

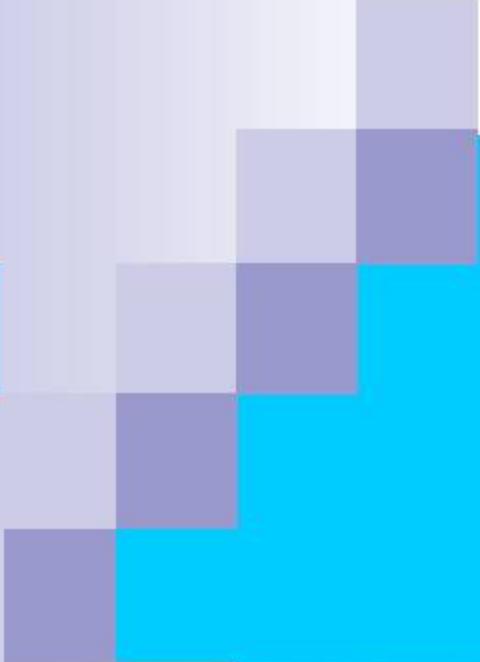


数据挖掘

Data Mining

第二课 数据预处理

■ 主讲人：丁兆云



数据挖掘

Data Mining

第二课 数据预处理

- 主讲人：丁兆云



内容提纲

- 2.1 数据质量
- 2.2 数据预处理
- 2.3 特征构造



2.1 数据质量



数据质量

- 被广泛接受的数据质量测量标准

- 准确性
- 完整性
- 一致性
- 合时性
- 可信度
- 解释性



2.2 数据预处理



2.2 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
- 数据缩减
 - 降维
 - Numerosity reduction
 - 数据压缩
- 数据转换和数据离散化
 - 正常化
 - 生成概念层次结构



2.2.1 数据清洗

- 在现实世界中的数据是“脏”的：
 - 不完整的：缺少属性值，缺乏某些属性值，或只包含总数据
 - 例如，职业=“ ”（丢失的数据）
 - 含嘈杂的噪音，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”



2.2.1-1如何处理丢失数据？

	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. 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, Mrs. 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, Mr. 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, Mr. 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, Mr. female		14	1	0	237736	30.0708		C
12	11	1	3	Sandström, female		4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, Mrs. female		58	0	0	113783	26.55	C103	S



