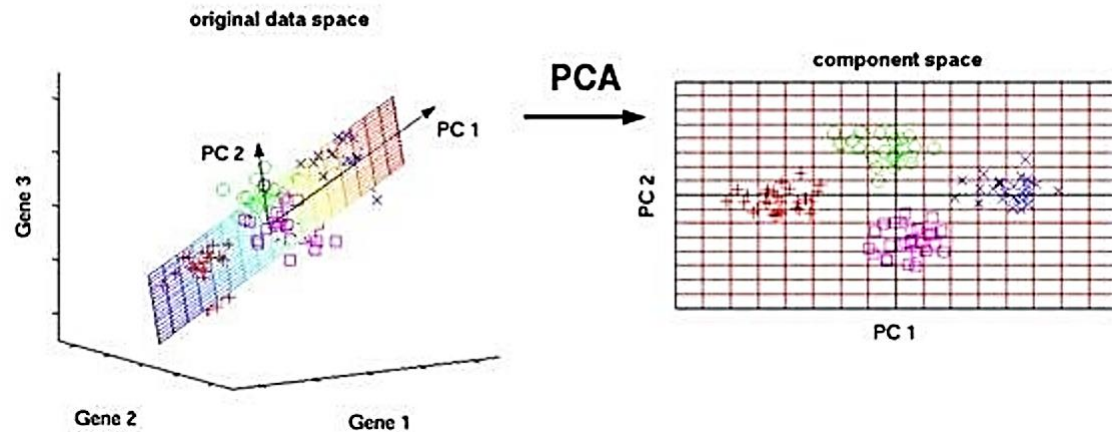
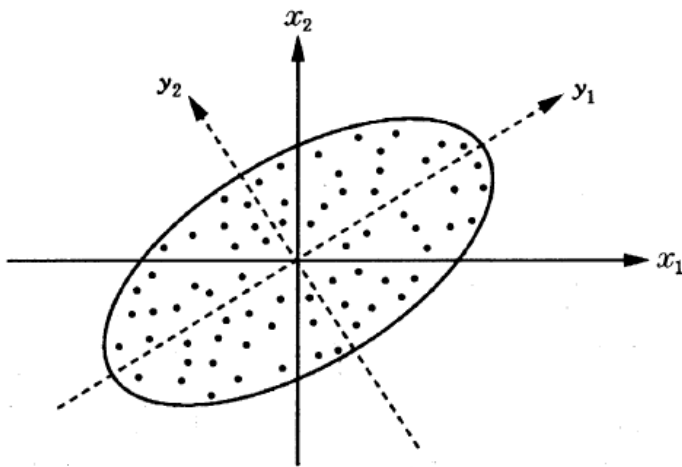
- **线性变换等价于坐标旋转**

变换的目的是为了使得 n 个样本点在 y_1 轴方向上的离散程度最大，既 y_1 的方差达最大。说明变量 y_1 代表了原始数据的绝大部分信息，对 y_2 忽略也无损大局，即由两个指标压缩成一个指标。

- **主成分分析几何意义：寻找主轴**

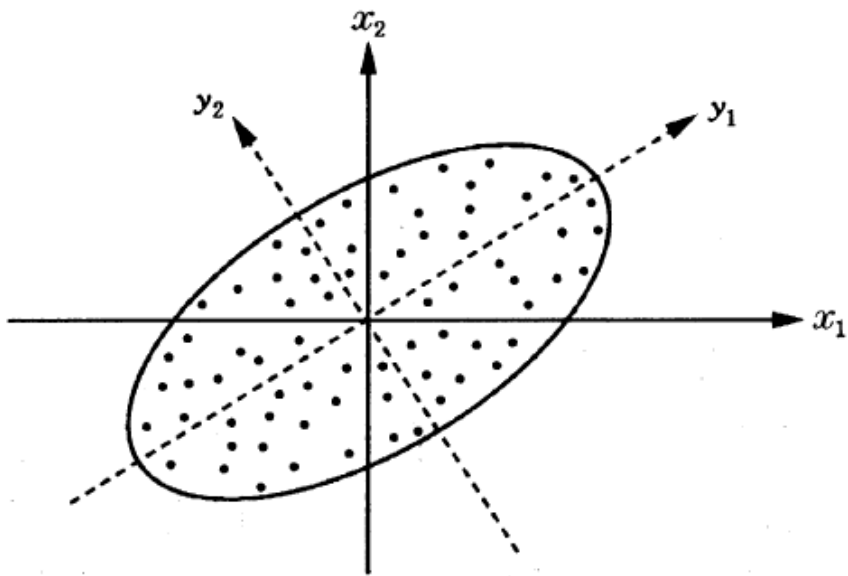
降维典型方法 —— 主成分几何意义

- 正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主成分。
- 选择越少的主成分，降维就越好。什么是标准呢？那就是这些被选的主成分所代表的**主轴的长度之和占了主轴长度总和的大部分**。



