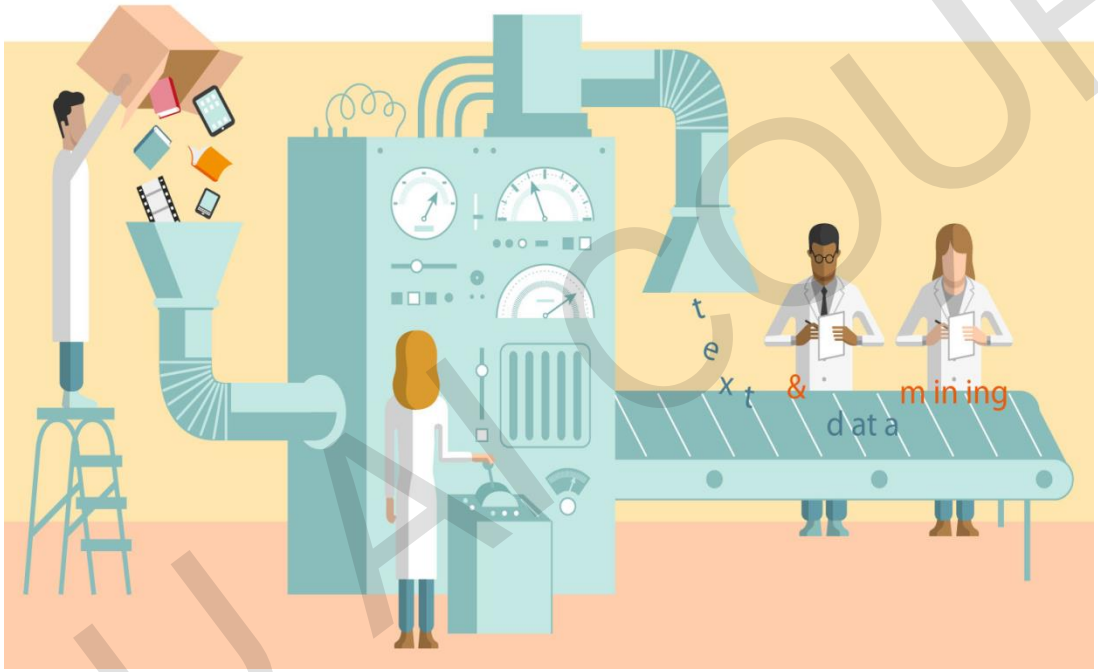
人工智能
认识数据

中山大學
SUN YAT-SEN UNIVERSITY

机器学习一般流程



数据 → 特征 → 模型→ 结果

Example

You receive an email from a medical researcher concerning a project that you are eager to work on.

Hi,

I've attached the data file.

Each line contains the information for a single patient and consists of five fields.

We want to predict the last field using the other fields.

Thanks and see you in a couple of days.

Example

The first few rows of the file are as follows:

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	165	24.0	0	427.6
⋮				

Nothing looks strange. You put your doubts aside and start the analysis.

Two days later you arrive for the meeting, and before the meeting, you strike up a conversation with a statistician who is working on the project.

Example

Statistician: So, you got the data for all the patients?

Data Miner: Yes. I haven't had much time for analysis, but I do have a few interesting results.

Statistician: Amazing. There were so many data issues with this set of patients that I couldn't do much.

Data Miner: Oh? I didn't hear about any possible problems.

Statistician: But surely you heard about what happened to field 4? It's supposed to be measured on a scale from 1 to 10, with 0 indicating a missing value, but because of a data entry error, all 10's were changed into 0's.

Data Miner: Interesting. Were there any other problems?

Statistician: Yes, fields 2 and 3 are basically the same, but I assume that you probably noticed that.

Data Miner: Yes, but these fields were only weak predictors of field 5.

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	165	24.0	0	427.6
⋮				

Example

Statistician: Anyway, given all those problems, I'm surprised you were able to accomplish anything.

Data Miner: True, but my results are really quite good. Field 1 is a very strong predictor of field 5. I'm surprised that this wasn't noticed before.

Statistician: What? Field 1 is just an identification number.

Data Miner: Nonetheless, my results speak for themselves.

Statistician: Oh, no! I just remembered. We assigned ID numbers after we sorted the records based on field 5. There is a strong connection, but it's meaningless. Sorry.

Lesson: Get to know your data!

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	165	24.0	0	427.6
.				

What is Data?