2.2.1-1 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，它的效果非常差
- 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))
```



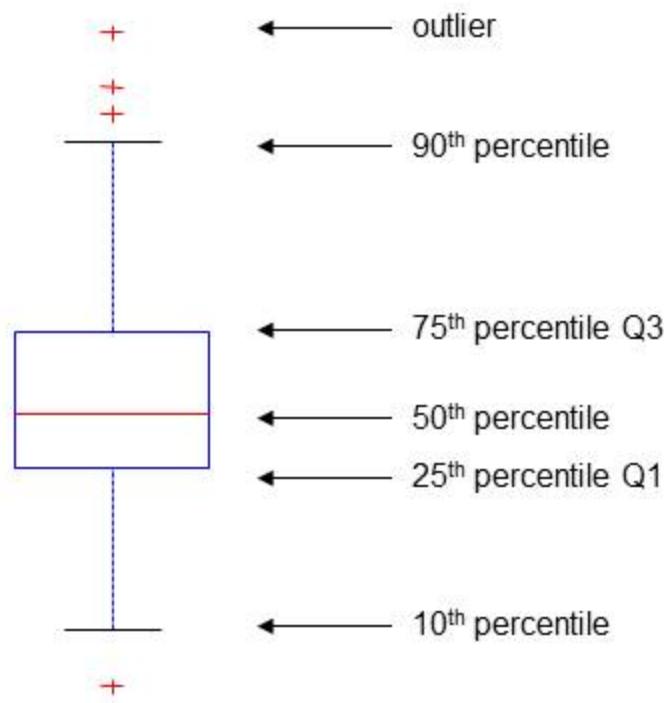
2.2.1-1如何处理丢失数据？

A	B	C	D	E	F	G	H	I	J	K	L
Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3 Braund, Mr. male		22	1	0	A/5 21171	7.25		S
2	2	1	1 Cumings, Mrs. (Catherine)	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3 Heikkinen, Sointu	female	26	0	0	STON/O2	7.925		S
4	4	1	1 Futrelle, Mrs. (Mme. Jeanne)	female	35	1	0	113803	53.1	C123	S
5	5	0	3 Allen, Mr. (William Henry)	male	35	0	0	373450	8.05		S
6	6	0	3 Moran, Mr. (James Joseph)	male	平均值	0	0	330877	8.4583		Q
7	7	0	1 McCarthy, Mrs. (Pamela Elizabeth)	male	54	0	0	17463	51.8625	E46	S
8	8	0	3 Palsson, Master. (Gosta Carlsson)	male	2	3	1	349909	21.075		S
9	9	1	3 Johnson, Mrs. (Mrs. John Bradley)	female	27	0	2	347742	11.1333		S
10	10	1	2 Nasser, Mrs. (Fahima)	female	14	1	0	237736	30.0708		C
11	11	1	3 Sandstrom, Mrs. (Margaret)	female	4	1	1	PP 9549	16.7	G6	S
12	12	1	1 Bonnell, Mrs. (Elizabeth)	female	58	0	0	113783	26.55	C103	S



2.2.1-2如何处理噪声数据？

- 盒状图检测离群数据：删除离群点





2.2.1-3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=“03/07/1997”
 - 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

数据库中某属性缺失值比较多时，数据清理采用的方法

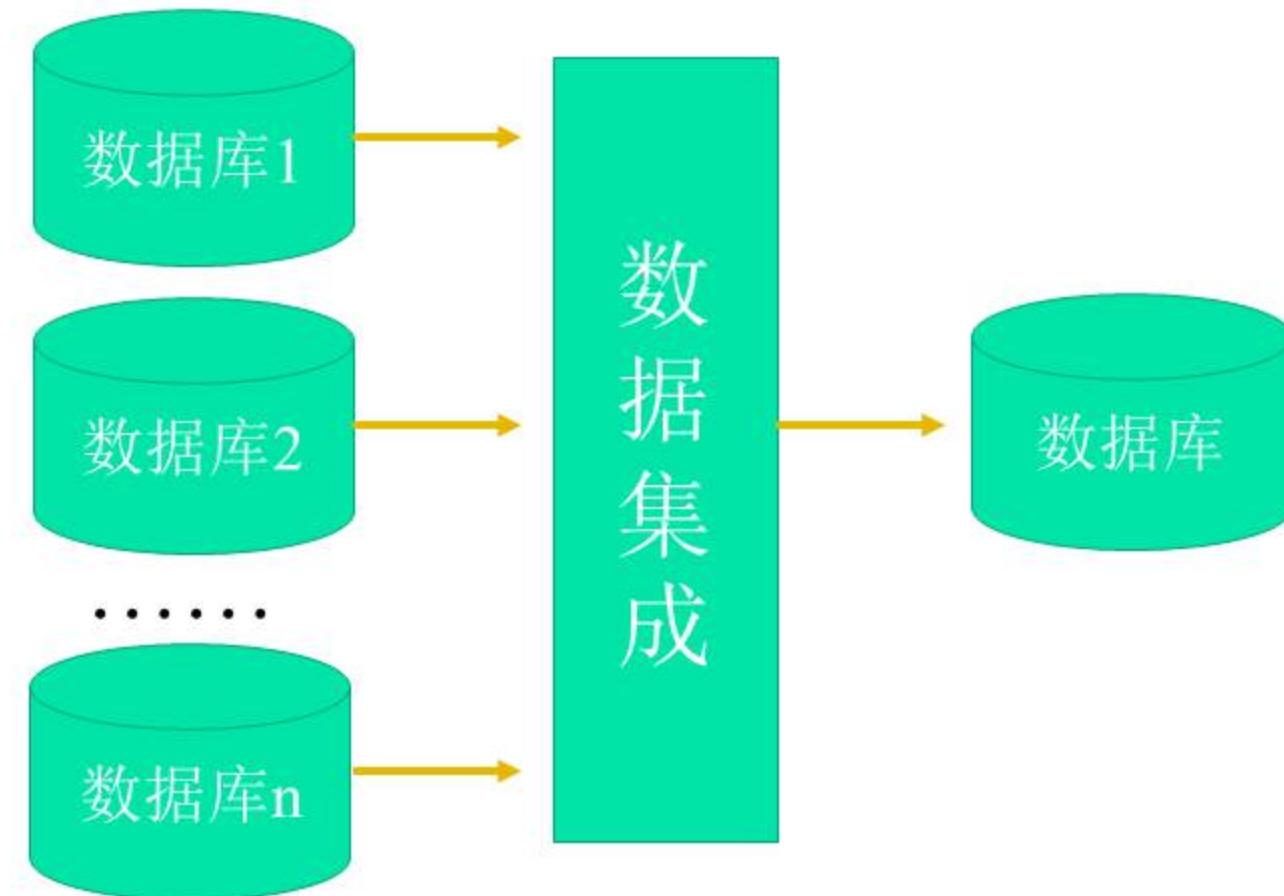
- A 忽略元组
- B 平均值填充
- C 盒状图法

提交



2.2.2 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





2.2.2 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：例如，**A.cust-id ≡ B.cust-#**
 - 整合来自不同来源的元数据

数据库A			数据库B		
cust-id	name	height	cust-id	name	height
1	丁兆云	1.68	1	dzy	5.51
2	张三	1.76	2	zs	5.77

数据集成

id	nameA	heightA	nameB	heightB
1	丁兆云	1.68	dzy	5.51
2	张三	1.76	zs	5.77



2.2.2 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:
 - 识别来自多个数据源的真实世界的实体，例如，**Bill Clinton = William Clinton**

数据库A			数据库B		
cust-id	name	height	cust-id	name	height
1	丁兆云	1.68	1	dzy	5.51
2	张三	1.76	2	zs	5.77

数据集成

id	name	heightA	heightB
1	丁兆云	1.68	5.51
2	张三	1.76	5.77



2.2.2 数据集成-数据冲突检测和解决

■ 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

■ 数据冲突检测和解决

- 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 数据库B

cust-id	name	height
1	丁兆云	1.68
2	张三	1.76

cust-id	name	height
1	dzy	5.51
2	zs	5.77

数据集成

id	name	height
1	丁兆云	1.68
2	张三	1.76



2.2.2 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
 - 1、模式集成：例如，A.cust-id≡B.cust-#
 - 2、实体识别问题：
 - 3、数据冲突检测和解决

数据库A			数据库B		
cust-id	name	height	cust-id	name	height
1	丁兆云	1.68	1	dzy	5.51
2	张三	1.76	2	zs	5.77

数据集成

id	name	height
1	丁兆云	1.68
2	张三	1.76



数据集成需要解决的问题

- A 模式集成
- B 实体识别
- C 数据冲突检测

提交



2.2.3 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，年收入

数据库A			数据库B		
cust-id	name	3000m	cust-id	name	5000m
1	丁兆云	13.24	1	dzy	25.35
2	张三	11.26	2	zs	21.27

数据集成

id	name	run
1	丁兆云	15.24
2	张三	12.14



2.2.3 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，年收入
- 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



2.2.4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	是否看书
张三	1
王五	1
马六	0
....	

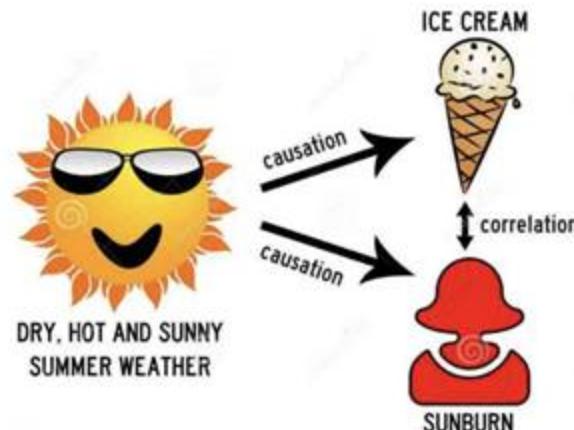


2.2.4 相关分析

■ χ^2 (chi-square) test

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

- 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





2.2.4 χ^2 (chi-square) test 举例

姓名	是否下棋
张三	1
王五	0
马六	0
....	

姓名	是否看书
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

填空题

1分



设置



	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

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

= [填空1]

正常使用填空题需3.0以上版本雨课堂

作答



2.2.4 X² (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



2.2.4 相关性数据集成

姓名	是否下棋
张三	1
王五	0
马六	0
....	

姓名	是否看书
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
....	