降维典型方法 —— 主成分几何意义

- 从几何上看，**找主成分的问题**，就是找出P维空间中椭球体的**主轴问题**
- 从数学上可以证明，它们分别是**协相关矩阵的前m个最大特征值所对应的特征向量**



$$\begin{cases} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{cases}$$

降维典型方法 —— 主成分计算步骤

(一) 计算相关系数矩阵

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

r_{ij} ($i, j=1, 2, \dots, p$) 为原变量 x_i 与 x_j 的相关系数, $r_{ij}=r_{ji}$,

其计算公式为:

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}}$$

降维典型方法 —— 主成分计算步骤

(二) 计算特征值与特征向量:

- ① 解特征方程 $|\lambda I - R| = 0$, 常用雅可比法 (Jacobi) 求出特征值, 并使其按大小顺序排列 $\lambda_1 \geq \lambda_2 \geq \cdots, \geq \lambda_p \geq 0$;
- ② 分别求出对应于特征值 λ_i 的特征向量 $e_i (i = 1, 2, \cdots, p)$, 要求 $\|e_i\|=1$, 即 $\sum_{j=1}^p e_{ij}^2 = 1$, 其中 e_{ij} 表示向量 e_i 的第j个分量。

降维典型方法 —— 主成分计算步骤

③ 计算主成分贡献率及累计贡献率

▲ 贡献率: $f_i = \lambda_i / \sum_{i=1}^p \lambda_i$

▲ 累计贡献率: $\alpha_k = \sum_{i=1}^k f_i$

一般取累计贡献率达85—95%的特征值 $\lambda_1, \lambda_2, \dots, \lambda_m$ 所对应的第一、第二、...、第m ($m \leq p$) 个主成分。

主成分	特征值	贡献率(%)	累积贡献率(%)
z_1	4.661	51.791	51.791
z_2	2.089	23.216	75.007
z_3	1.043	11.589	86.596
z_4	0.507	5.638	92.234
z_5	0.315	3.502	95.736
z_6	0.193	2.14	97.876
z_7	0.114	1.271	99.147
z_8	0.0453	0.504	99.65
z_9	0.0315	0.35	100

降维典型方法 —— 主成分计算步骤

④ 计算主成分值

前 k 个主成分值
$$\begin{aligned} z &= (Xe_1, Xe_2, \dots, Xe_k) \\ &= (z_1, z_2, \dots, z_k) \end{aligned}$$

$$\begin{cases} z_1 = e_{11}x_1 + e_{12}x_2 + \dots + e_{1p}x_p \\ z_2 = e_{21}x_1 + e_{22}x_2 + \dots + e_{2p}x_p \\ \vdots \\ z_k = e_{k1}x_1 + e_{k2}x_2 + \dots + e_{kp}x_p \end{cases}$$

降维典型方法 —— 主成分计算示例

- 某农业生态经济系统做主成分分析

样本序号	X ₁ 人口密度 (人/km ²)	X ₂ 人均耕地 面积 (ha)	X ₃ 森林 覆盖率(%)	X ₄ 农民人均 纯收入 (元/人)	X ₅ 人均粮食 产量 (kg/人)	X ₆ 经济作物 占农作物 播面比例 (%)	X ₇ 耕地占 土地面积 比率 (%)	X ₈ 果园与 林地面积 之比 (%)	X ₉ 灌溉田占 耕地面积 之比 (%)
1	363.912	0.352	16.101	192.11	295.34	26.724	18.492	2.231	26.262
2	141.503	1.684	24.301	1752.35	452.26	32.314	14.464	1.455	27.066
3	100.695	1.067	65.601	1181.54	270.12	18.266	0.162	7.474	12.489
4	143.739	1.336	33.205	1436.12	354.26	17.486	11.805	1.892	17.534
5	131.412	1.623	16.607	1405.09	586.59	40.683	14.401	0.303	22.932
6	68.337	2.032	76.204	1540.29	216.39	8.128	4.065	0.011	4.861
7	95.416	0.801	71.106	926.35	291.52	8.135	4.063	0.012	4.862
8	62.901	1.652	73.307	1501.24	225.25	18.352	2.645	0.034	3.201
9	86.624	0.841	68.904	897.36	196.37	16.861	5.176	0.055	6.167
10	91.394	0.812	66.502	911.24	226.51	18.279	5.643	0.076	4.477