Collection of data objects and their attributes

An attribute is a property or characteristic of an object

- Examples: eye color of a person, temperature, etc.
- **Attribute** is also known as **variable, field, characteristic,** or **feature**

A collection of attributes describe an object

- **Object** is also known as **record, point, case, sample, entity**, or **instance**

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Values

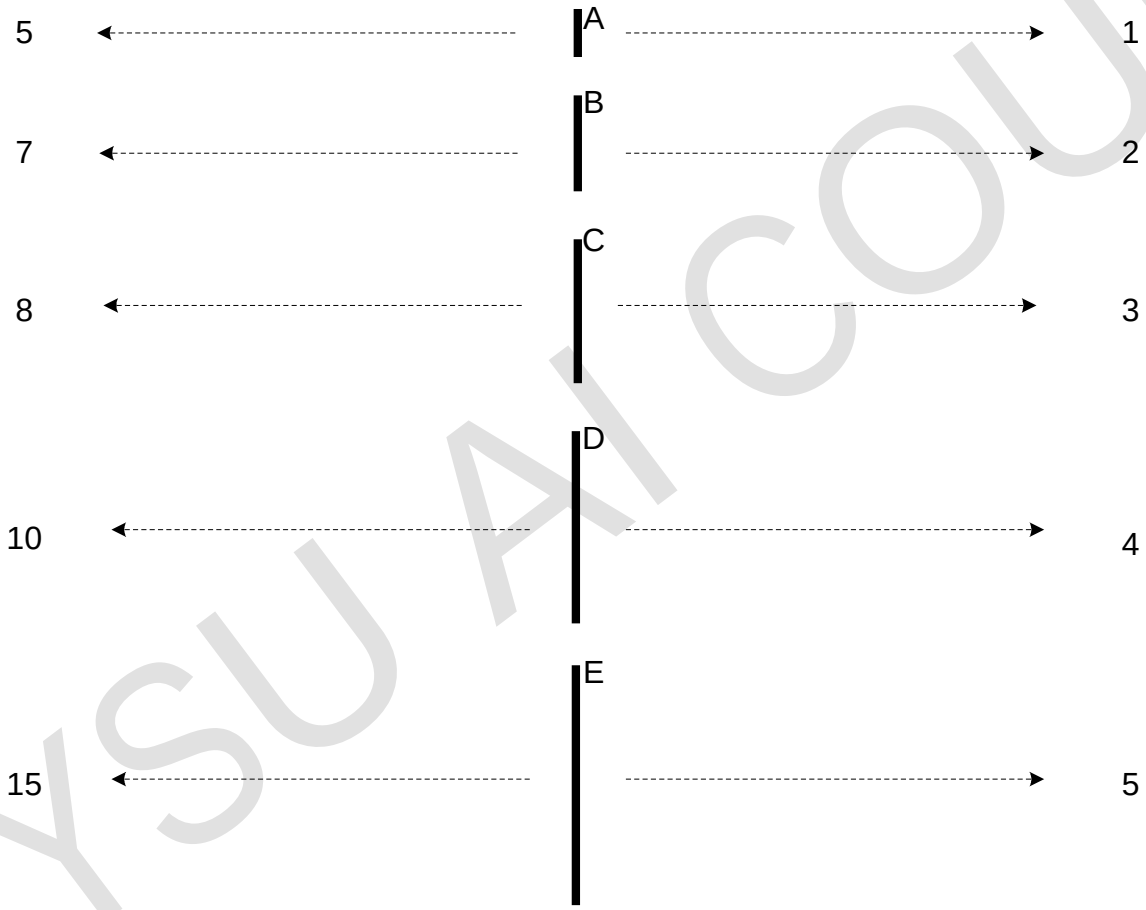
Attribute values are numbers or symbols assigned to an attribute

Distinction between attributes and attribute values

- Same attribute can be mapped to different attribute values
  - Example: height can be measured in feet or meters
- Different attributes can be mapped to the same set of values
  - Example: Attribute values for ID and age are integers
  - But properties of attribute values can be different
    - ID has no limit but age has a maximum and minimum value

Measurement of Length

The way you measure an attribute is somewhat may not match the attributes properties.