2.2.5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
....	



2.2.5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是各自属性的具体值， 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
....	



2.2.5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是各自属性的具体值， 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
....	



2.2.5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是各自属性的具体值， 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
....	



2.2.5相关分析

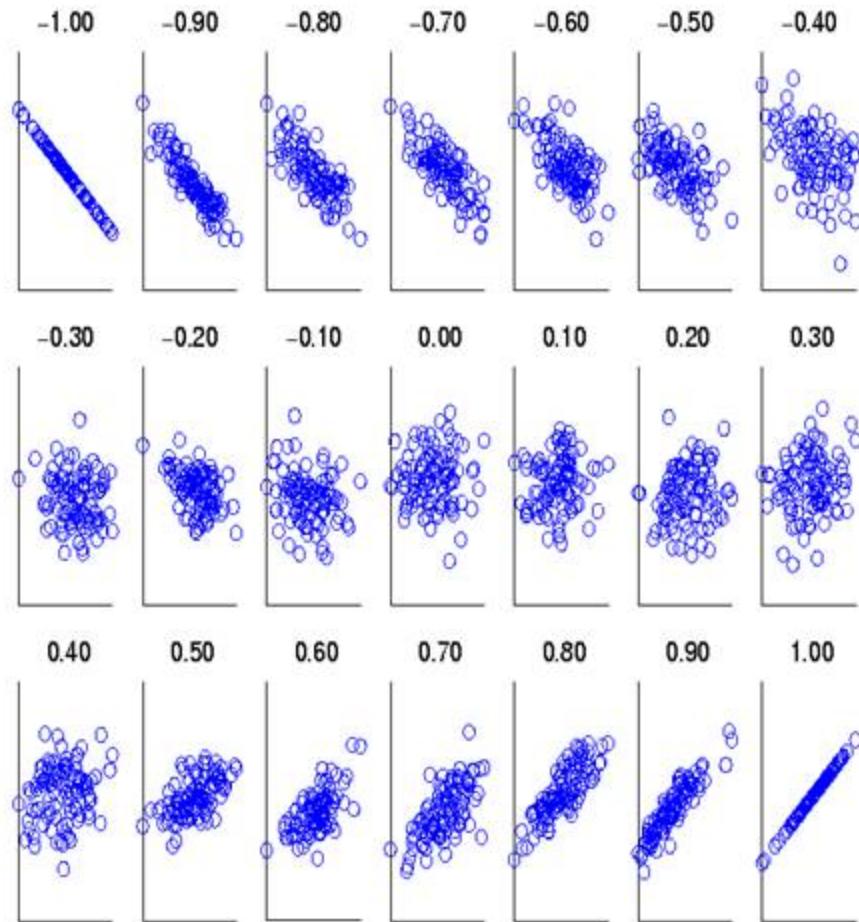
■ 相关系数（也称为皮尔逊相关系数）

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

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



2.2.5相关分析-视觉评估相关



散点图显示的相似性，从-1到1。



2.2.5相关分析-课后加深理解作业

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

- 假设五个同学单杠和俯卧撑的次数的以下值: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)。



2.2.6 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。

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
....	



2.2.6 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: $Cov(p, q) < 0$
 - 独立性: $Cov(p, q) = 0$
- 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



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

- 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，他们的价格协方差等于：[\[填空1\]](#)

正常使用填空题需3.0以上版本雨课堂



2.2.6 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。



2.2.7 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



2.2.7 数据规约策略-降维

- 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
...



2.2.7 数据规约策略-降维

- 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						
2						
3						
4						
5						
6						
7						
8						
9						
...						



2.2.7 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，数据被绝大部分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

- 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



2.2.7 数据规约策略-为什么降维

- 原因

- 类似神经网络的机器学习方法，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

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \cdots + w_dx_d - t)$$

- **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**。



2.2.7 数据规约策略-降维

- 原因

- 类似神经网络的机器学习方法，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

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \cdots + w_dx_d - t)$$

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



2.2.8 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
...



2.2.8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						
2						
3						
4						
5						
6						
7						
8						
9						
...						



2.2.8 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						
2						
3						
4						
5						
6						
7						
8						
9						
...						

理科成绩 文科成绩



2.2.8 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						
2						
3						
4						
5						
6						
7						
8						
9						
...						

理科成绩 文科成绩



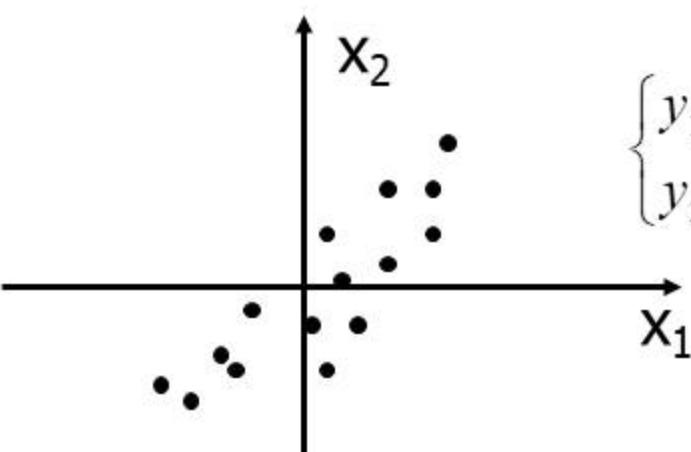
2.2.8 主成分计算

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

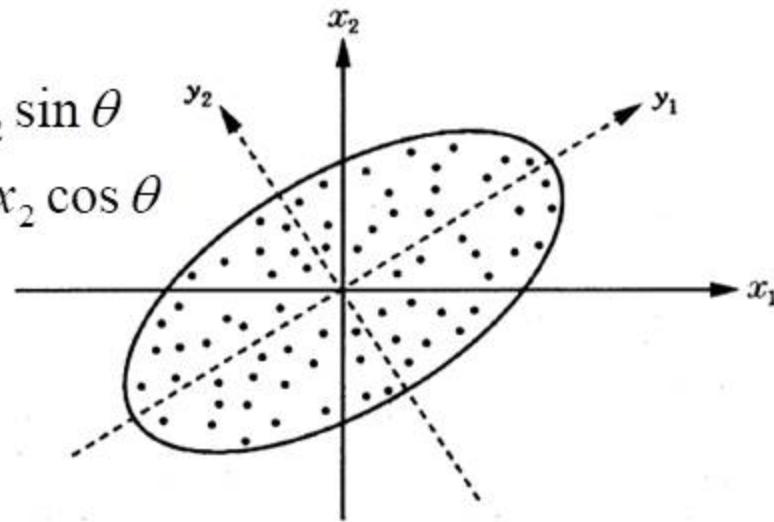
$$\left\{ \begin{array}{l} z_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \vdots \\ z_m = l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{array} \right.$$



2.2.8 主成分几何意义



$$\begin{cases} y_1 = x_1 \cos \theta + x_2 \sin \theta \\ y_2 = -x_1 \sin \theta + x_2 \cos \theta \end{cases}$$



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。

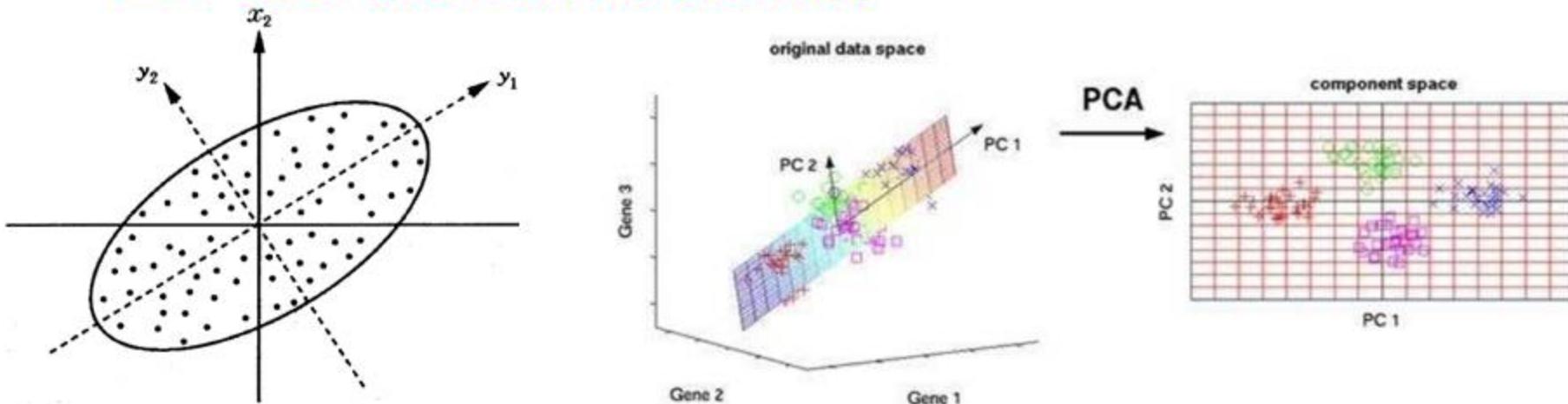
主成分分析几何意义：寻找主轴

zyding@nudt.edu.cn



2.2.8 主成分几何意义

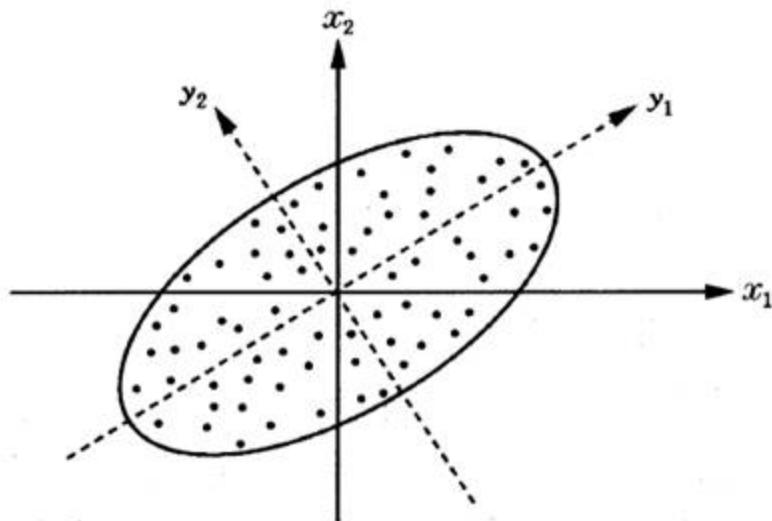
- 正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主成分。
- 选择越少的主成分，降维就越好。什么是标准呢？那就是这些被选的主成分所代表的**主轴的长度之和占了主轴长度总和的大部分**。





2.2.8 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**相关矩阵的m个较大的特征值**所对应的**特征向量**