降维典型方法 —— 主成分计算示例

- 某农业生态经济系统做主成分分析

样本序号	X ₁ 人口密度 (人/km ²)	X ₂ 人均耕地 面积 (ha)	X ₃ 森林 覆盖率(%)	X ₄ 农民人均 纯收入 (元/人)	X ₅ 人均粮食 产量 (kg/人)	X ₆ 经济作物 占农作物 播面比例 (%)	X ₇ 耕地占 土地面积 比率 (%)	X ₈ 果园与 林地面积 之比 (%)	X ₉ 灌溉田占 耕地面积 之比 (%)
11	76.912	0.858	50.302	103.52	217.09	19.793	4.881	0.001	6.165
12	51.274	1.041	64.609	968.33	181.38	4.005	4.066	0.015	5.402
13	68.831	0.836	62.804	957.14	194.04	9.11	4.484	0.002	5.79
14	77.301	0.623	60.102	824.37	188.09	19.409	5.721	5.055	8.413
15	76.948	1.022	68.001	1255.42	211.55	11.102	3.133	0.01	3.425
16	99.265	0.654	60.702	1251.03	220.91	4.383	4.615	0.011	5.593
17	118.505	0.661	63.304	1246.47	242.16	10.706	6.053	0.154	8.701
18	141.473	0.737	54.206	814.21	193.46	11.419	6.442	0.012	12.945
19	137.761	0.598	55.901	1124.05	228.44	9.521	7.881	0.069	12.654
20	117.612	1.245	54.503	805.67	175.23	18.106	5.789	0.048	8.461
21	122.781	0.731	49.102	1313.11	236.29	26.724	7.162	0.092	10.078

降维典型方法 —— 主成分计算示例

步骤如下：

① 计算相关系数矩阵

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₇	x ₈	x ₉
x ₁	1	-0.327	-0.714	-0.336	0.309	0.408	0.79	0.156	0.744
x ₂	-0.327	1	-0.035	0.644	0.42	0.255	0.009	-0.078	0.094
x ₃	-0.714	-0.035	1	0.07	-0.74	-0.755	-0.93	-0.109	-0.924
x ₄	-0.336	0.644	0.07	1	0.383	0.069	-0.046	-0.031	0.073
x ₅	0.309	0.42	-0.74	0.383	1	0.734	0.672	0.098	0.747
x ₆	0.408	0.255	-0.755	0.069	0.734	1	0.658	0.222	0.707
x ₇	0.79	0.009	-0.93	-0.046	0.672	0.658	1	-0.03	0.89
x ₈	0.156	-0.078	-0.109	-0.031	0.098	0.222	-0.03	1	0.29
x ₉	0.744	0.094	-0.924	0.073	0.747	0.707	0.89	0.29	1

降维典型方法 —— 主成分计算示例

步骤如下：

② 由相关系数矩阵计算特征值，以及各个主成分的贡献率与累计贡献率。第一，第二，第三主成分的累计贡献率已高达86.596%（大于85%），故只需要求出第一、第二、第三主成分 z_1 ， z_2 ， z_3 即可

主成分	特征值	贡献率(%)	累积贡献率(%)
z_1	4.661	51.791	51.791
z_2	2.089	23.216	75.007
z_3	1.043	11.589	86.596
z_4	0.507	5.638	92.234
z_5	0.315	3.502	95.736
z_6	0.193	2.14	97.876
z_7	0.114	1.271	99.147
z_8	0.0453	0.504	99.65
z_9	0.0315	0.35	100

降维典型方法 —— 主成分计算示例

样本序号	x_1 人口密度 (人/km ²)	x_2 人均耕地 面积 (ha)	x_3 森林 覆盖率(%)	x_4 农民人均 纯收入 (元/人)	x_5 人均粮食 产量 (kg/人)	x_6 经济作物 占农作物 播面比例 (%)	x_7 耕地占 土地面积 比率 (%)	x_8 果园与 林地面积 之比 (%)	x_9 灌溉田占 耕地面积 之比 (%)
1	363.912	0.352	16.101	192.11	295.34	26.724	18.492	2.231	26.262
2	141.503	1.684	24.301	1752.35	452.26	32.314	14.464	1.455	27.066
3	100.695	1.067	65.601	1181.54	270.12	18.266	0.162	7.474	12.489
4	143.739	1.336	33.205	1436.12	354.26	17.486	11.805	1.892	17.534
5	131.412	1.623	16.607	1405.09	586.59	40.683	14.401	0.303	22.932

	z_1	z_2	z_3
x_1	0.739	-0.532	-0.0061
x_2	0.123	0.887	-0.0028
x_3	-0.964	0.0096	0.0095
x_4	0.0042	0.868	0.0037
x_5	0.813	0.444	-0.0011
x_6	0.819	0.179	0.125
x_7	0.933	-0.133	-0.251