Types of Attributes

There are different types of attributes

- Nominal 定类变量
  - Examples: ID numbers, eye color, zip codes
- Ordinal 定序变量
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
- Interval 定距变量
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
- Ratio 定比变量
  - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

The type of an attribute depends on which of the following properties it possesses:

- Distinctness: = ≠
- Order: < >
- Addition: + -
- Multiplication: * /

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any one-to-one mapping	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants 华氏度 = 摄氏度 × 1.8 + 32	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value* 1 meter = 3 feet	Length can be measured in meters or feet.

Discrete and Continuous Attributes

Discrete Attribute

- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Often represented as integer variables.
- Note: binary attributes are a special case of discrete attributes

Continuous Attribute

- Has real numbers as attribute values
- Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

Important Characteristics of Structured Data

- **Dimensionality**
  - **Curse of Dimensionality**
- **Sparsity**
  - **Only presence counts**
- **Resolution**
  - **Patterns depend on the scale**

Types of data sets

Record

- Data Matrix
- Document Data
- Transaction Data

Graph

- World Wide Web
- Molecular Structures

Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

Record Data

Data that consists of a collection of records, each of which consists of a fixed set of attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

Each document becomes a `term' vector,

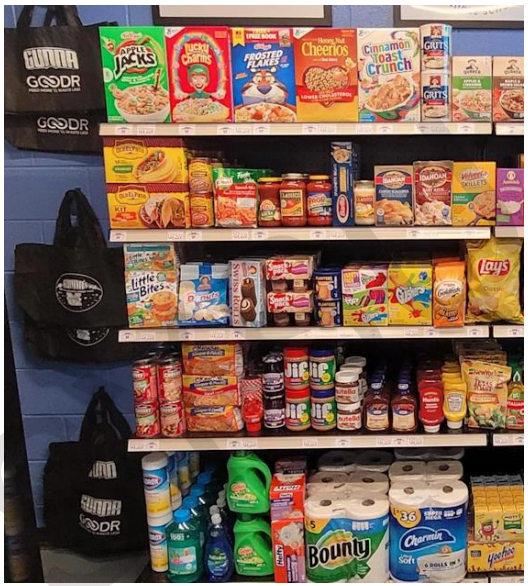
- each term is a component (attribute) of the vector,
- the value of each component is the number of times the corresponding term occurs in the document.

	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

Transaction Data

A special type of record data, where

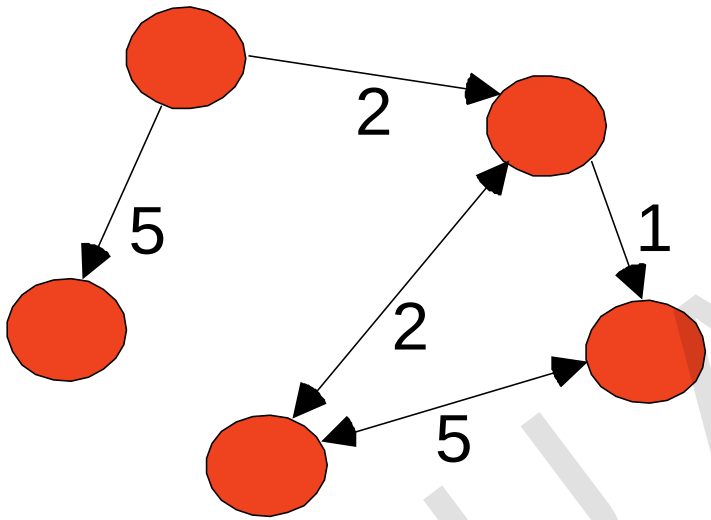
- Each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.