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



2.2.8 主成分计算步骤

(一) 计算相关系数矩阵

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

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:

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



2.2.8 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 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$;

② 分别求出对应于特征值 λ_i 的特征向量

e_i ($i = 1, 2, \dots, p$) , 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。



2.2.8 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



2.2.8 主成分计算步骤

④ 计算主成分值

前 k 个主成分值

$$\begin{aligned} z &= (Xe_1, Xe_2, \dots, Xe_k) \\ &= (z_1, z_2, \dots, z_k) \end{aligned}$$

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

	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
x_8	0.197	-0.1	0.97
x_9	0.964	-0.0025	0.0092



2.2.8 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 : 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种面比例(%)	x_7 : 耕地占土地面积比率(%)	x_8 : 果园与耕地面积之比	x_9 : 灌溉田占耕地面积之比(%)
1	363.91	0.352	16.101	192.11	295.34	26.724	18.492	2.231	26.262
2	141.5	1.684	24.301	1752.35	452.26	32.314	14.464	1.455	27.066
3	100.7	1.067	65.601	1181.54	270.12	18.266	0.162	7.474	12.489
4	143.74	1.336	33.205	1436.12	354.26	17.486	11.805	1.892	17.534
5	131.41	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
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	118.51	0.661	63.304	1246.47	242.16	10.706	6.053	0.154	8.701
18	141.47	0.737	54.206	814.21	193.46	11.419	6.442	0.012	12.945
19	137.76	0.598	55.901	1124.05	228.44	9.521	7.881	0.069	12.654
20	117.61	1.245	54.503	805.67	175.23	18.106	5.789	0.048	8.461
21	122.78	0.731	49.102	1313.11	236.29	26.724	7.162	0.092	10.078



2.2.8 主成分计算示例

步骤如下：（1）计算相关系数矩阵

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

	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	-0.05	-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



2.2.8 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



2.2.8 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。

	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 作物播种 比例(%)	x_7 : 耕地 占土地面 积比率 (%)	x_8 : 林 园与林 地面积 之比 (%)	x_9 : 滴灌 田占耕 地面积 之比 (%)
1	363.91	0.352	16.101	192.11	295.34	26.724	18.492	2.231	26.262
2	141.5	1.684	24.301	1752.35	452.26	32.314	14.464	1.455	27.066
3	100.7	1.067	65.601	1181.54	270.12	18.266	0.162	7.474	12.489
4	143.74	1.336	33.205	1436.12	354.26	17.486	11.805	1.892	17.534
5	131.41	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
x_8	0.197	-0.1	0.97
x_9	0.964	-0.0025	0.0092



2.2.8 主成分计算示例

分析：③第三主成分z3，与x8呈显出的正相关程度最高，其次是x6，而与x7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。

	x_1 : 人口密度	x_2 : 人均耕地面积	x_3 : 森林覆盖率(%)	x_4 : 农民人均纯收入(元/人)	x_5 : 人均均粮食产量(kg)	x_6 : 经济作物占农作物播种面积比例(%)	x_7 : 耕地占土地面积之比(%)	x_8 : 林地占耕地面积之比(%)	x_9 : 淌溉地面积占耕地面积之比(%)
样本序号	(人/ km^2)	(ha)	(%)	(元/人)	(kg)	(%)	(%)	(%)	(%)
1	363.91	0.352	16.101	192.11	295.34	26.724	18.492	2.231	26.262
2	141.5	1.684	24.301	1752.35	452.26	32.314	14.464	1.455	27.066
3	100.7	1.067	65.601	1181.54	270.12	18.266	0.162	7.474	12.489
4	143.74	1.336	33.205	1436.12	354.26	17.486	11.805	1.892	17.534
5	131.41	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
X_8	0.197	-0.1	0.97
X_9	0.964	-0.0025	0.0092



2.2.8 主成分分析编程实践

<https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#sklearn.decomposition.PCA>

https://blog.csdn.net/lynn_001/article/details/86741284

编程升级：课后同学们参考上述链接写写代码完成实践，不计入平时分，但是这个实践很有用！！！

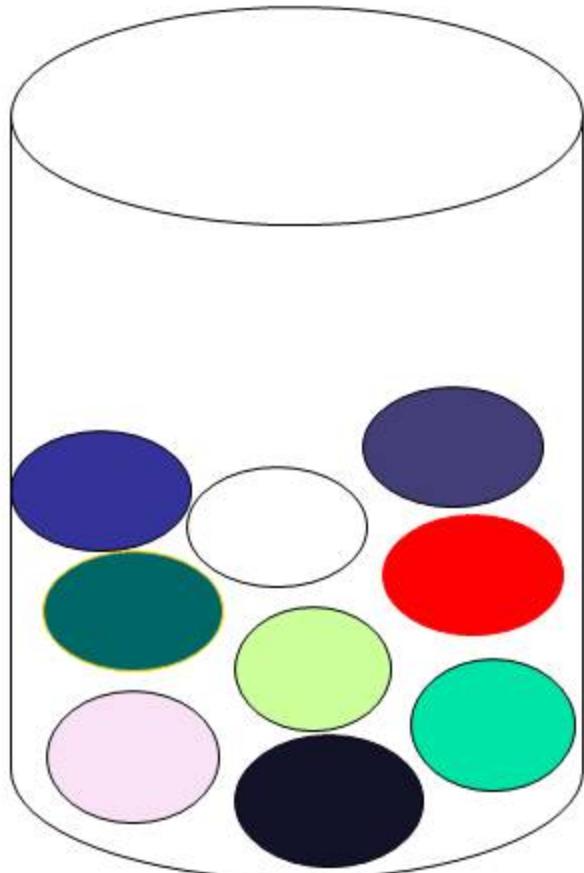


2.2.9 数据规约策略

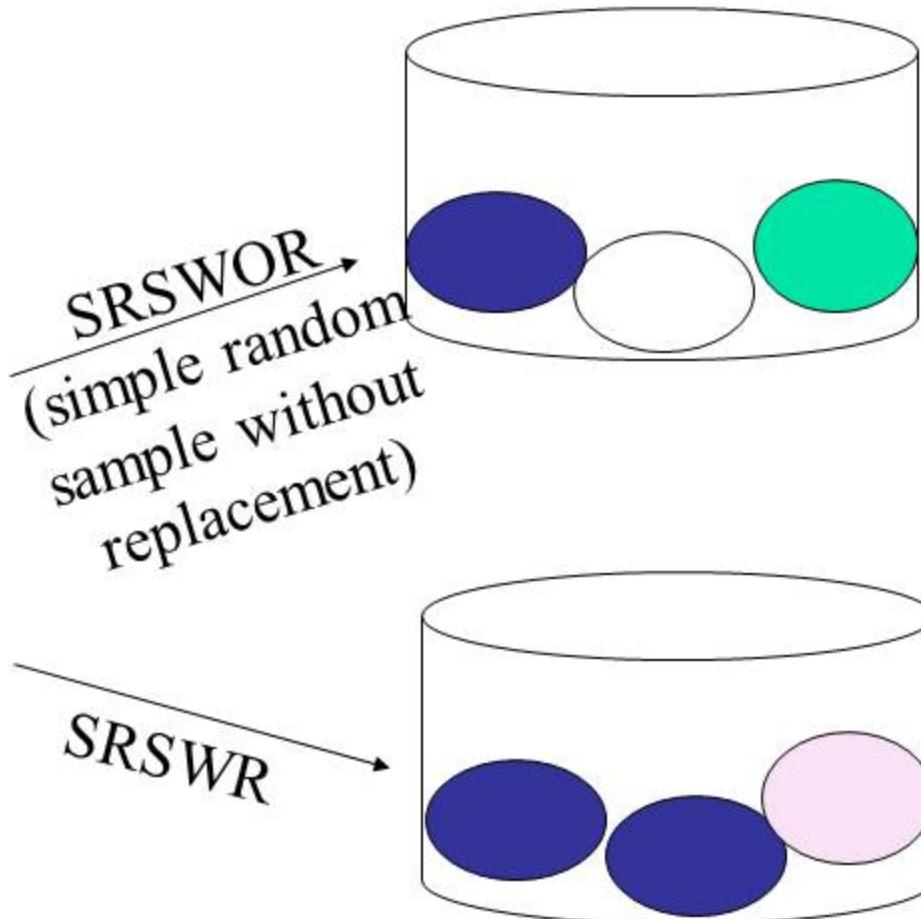
- 为什么数据规约（**data reduction**）？
 - 由于数据仓库可以存储TB的数据，因此在一个完整的数据集上运行时，复杂的数据分析可能需要一个很长的时间
- 降维
- 降数据
- 数据压缩



2.2.9 降数据典型方法-抽样法