降维典型方法 —— 主成分计算示例

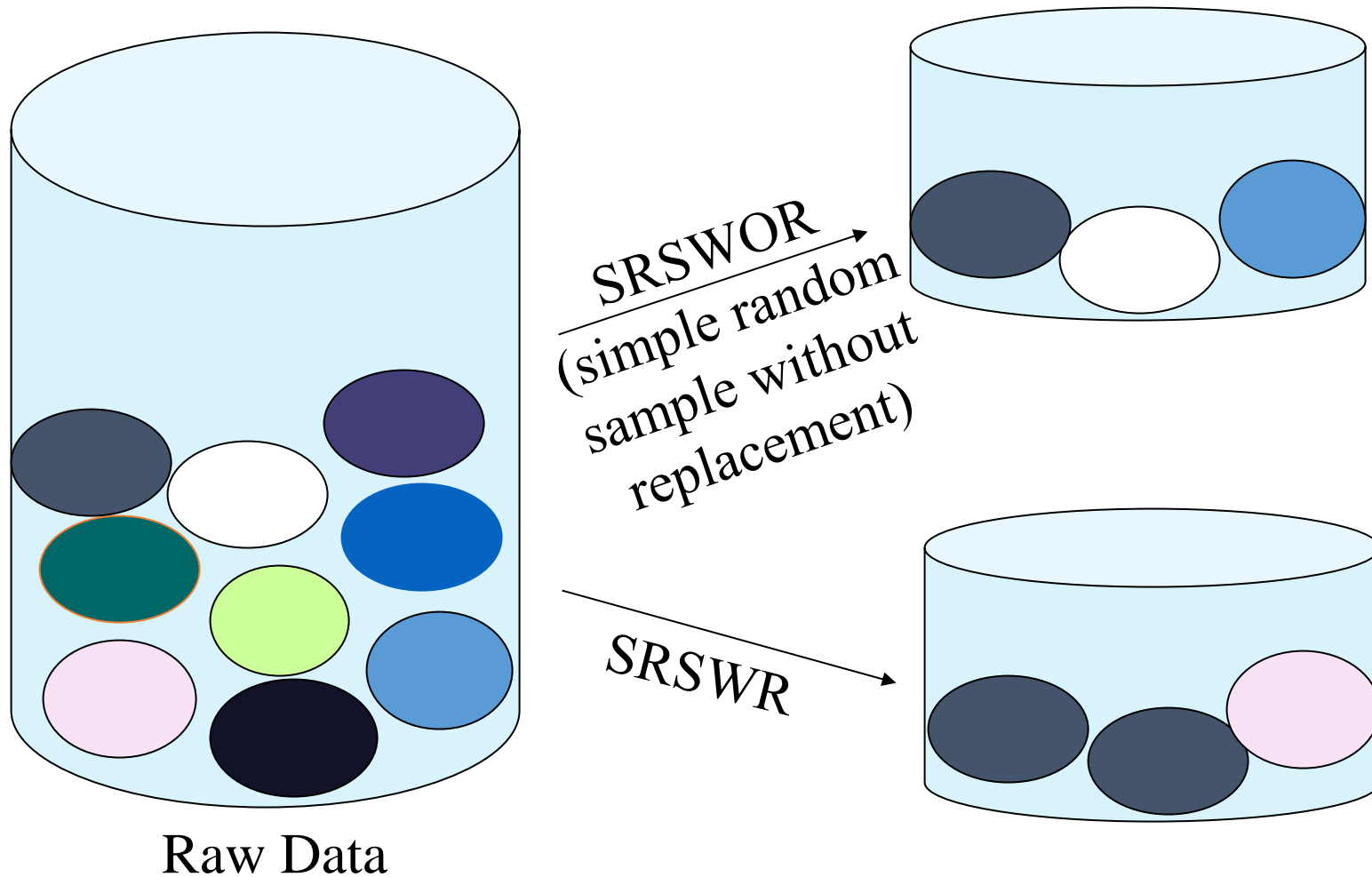
分析：

- ① 第一主成分 z_1 与 x_1 , x_5 , x_6 , x_7 , x_9 呈显出较强的正相关, 与 x_3 呈显出较强的负相关, 而这几个变量则综合反映了生态经济结构状况, 因此可以认为第一主成分 z_1 是生态经济结构的代表。
- ② 第二主成分 z_2 与 x_2 , x_4 , x_5 呈显出较强的正相关, 与 x_1 呈显出较强的负相关, 其中, 除了 x_1 为人口总数外, x_2 , x_4 , x_5 都反映了人均占有资源量的情况, 因此可以认为第二主成分 z_2 代表了人均资源量。
- ③ 第三主成分 z_3 , 与 x_8 呈显出的正相关程度最高, 其次是 x_6 , 而与 x_7 呈负相关, 因此可以认为第三主成分在一定程度上代表了农业经济结构。