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

Graph Data

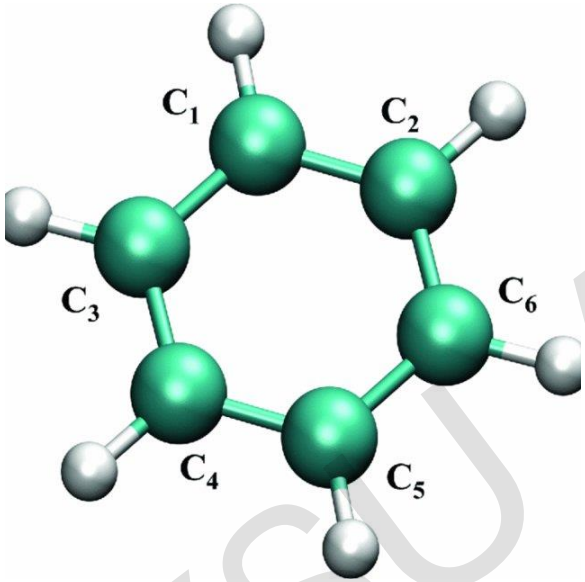
Examples: Generic graph and HTML Links

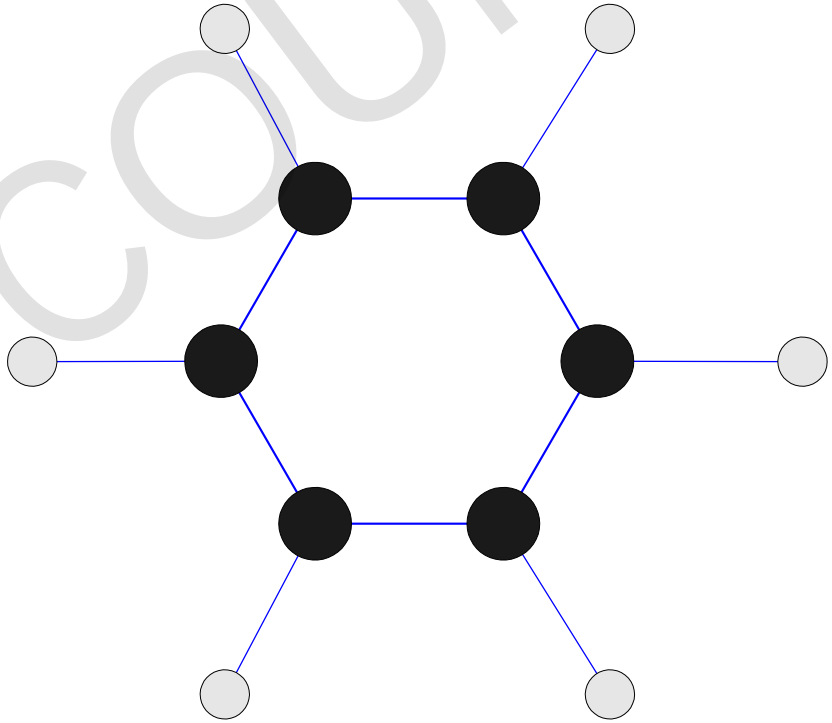



```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Chemical Data

Benzene Molecule（苯分子）: C_6H_6