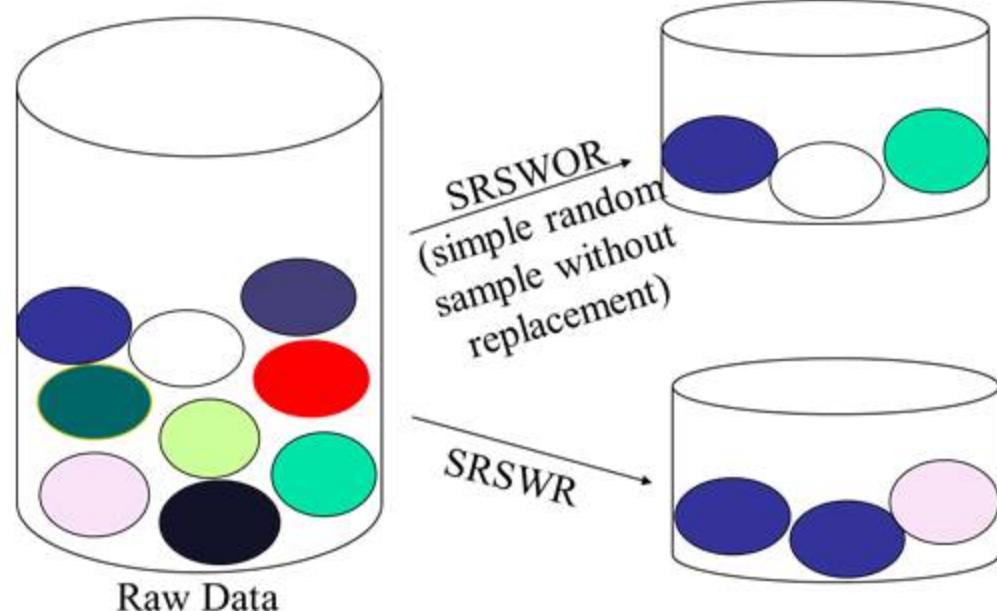
Raw Data





2.2.9 抽样类型

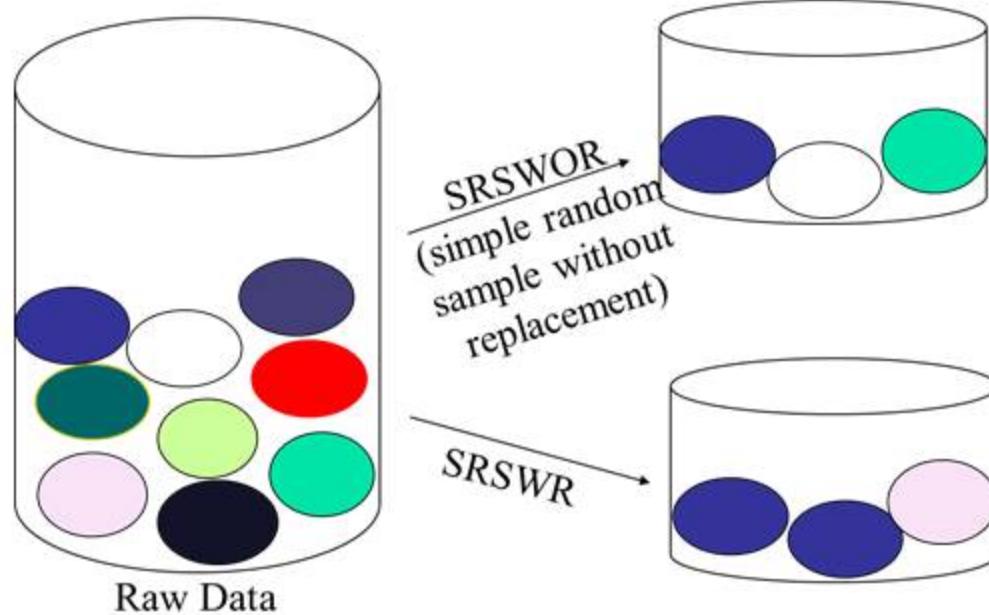
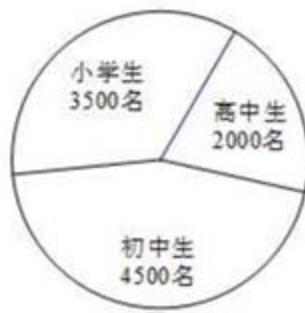
- 简单随机抽样(Simple Random Sampling)
 - 相等的概率选择
 - 不放回抽样(Sampling without replacement)
 - 一旦对象被选中，则将其删除
 - 有放回抽样(Sampling with replacement)
 - 选择对象不会被删除





2.2.9 采样类型

- 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
- 分层抽样
 - 每组抽相同个数
 - 用于偏斜数据





2.2.9 样本大小对数据质量的影响

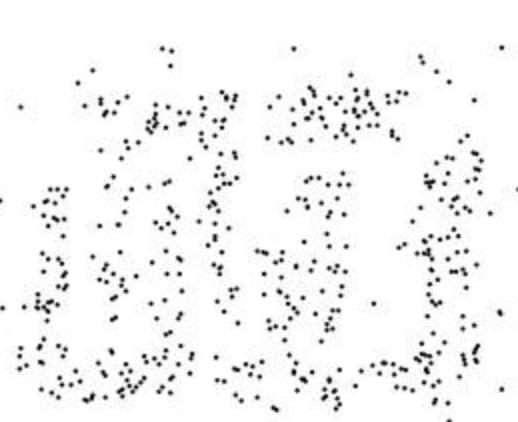
- 从8000个点分别抽2000和500个点
 - 2000个点的样本保留了数据集的大部分结构
 - 500个点的样本丢失了许多结构



8000 points



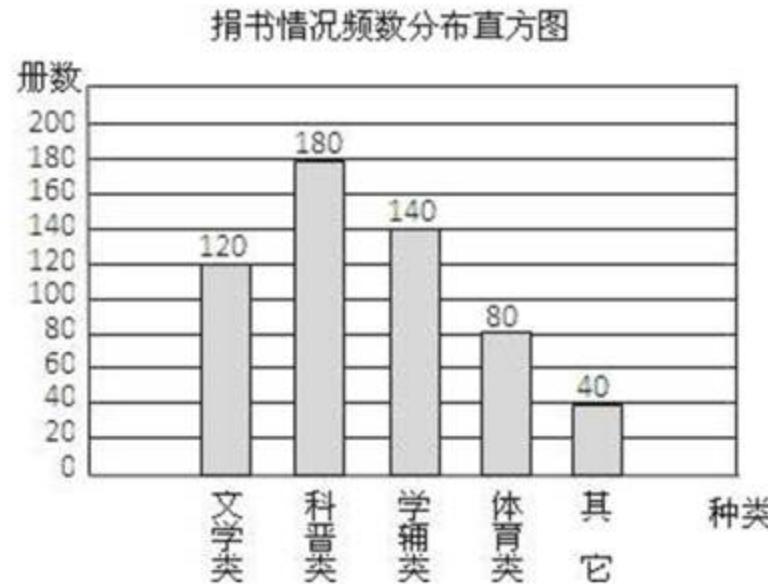
2000 Points



500 Points

图所示数据适合采用哪种抽样方法

- A 不放回抽样
- B 有放回抽样
- C 分层抽样



提交

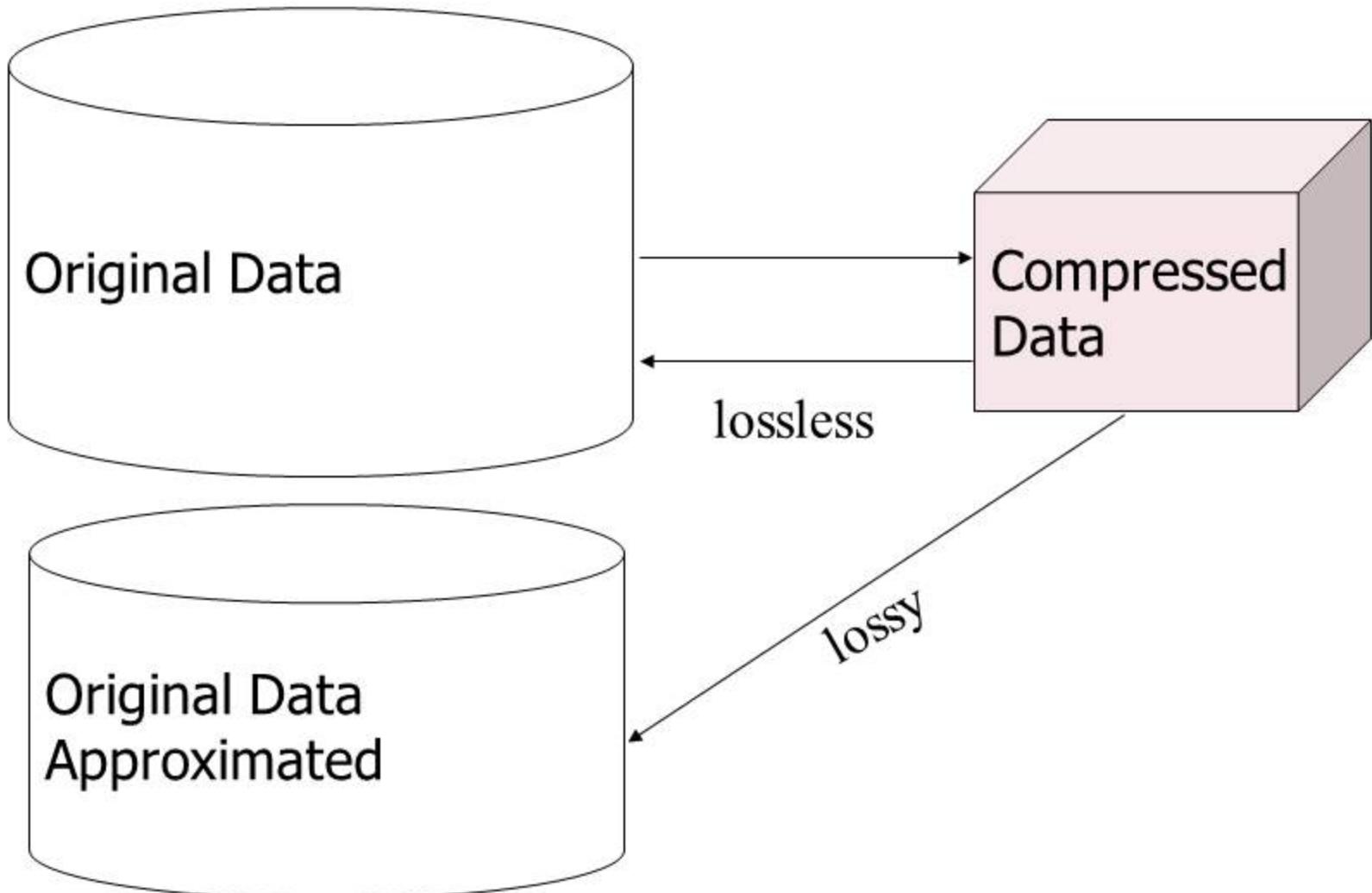


2.2.10 数据规约策略

- 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



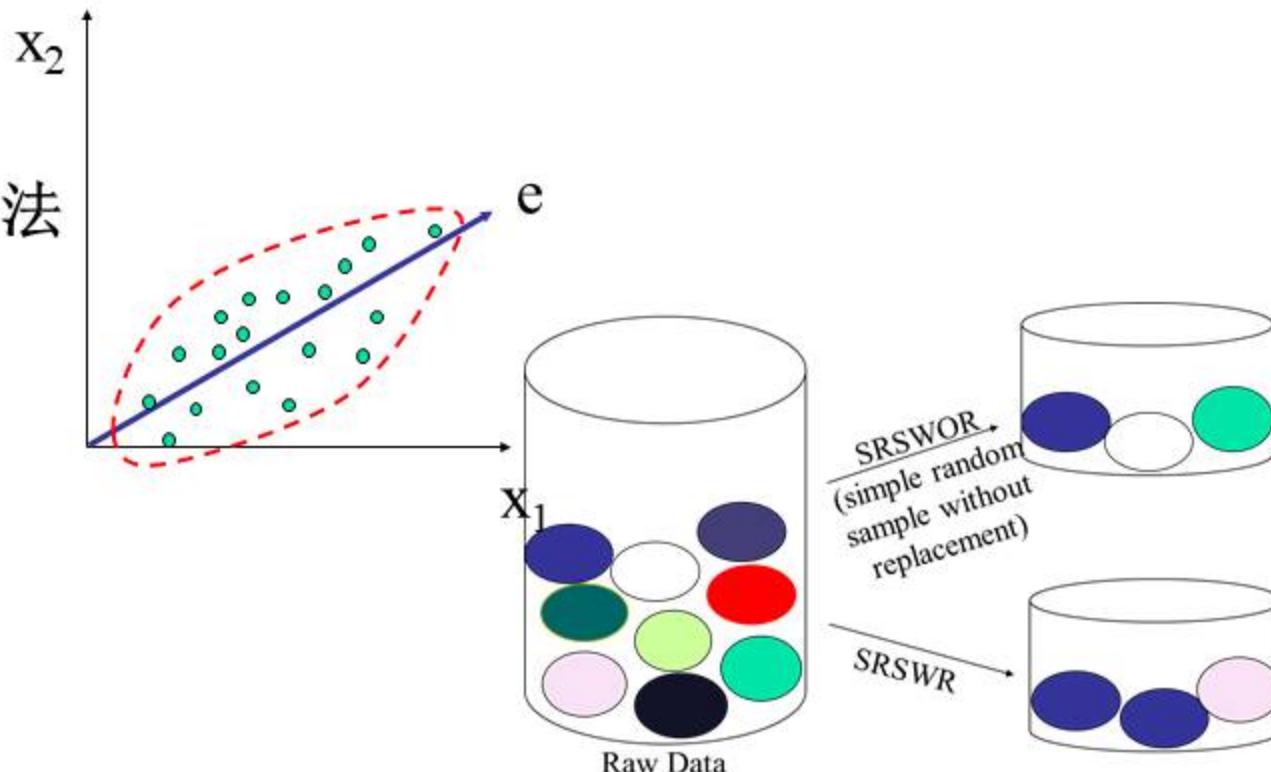
2.2.10数据压缩





2.2.11 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





2.2.12 数据转换和离散化

- 函数映射指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



2.2.12 数据转换和离散化-规范化方法

■ 最小-最大规范化

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

- 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
...



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： [填空1]

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

正常使用填空题需3.0以上版本雨课堂



2.2.12 数据转换和离散化-规范化方法

■ 最小-最大规范化

$$v' = \frac{v - \min_A}{\max_A - \min_A} (new_max_A - new_min_A) + new_min_A$$

■ v即需要规范的数据

■ 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
...



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： [填空1]

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

正常使用填空题需3.0以上版本雨课堂



2.2.12 数据转换和离散化-规范化方法

- 最小-最大规范化

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

- v 即需要规范的数据

- z-分数规范化

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

- 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



2.2.13 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



2.2.13 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



2.2.13 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
 - 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



2.2.13 数据转换和离散化-离散化方法

■ 非监督离散化法

- 等宽法

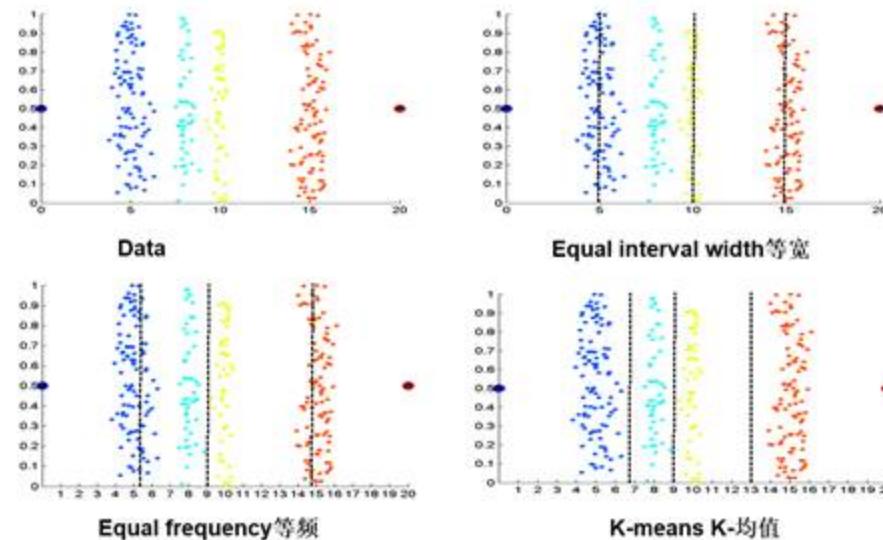
- 根据属性的值域来划分，使每个区间的宽度相等

- 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





下列哪些是非监督数据离散化方法

- A 等宽法
- B 等频法
- C 聚类法
- D 决策树法

提交



2.2.1 数据规约编程实践

<https://scikit-learn.org/stable/modules/preprocessing.html#standardization-or-mean-removal-and-variance-scaling>

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

编程升级：课后同学们参考上述链接写写代码完成实践，不计入平时分，但是这个实践很有用！！！



2.2数据预处理-第4次课后作业

- 第四次课后作业-在educoder平台上完成作业
 - <https://www.educoder.net/shixuns/2h9j74o6/challenges>
 - <https://www.educoder.net/shixuns/rbnaxywe/challenges>

提交作业截至时间： **2020年2月28日**



2.3 特征构造



2.3 特征构造

- 机器学习行内有句被奉为真理的话，**数据和特征**决定了机器学习的上限，**模型和算法**只是逼近这个上限而已
 - 1、当数据质量不高、特征信息不明显，通常需要构造新特征
 - 基本特征构造法
 - 时间类型数据特征构造法
 - 时间序列数据特征构造法
 - 2、当数据中既有连续数据，又有离散数据时，当采用类似神经网络这种优化机器学习方法，则需要
 - 将离散数据特征进行哑编码



2.3.1 特征构造-基本特征构造法

- 原始数据集的特征具有必要的信息，但其形式不适合数据挖掘算法
- 由原特征构造的新特征可能比原特征更有用
- 例：文物数据库
 - 每件文物的特征包括：体积和质量，以及其他信息
 - 文物材质(类)：木材、陶土、青铜、黄金
 - 原特征不适合分类
 - 构造新特征：密度=质量/体积

$$p = \frac{m}{v}$$

mass
density
volume



2.3.1 特征构造-基本特征构造法

- 单调变换（如对数变换、指数变换等）

- 不适用于决策树类算法，

对于决策树而言，指数、对数变换等之间没有差异

月收入		
1亿	\log_{10}	8
10000		4
1000		3



2.3.1 特征构造-基本特征构造法

- 单调变换 (如对数变换、指数变换等)

- 不适用于决策树类算法，

对于决策树而言，指数、对数变换等之间没有差异

月收入	
1亿	\log_{10}
10000	8
1000	4
	3

- 线性组合 (linear combination) $s = w_1x_1 + w_2x_2 + \dots + w_kx_k$

- 仅适用于决策树以及基于决策树的ensemble (如gradient boosting, random forest)，因为常见的决策树模型不擅长捕获不同特征之间的相关性
 - 不适用于SVM、线性回归、神经网络等



2.3.1 特征构造-基本特征构造法

- 单调变换 (如对数变换、指数变换等)

- 不适用于决策树类算法，

对于决策树而言，指数、对数变换等之间没有差异