数据规约策略

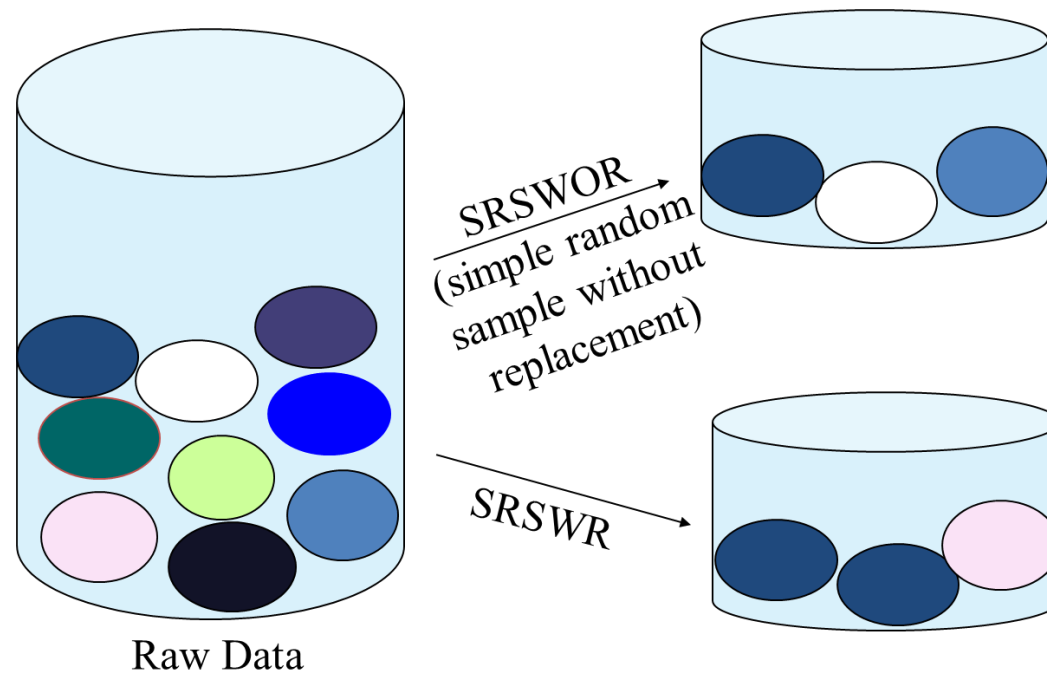
- 为什么数据规约 (data reduction) ?
 - 由于数据仓库可以存储TB的数据，因此在一个完整的数据集上运行时，复杂的数据分析可能需要一个很长的时间
- 降维
- 降数据
- 数据压缩