Ordered Data: Sequential Data

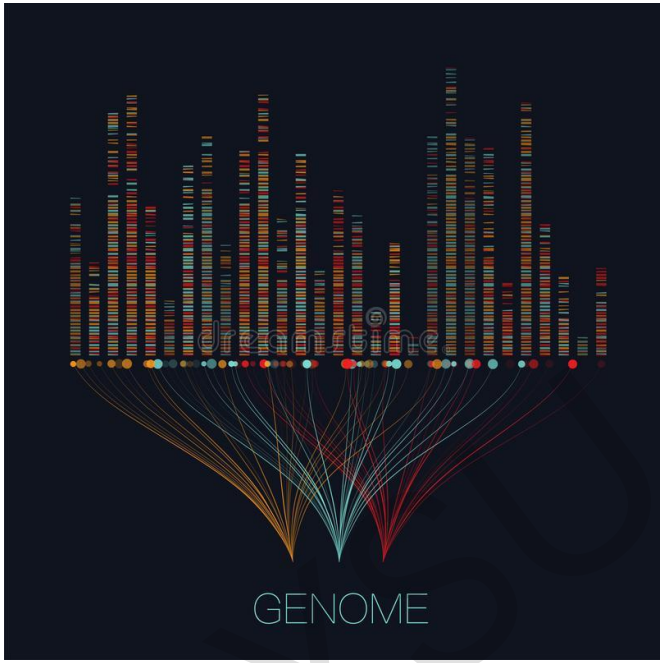
Sequential Data

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Ordered Data: Sequence Data

Genomic sequence data

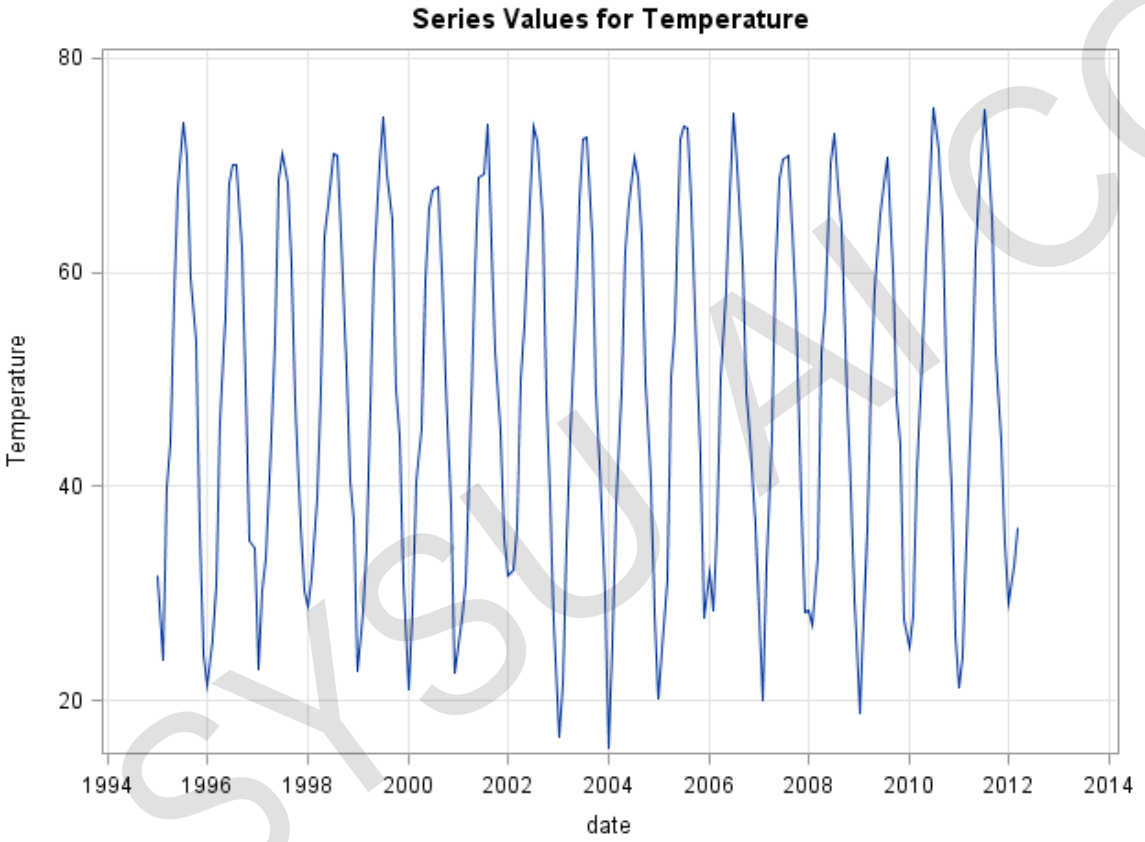


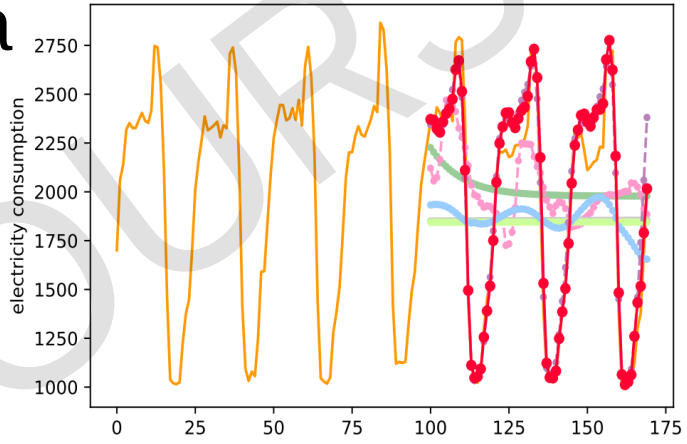