月收入	
1亿	\log_{10}
10000	8
1000	4
	3

- 线性组合 (linear combination) $s = w_1x_1 + w_2x_2 + \dots + w_kx_k$

- 仅适用于决策树以及基于决策树的ensemble (如gradient boosting, random forest)，因为常见的决策树模型不擅长捕获不同特征之间的相关性
 - 不适用于SVM、线性回归、神经网络等

- 比例特征 (ratio feature) : x_1/x_2

- 绝对值 (absolute value)

- 最大值 $\max(x_1, x_2)$ ，最小值 $\min(x_1, x_2)$



决策树算法特征构造，适合采用什么方法

- A 单调变换
- B 线性组合
- C 绝对值

提交



2.3.2 特征构造-时间类型数据特征构造法

- 机器学习行内有句被奉为真理的话，**数据和特征**决定了机器学习的上限，**模型和算法**只是逼近这个上限而已
 - 1、当数据质量不高、特征信息不明显，通常需要构造新特征
 - 基本特征构造法
 - **时间类型数据特征构造法**
 - 时间序列数据特征构造法
 - 2、当数据中既有连续数据，又有离散数据时，当采用类似神经网络这种优化机器学习方法，则需要
 - 将离散数据特征进行哑编码



2.3.2特征构造-时间类型数据特征构造法

题目内容

光伏发电具有波动性和间歇性，大规模光伏电站并网运行可能对电力系统的安全稳定经济运行造成影响。对光伏电站的输出功率进行准确率预测，有助于调度部门统筹安排常规能源和光伏发电的协调配合，及时调整调度计划，合理安排电网运行方式。因此，本题旨在通过利用气象信息、历史数据、组件信息等，通过机器学习、人工智能方法，预测未来发电功率，为进一步为光伏发电功率提供准确的预测结果。





2.3.2 特征构造-时间类型数据特征构造法

题目数据

训练集数据提供了4个电场的脱敏后的环境数据和电场实际辐照度和电场发电功率。

测试集数据提供了4个电场的脱敏后的环境数据，需要利用这些数据预测每个时间点的光伏发电功率。

时间	辐照度	风速	温度	压强	湿度	实发辐照度	实际功率
2016-04-01 00:	-1	-0.70755	-0.09091	-0.0303	-0.15789	0	-0.01933
2016-04-01 00:	-1	-0.70755	-0.09091	-0.0303	-0.15789	0	-0.01933
2016-04-01 00:	-1	-0.71698	-0.10707	-0.0303	-0.13684	0	-0.021
2016-04-01 00:	-1	-0.72642	-0.12323	0.030303	-0.09474	0	-0.022
2016-04-01 01:	-1	-0.73585	-0.13535	0.030303	-0.07368	0	-0.022
2016-04-01 01:	-1	-0.75472	-0.14747	0.030303	-0.05263	0	-0.022
2016-04-01 01:	-1	-0.75472	-0.16364	0.030303	-0.01053	0	-0.02067
2016-04-01 01:	-1	-0.75472	-0.1798	0.030303	0.010526	0	-0.02067
2016-04-01 02:	-1	-0.76415	-0.19192	0.030303	0.031579	0	-0.02067
2016-04-01 02:	-1	-0.76415	-0.20404	0.030303	0.073684	0	-0.022
2016-04-01 02:	-1	-0.77358	-0.21616	0.030303	0.094737	0	-0.022
2016-04-01 02:	-1	-0.77358	-0.22424	0.030303	0.115789	0	-0.022
2016-04-01 03:	-1	-0.77358	-0.23636	0.030303	0.136842	0	-0.021
2016-04-01 03:	-1	-0.78302	-0.24848	0.030303	0.157895	0	-0.02067
2016-04-01 03:	-1	-0.77358	-0.25657	0.030303	0.178947	0	-0.022
2016-04-01 03:	-1	-0.77358	-0.26465	-0.0303	0.178947	0	-0.022
2016-04-01 04:	-1	-0.77358	-0.27677	-0.0303	0.2	0	-0.022
2016-04-01 04:	-1	-0.77358	-0.28485	-0.0303	0.2	0	-0.019
2016-04-01 04:	-1	-0.77358	-0.28889	-0.0303	0.2	0	-0.022
2016-04-01 04:	-1	-0.77358	-0.29697	-0.0303	0.2	0	-0.022
2016-04-01 05:	-1	-0.77358	-0.30505	-0.0303	0.2	0	-0.02067
2016-04-01 05:	-1	-0.77358	-0.30909	-0.0303	0.2	0	-0.022
2016-04-01 05:	-1	-0.77358	-0.31313	-0.0303	0.178947	0	-0.02067
2016-04-01 05:	-1	-0.77358	-0.31717	-0.0303	0.178947	0	-0.022
2016-04-01 06:	-1	-0.78302	-0.32121	-0.0303	0.178947	0	-0.02067
2016-04-01 06:	-1	-0.78302	-0.32525	-0.0303	0.157895	0	-0.022



2.3.2 特征构造-时间类型数据特征构造法

1. 由于光照在一天中不同时刻强度不同从而导致光伏功率差别很大，所以可以对一天时间划分为两段，即上午10点到下午15点为强光照时间，其余时间为弱光照时间，并分别打上0和1的标签，增加一维时刻特征。

2. 光照强度也可能因为季节的变化而变化，所以可以将一年内12个月份划分为4个季节，定3, 4, 5月为春季，6, 7, 8月为夏季，9, 10, 11月为秋季，12, 1, 2月为冬季，分别对四个对应的季节打上0, 1, 2, 3的标签，增加一维季节特征

3.....

季节变换特征
光照强度特征

时间	辐照度	风速	温度	压强	湿度	实发辐照度	实际功率	month	day	hour	type
2016-04-01 00:	-1	-0.70755	-0.09091	-0.0303	-0.15789	0	-0.01933	4	1	0	1
2016-04-01 00:	-1	-0.70755	-0.09091	-0.0303	-0.15789	0	-0.01933	4	1	0	1
2016-04-01 00:	-1	-0.71698	-0.10707	-0.0303	-0.13684	0	-0.021	4	1	0	1
2016-04-01 00:	-1	-0.72642	-0.12323	0.030303	-0.09474	0	-0.022	4	1	0	1
2016-04-01 01:	-1	-0.73585	-0.13535	0.030303	-0.07368	0	-0.022	4	1	1	1
2016-04-01 01:	-1	-0.75472	-0.14747	0.030303	-0.05263	0	-0.022	4	1	1	1
2016-04-01 01:	-1	-0.75472	-0.16364	0.030303	-0.01053	0	-0.02067	4	1	1	1
2016-04-01 01:	-1	-0.75472	-0.1798	0.030303	0.010526	0	-0.02067	4	1	1	1
2016-04-01 02:	-1	-0.76415	-0.19192	0.030303	0.031579	0	-0.02067	4	1	2	1
2016-04-01 02:	-1	-0.76415	-0.20404	0.030303	0.073684	0	-0.022	4	1	2	1
2016-04-01 02:	-1	-0.77358	-0.21616	0.030303	0.094737	0	-0.022	4	1	2	1