降数据典型方法 —— 抽样法



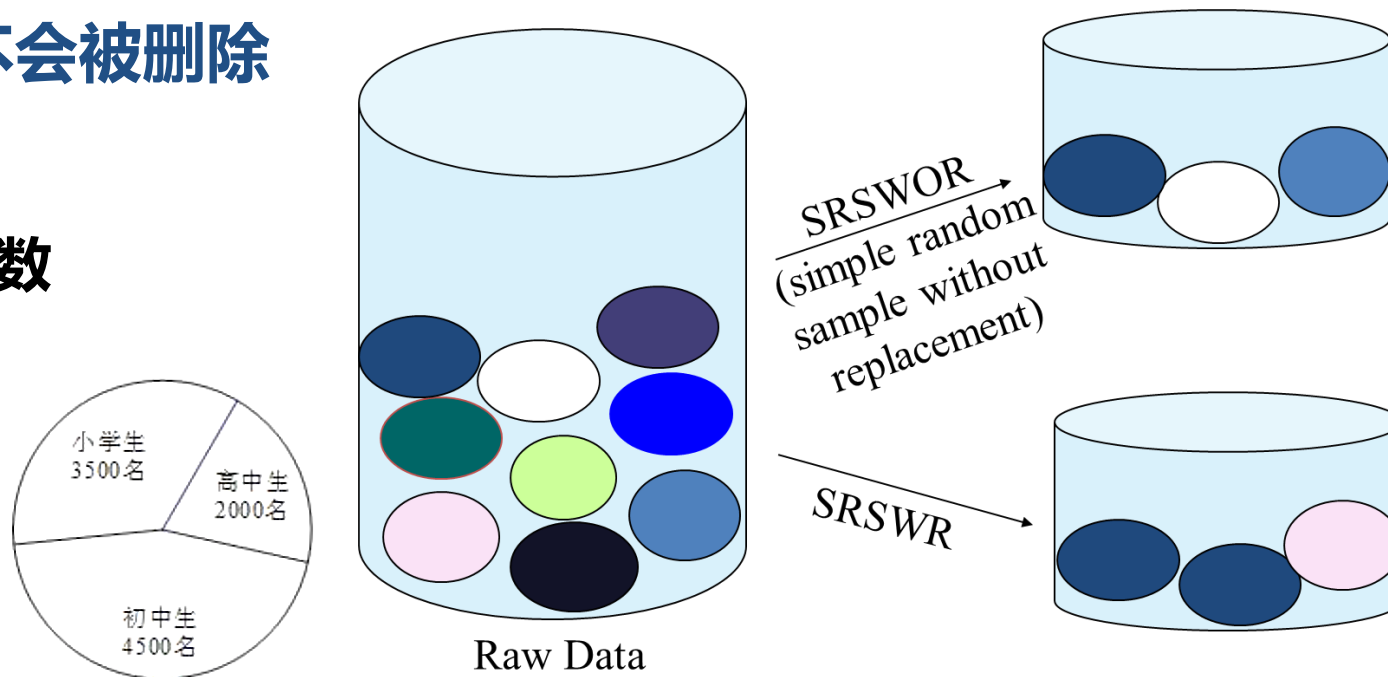
抽样类型

- 简单随机抽样(Simple Random Sampling)
 - 相等的概率选择
 - 不放回抽样(Sampling without replacement)
 - 一旦对象被选中，则将其删除
 - 有放回抽样(Sampling with replacement)
 - 选择对象不会被删除



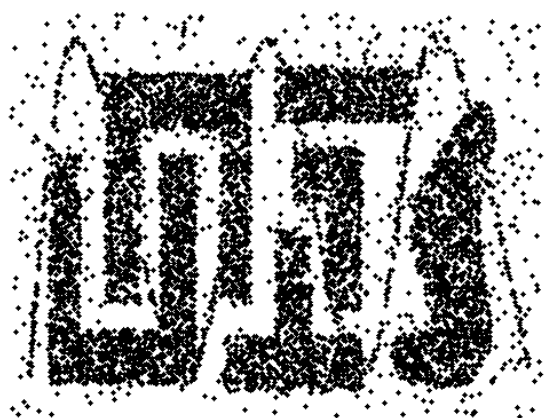
抽样类型

- 简单随机抽样(Simple Random Sampling)
 - 相等的概率选择
 - 不放回抽样(Sampling without replacement)
 - 一旦对象被选中，则将其删除
 - 有放回抽样(Sampling with replacement)
 - 选择对象不会被删除
- 分组抽样
 - 每组抽相近个数
 - 用于偏斜数据