```
ATCTCTTGGCTCCAGCATCGATGAAGAACGCA
TCATTTAGAGGAAGTAAAAGTCGTAACAAGGT
GAACTGTCAAAACTTTTAACAACGGATCTCTT
TGTTGCTTCGGCGGCGCCCGCAAGGGTGCCCG
GGCCTGCCGTGGCAGATCCCCAACGCCGGGCC
TCTCTTGGCTCCAGCATCGATGAAGAACGCAG
CAGCATCGATGAAGAACGCAGCGAAACGCGAT
CGATACTTCTGAGTGTTCTTAGCGAACTGTCA
CGGATCTCTTGGCTCCAGCATCGATGAAGAAC
ACAACGGATCTCTTGGCTCCAGCATCGATGAA
CGGATCTCTTGGCTCCAGCATCGATGAAGAAC
GATGAAGAACGCAGCGAAACGCGATATGTAAT
```

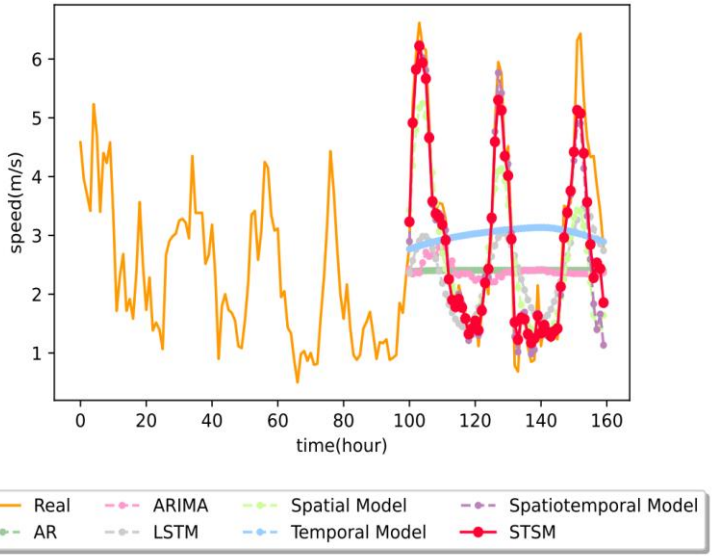
Ordered Data: Time Series Data