2016-04-01 02:	-1	-0.77358	-0.22424	0.030303	0.115789	0	-0.022	4	1	2	1
2016-04-01 03:	-1	-0.77358	-0.23636	0.030303	0.136842	0	-0.021	4	1	3	1
2016-04-01 03:	-1	-0.78302	-0.24848	0.030303	0.157895	0	-0.02067	4	1	3	1
2016-04-01 03:	-1	-0.77358	-0.25657	0.030303	0.178947	0	-0.022	4	1	3	1



2.3.3 特征构造-时间序列数据特征构造法

- 机器学习行内有句被奉为真理的话，**数据和特征**决定了机器学习的上限，**模型和算法**只是逼近这个上限而已
 - 1、当数据质量不高、特征信息不明显，通常需要构造新特征
 - 基本特征构造法
 - 时间类型数据特征构造法
 - **时间序列数据特征构造法**
 - 2、当数据中既有连续数据，又有离散数据时，当采用类似神经网络这种优化机器学习方法，则需要
 - 将离散数据特征进行哑编码



2.3.3 特征构造-时间序列数据特征构造法

轴承故障检测：任务介绍

- 轴承有3种故障：外圈故障，内圈故障，滚珠故障，外加正常的工作状态。如表1所示，结合轴承的3种直径（直径1，直径2，直径3），轴承的工作状态有10类：

	外圈故障	内圈故障	滚珠故障	正常
直径 1	1	2	3	
直径 2	4	5	6	0
直径 3	7	8	9	

表 1 轴承的故障类别



2.3.3 特征构造-时间序列数据特征构造法

■ 数据集：

- **1.train.csv**, 训练集数据, 1到**6000**为按**时间序列连续采样的振动信号数值**, 每行数据是一个样本, 共**792**条数据, 第一列**id**字段为样本编号, 最后一列**label**字段为标签数据, 即轴承的工作状态, 用数字**0**到**9**表示。
- **2.test_data.csv**, 测试集数据, 共**528**条数据, 除无**label**字段外, 其他字段同训练集。

id	1	2	3	5999	6000	HVV
1	0.5636499	1.069229242	-0.837759182		-0.018273952	0.021522655	7

	外圈故障	内圈故障	滚珠故障	正常
直径 1	1	2	3	0
直径 2	4	5	6	
直径 3	7	8	9	

表 1 轴承的故障类别



2.3.3 特征构造-时间序列数据特征构造法

■ 数据集：

- 1. **train.csv**, 训练集数据, 1到6000为按**时间序列连续采样的振动信号数值**, 每行数据是一个样本, 共792条数据, 第一列**id**字段为样本编号, 最后一列**label**字段为标签数据, 即轴承的工作状态, 用数字0到9表示。
- 2. **test_data.csv**, 测试集数据, 共528条数据, 除无**label**字段外, 其他字段同训练集。

id	1	2	3	5999	6000	HVV
1	0.5636499	1.069229242	-0.837759182		-0.018273952	0.021522655	7

	外圈故障	内圈故障	滚珠故障	正常
直径1	1	2	3	0
直径2	4	5	6	
直径3	7	8	9	

表1 轴承的故障类别

原始数据无法直接拿来预测轴承故障



2.3.3 特征构造-时间序列数据特征构造法

id	1	2	3	5999	6000	HVV
1	0.5636499	1.069229242	-0.837759182		-0.018273952	0.021522655	7

- 振动的幅值、频率和相位是振动的三个基本参数，称为振动三要素
 - 幅值：幅值是振动强度的标志
 - 频率：不同的频率成分反映系统内不同的振源
 - 相位：利用相位关系确定共振点、测量振型、旋转件动平衡、有源振动控制、降噪等



2.3.3 特征构造-时间序列数据特征构造法

id	1	2	3	5999	6000	HVV
1	0.5636499	1.069229242	-0.837759182		-0.018273952	0.021522655	7

- 振动的幅值、频率和相位是振动的三个基本参数，称为振动三要素
 - 幅值：幅值是振动强度的标志
 - 频率：不同的频率成分反映系统内不同的振源
 - 相位：利用相位关系确定共振点、测量振型、旋转件动平衡、有源振动控制、降噪等

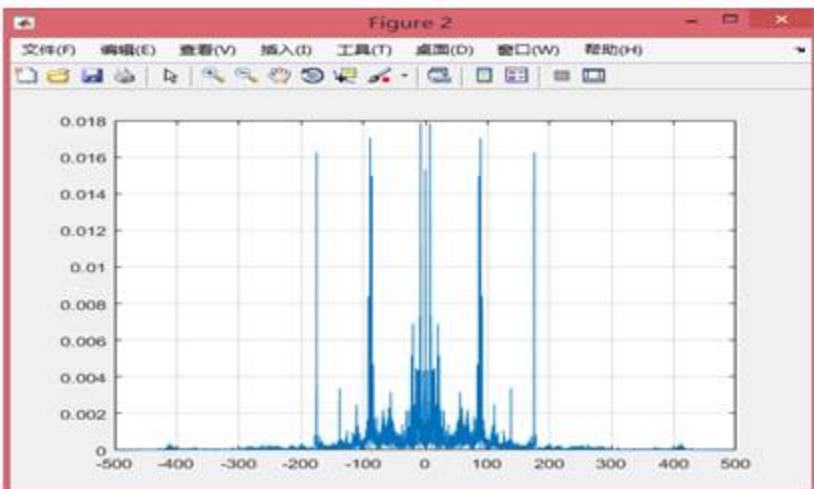
振动三要素能够体现轴承故障不同特征



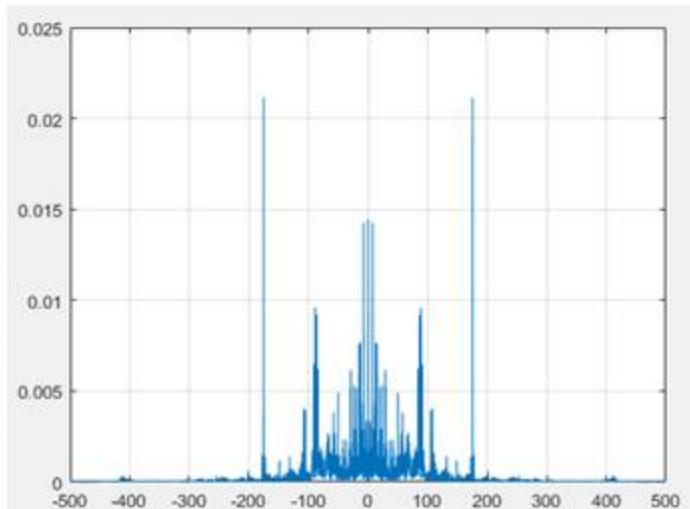
2.3.3 特征构造-时间序列数据特征构造法

id	1	2	3	5999	6000	HVV
1	0.5636499	1.069229242	-0.837759182		-0.018273952	0.021522655	7

- 振动三要素特征构造（幅值、频率和相位）
 - 傅里叶变换(Fourier transform)
 - 小波变换(Wavelet transform)



对比2种
轴承故障，
频谱具
有差异性





时间序列数据特征构造方法主要包括

- A 傅里叶变换
- B 小波变换
- C 单调变换

提交



2.3.4 特征构造-离散数据特征哑编码

- 机器学习行内有句被奉为真理的话，**数据和特征**决定了机器学习的上限，**模型和算法**只是逼近这个上限而已
 - 1、当数据质量不高、特征信息不明显，通常需要构造新特征
 - 基本特征构造法
 - 时间类型数据特征构造法
 - 时间序列数据特征构造法
 - 2、当数据中既有连续数据，又有离散数据时，当采用类似神经网络这种优化机器学习方法，则需要
 - 将离散数据特征进行哑编码



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	红色	
2	白色	
3	黑色	
4	

- 利用“颜色”等特征来预测品味等级
 - 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	红色	
2	白色	
3	黑色	
4	

- 利用“颜色”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换

- 要求特征为连续的值



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	红色	
2	白色	
3	黑色	
4	

- 利用“颜色”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换

- 要求特征为连续的值

方法1：如果属性具有m个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	0 (红色)	
2	1 (白色)	
3	2 (黑色)	
4	

- 利用“颜色”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换

- 要求特征为连续的值

方法1：如果属性具有m个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	0 (红色)	方法 1 通过将字符类的离散数据用 0、1、...、2 来表示，构造成神经网络可以识别的连续数据	
2	1 (白色)	
3	2 (黑色)	
4	

- 利用“颜色”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换

- 要求特征为连续的值

方法**1**: 如果属性具有**m**个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	0 (红色)	方法 1 通过将字符类的离散数据用 0、1、...、2 来表示，构造成神经网络可以识别的连续数据	
2	1 (白色)	
3	2 (黑色)	
4	该方法是否可行？	

- 利用“颜色”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换

- 要求特征为连续的值