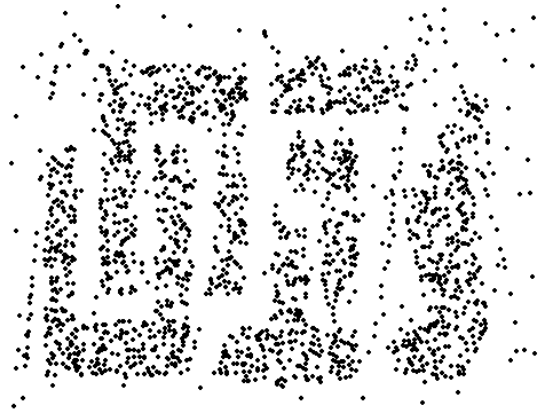


样本大小对数据质量的影响

- 从8000个点分别抽2000和500个点
 - 2000个点的样本保留了数据集的大部分结构
 - 500个点的样本丢失了许多结构



8000 points



2000 Points

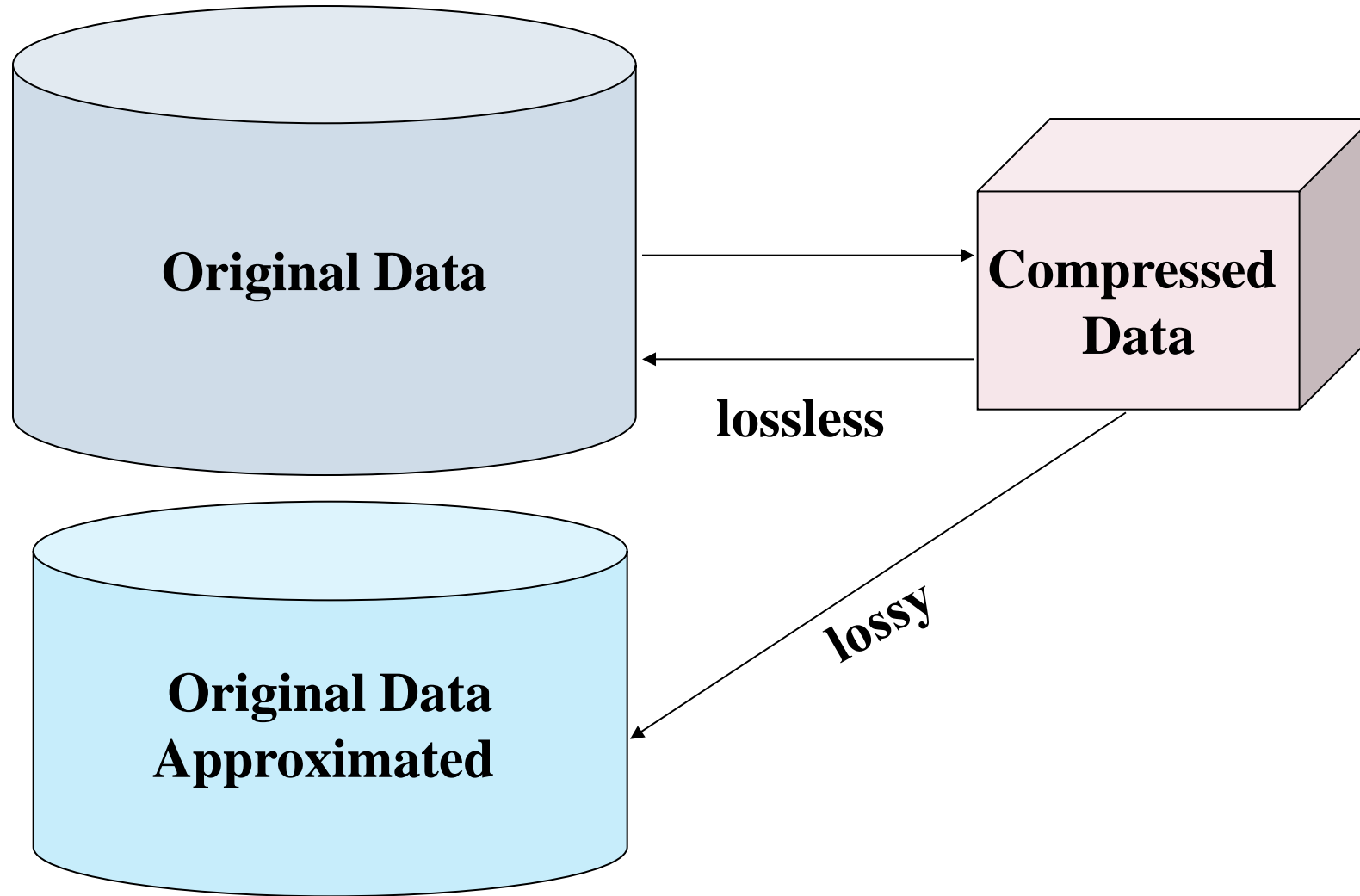


500 Points

数据规约策略

- 为什么数据规约 (data reduction) ?
 - 由于数据仓库可以存储TB的数据，因此在一个完整的数据集上运行时，复杂的数据分析可能需要一个很长的时间
- 降维
- 降数据
- 数据压缩

数据压缩



数据规约策略小结

- 为什么数据规约 (data reduction) ?
 - 由于数据仓库可以存储TB的数据，因此在一个完整的数据集上运行时，复杂的数据分析可能需要一个很长的时间