Special type of sequential data

Temporal autocorrelation

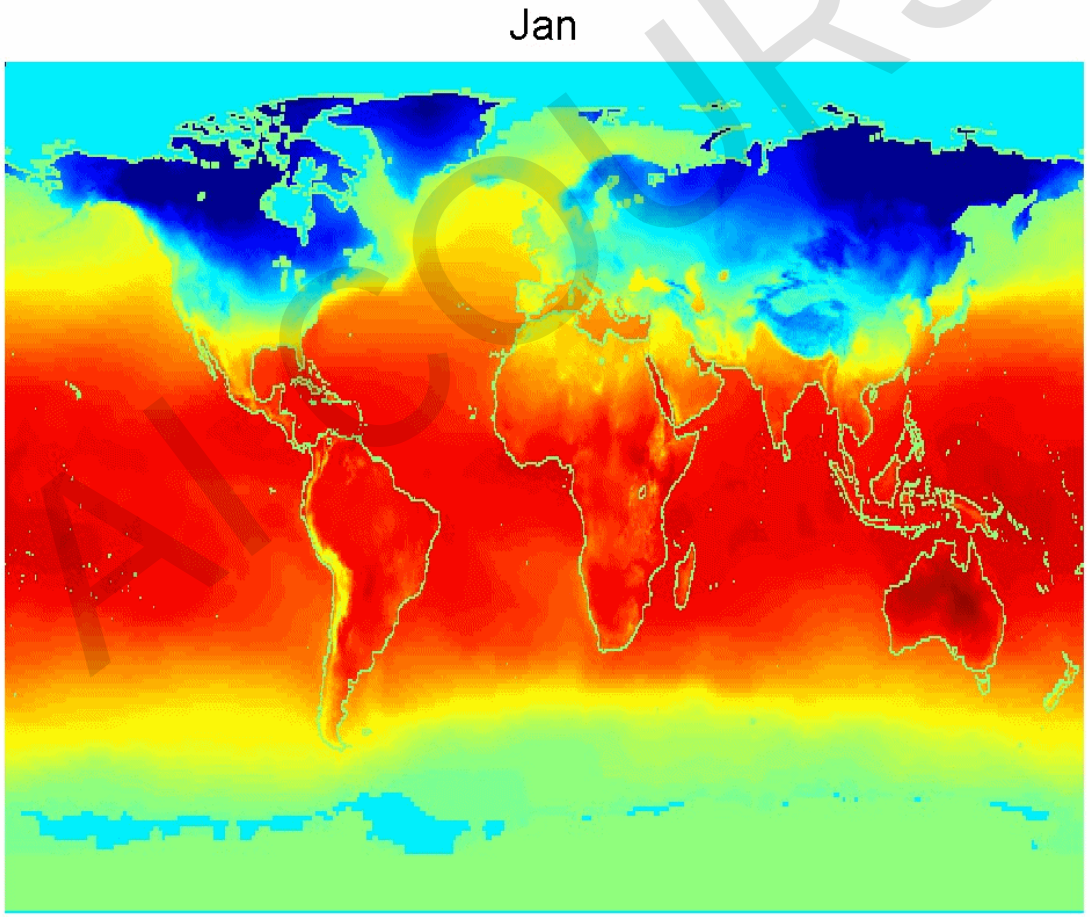






Ordered Data: Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Data Quality

What kinds of data quality problems?

How can we detect problems with the data?

What can we do about these problems?

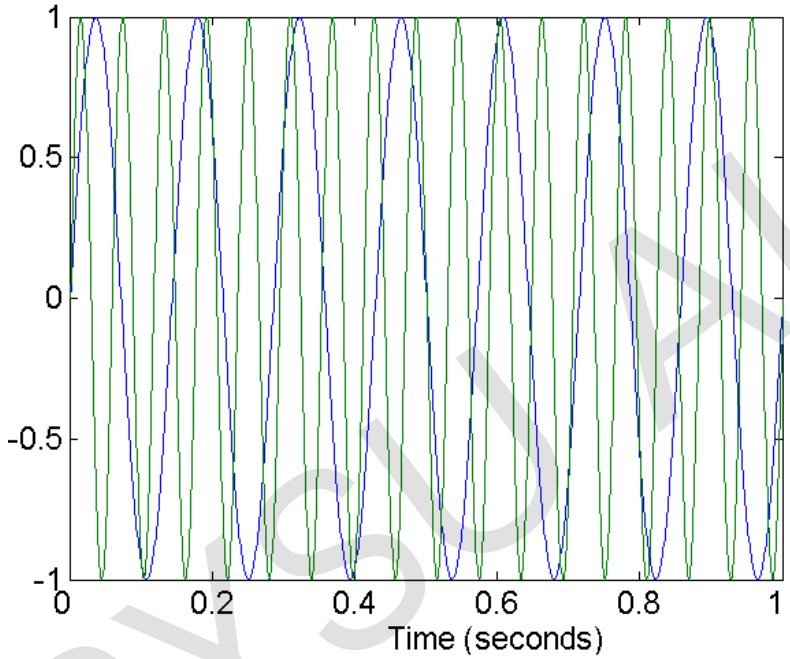
Examples of data quality problems:

- Noise and outliers
- missing values
- duplicate data

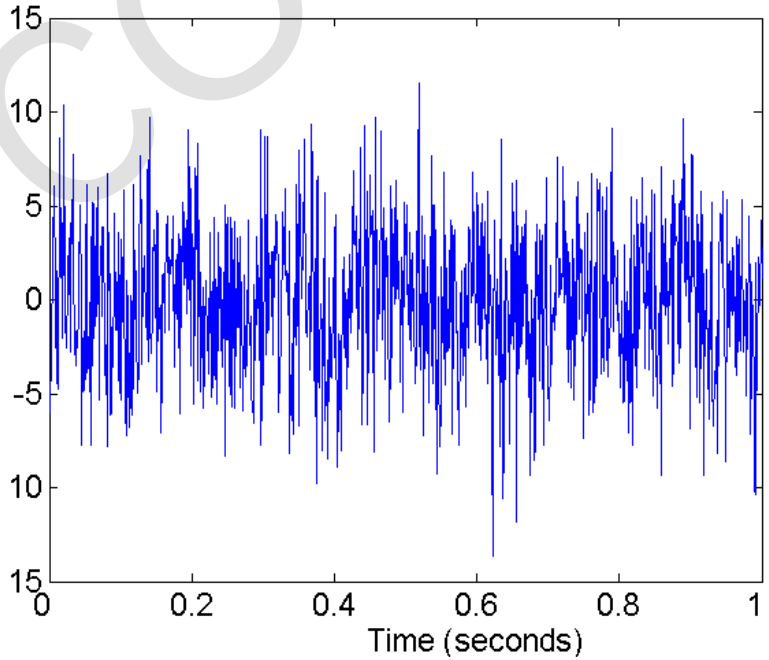
Noise