方法**1**: 如果属性具有**m**个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数



该方法是否可行?

A 是

B 否

提交



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	0 (红色)	方法 1 通过将字符类的离散数据用 0、1、2 来表示，构造成神经网络可以识别的连续数据	
2	1 (白色)	
3	2 (黑色)	
4	用 0、1、2 来表示红、白、黑，则缺省的认为红<白<黑	

- 利用“颜色”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换

- 要求特征为连续的值

方法**1**: 如果属性具有**m**个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	0 (红色)	方法 1 通过将字符类的离散数据用 0、1、...、2 来表示，构造成神经网络可以识别的连续数据	
2	1 (白色)	
3	2 (黑色)	
4	

- 利用“颜色”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换

- 要求特征为连续的值

方法**1**: 如果属性具有**m**个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	0 (红色)	
2	1 (白色)	
3	2 (黑色)	
4	

- 利用“颜色”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换

- 要求特征为连续的值

方法2: 离散数据特征哑编



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	0 (红色)	
2	1 (白色)	
3	2 (黑色)	
4	

- 把 m 个整数都转换成一个二进制数
 - 需要 $n = \lceil \log_2 m \rceil$ 个二进位表示这些整数
- 用 n 个二元属性表示这些二进制数
- 例：5种颜色的分类变量需要三个二元变量
 - x_1 、 x_2 、 x_3

	x_1	x_2	x_3
红	0	0	0
白	0	0	1
黑	0	1	0
蓝	0	1	1
绿	1	0	0



2.3.4 特征构造-离散数据特征哑编码

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	0 (红色)	
2	1 (白色)	该方法也适用于将多分类问题变成多个二分类问题
3	2 (黑色)	
4	

- 把 m 个整数都转换成一个二进制数
 - 需要 $n = \lceil \log_2 m \rceil$ 个二进位表示这些整数
- 用 n 个二元属性表示这些二进制数
- 例：5种颜色的分类变量需要三个二元变量
 - x_1 、 x_2 、 x_3

	x_1	x_2	x_3
红	0	0	0
白	0	0	1
黑	0	1	0
蓝	0	1	1
绿	1	0	0



2.3.4 特征构造-离散数据特征哑编码

用户id	舱位等级	其他特征	品味等级（类别标签）
1	头等舱	
2	一等舱	
3	二等舱	
4	

- 利用“舱位等级”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换
 - 要求特征为连续的值



2.3.4 特征构造-离散数据特征哑编码

用户id	舱位等级	其他特征	品味等级（类别标签）
1	头等舱	方法1：如果属性具有m个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数
2	一等舱
3	二等舱
4	方法2：离散数据特征哑编

- 利用“舱位等级”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换
 - 要求特征为连续的值



2.3.4 特征构造-离散数据特征哑编码

用户id	舱位等级	其他特征	品味等级（类别标签）
1	头等舱	方法1：如果属性具有m个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数
2	一等舱
3	二等舱
4	方法2：离散数据特征哑编 采用哪个方法呢？

- 利用“舱位等级”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换
 - 要求特征为连续的值



利用“舱位等级”等特征来预测品味等级的特征
构造采用哪个方法

- A 方法1
- B 方法2

提交



2.3.4 特征构造-离散数据特征哑编码

用户id	舱位等级	其他特征	品味等级（类别标签）
1	头等舱	方法1：如果属性具有m个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数	
2	一等舱		
3	二等舱	
4	

- 利用“舱位等级”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换
 - 要求特征为连续的值



2.3.4 特征构造-离散数据特征哑编码

用户id	舱位等级	其他特征	品味等级（类别标签）
1	头等舱	方法 1 : 如果属性具有m个值，则将每个原始值唯一地映射到区间[0, m-1]中的一个整数
2	一等舱	舱位等级本身是有序的
3	二等舱
4

- 利用“舱位等级”等特征来预测品味等级

- 假设采用神经网络分类器

$$\hat{y} = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_dx_d - t)$$

- 神经网络分类器为各个特征的线性变换
 - 要求特征为连续的值



2.3.4 特征构造-离散数据特征哑编码小结

用户id	衣服颜色	其他特征	品味等级（类别标签）
1	红色	
2	白色	
3	黑色	
4	

- 对**标称类（无序）离散数据**连续化特征构造通常采用**哑编码**方法
- 对**序数类离散数据**连续化特征构造可以采用直接用**[0,m-1]**的整数



2.3 特征构造

- 机器学习行内有句被奉为真理的话，**数据和特征**决定了机器学习的上限，**模型和算法**只是逼近这个上限而已
 - 1、当数据质量不高、特征信息不明显，通常需要构造新特征
 - 基本特征构造法
 - 时间类型数据特征构造法
 - 时间序列数据特征构造法
 - 2、当数据中既有连续数据，又有离散数据时，当采用类似神经网络这种优化机器学习方法，则需要
 - 将离散数据特征进行哑编码



内容提纲

- 2.1 数据质量
- 2.2 数据预处理
- 2.3 特征构造



Any Questions?

谢谢！



丢失数据处理

- ```
import numpy as np
from sklearn.preprocessing import Imputer
imp = Imputer(missing_values='NaN', strategy='mean', axis=0)
imp.fit([[1, 2], [np.nan, 3], [7, 6]])
X = [[np.nan, 2], [6, np.nan], [7, 6]]
print(imp.transform(X))
```



# 最小-最大规范化

---

- ```
from sklearn import preprocessing
import numpy as np
X = np.array([[ 1., -1.,  2.],
              [ 2.,  0.,  0.],
              [ 0.,  1., -1.]])
min_max_scaler=preprocessing.MinMaxScaler()
X_train_minmax=min_max_scaler.fit_transform(X)
print(X_train_minmax)
```



Z分数规范化

- ```
from sklearn import preprocessing
import numpy as np
X = np.array([[1., -1., 2.],
 [2., 0., 0.],
 [0., 1., -1.]])
X_scaled = preprocessing.scale(X)
print(X_scaled)
```