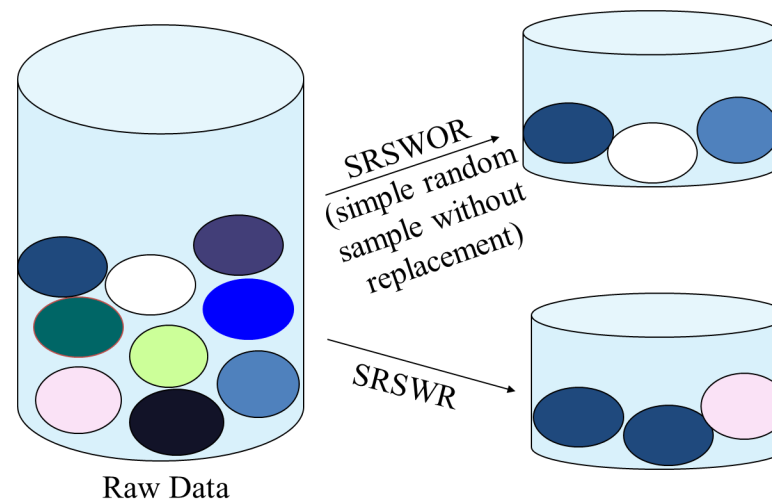
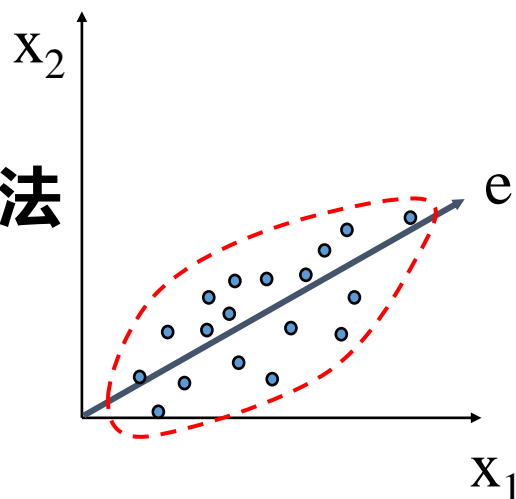
- 降维

- PCA主成分法

- 降数据

- 抽样法

- 数据压缩



目录

01 数据清洗

02 数据集成

03 数据规约

04 数据转换

数据转换

- **函数映射：** 给定的属性值更换了一个新的表示方法，每个旧值与新的值可以被识别
- **方法**
 - **规范化：** 按比例缩放到一个具体区间
 - 最小 - 最大规范化
 - z-得分正常化
 - 小数定标规范化
 - **离散化**

数据转换 —— 规范化方法

- 最小-最大规范化

$$v' = \frac{v - \min A}{\max A - \min A} (\text{new_max}A - \text{new_min}A) + \text{new_min}A$$

- v即需要规范的数据

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...

填空题

一组数据的最小值为12,000，最大值为98,000，将数据规范到[0,1]，则73,000规范化的值为：【0.8488】

$$v' = \frac{v - \min A}{\max A - \min A} (\text{new_max}A - \text{new_min}A) + \text{new_min}A$$

数据转换 —— 规范化方法

- z-分数规范化

$$v' = \frac{v - \text{均值}_A}{\text{标准差}_A}$$

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...

填空题

一组数据的均值为54,000，标准差为16,000，则 73,000规范化的值为：

[1.1875]

$$v' = \frac{v - \text{均值}_A}{\text{标准差}_A}$$

数据转换 —— 规范化方法

- 小数定标：移动属性A的小数点位置 (移动位数依赖于属性A的最大值)


$$v' = \frac{v}{10^j} \quad j \text{ 为使 } \text{Max}(|v|) < 1 \text{ 的最小整数}$$

一组数据的最小值为12,000，最大值为98,000， j 值取5

数据转换和离散化 —— 离散化方法

- 为什么需要离散化
 - 部分数据挖掘算法只使用于离散数据

id	收入
1	115
2	110
3	70
4	112
5	90
6	60
7	118
8	85
9	75
10	80



id	收入
1	高
2	高
3	低
4	高
5	中
6	低
7	高
8	中
9	低
10	中

数据转换和离散化 —— 离散化方法

- 非监督离散

- 等宽法

- 根据属性的值域来划分，使每个区间的宽度相等

id	收入
1	115
2	110
3	70
4	112
5	90
6	60
7	118
8	85
9	75
10	80

等宽划分

区间	离散类别
[60,80)	低
[80,100)	中
[100,120)	高



id	收入
1	高
2	高
3	低
4	高
5	中
6	低
7	高
8	中
9	低
10	中

数据转换——离散化方法

- 非监督离散
 - 等宽法
 - 根据属性的值域来划分，使每个区间的宽度相等
 - 等频法
 - 根据取值出现的频数来划分，将属性的值域划分成个小区间，并且要求落在每个区间的样本数目相等

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

数据转换——离散化方法

- 非监督离散化法

- 等宽法

- 根据属性的值域来划分，使每个区间的宽度相等

- 等频法

- 根据取值出现的频数来划分，将属性的值域划分成个小区间，并且要求落在每个区间的样本数目相等

- 聚类

- 利用聚类将数据划分到不同的离散类别