Noise refers to modification of original values

- Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



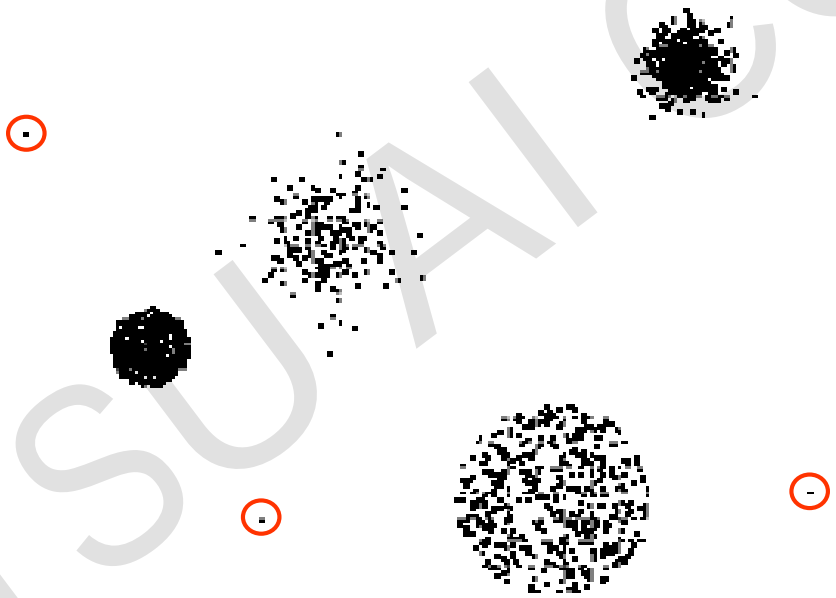
Two Sine Waves



Two Sine Waves + Noise

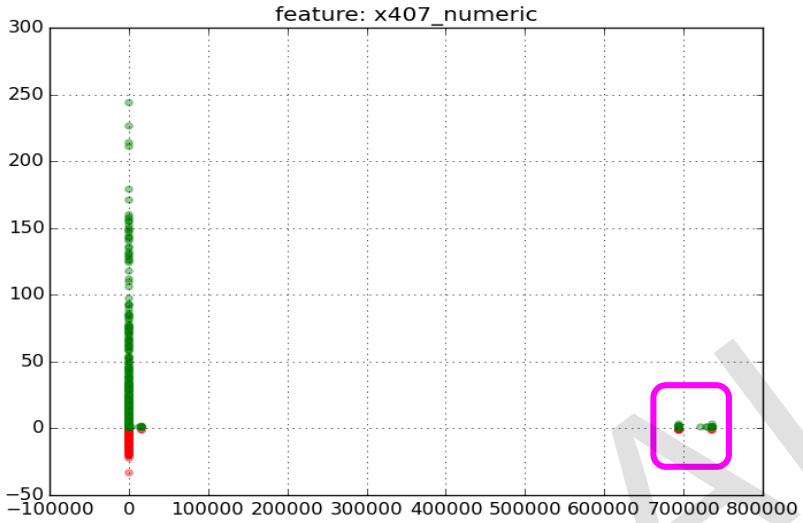
Outliers

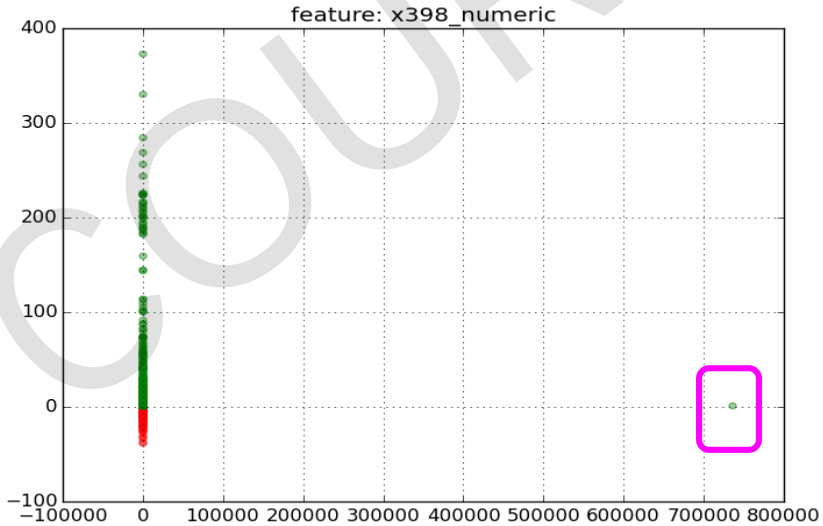
Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Outliers

X distribution map





边缘填充：If : |X – mean| > 3*std　Then: Outliers　X = mean+/-3*std

均值填充：If : |X – mean| > 3*std　Then: Outliers　X = mean

Missing Values

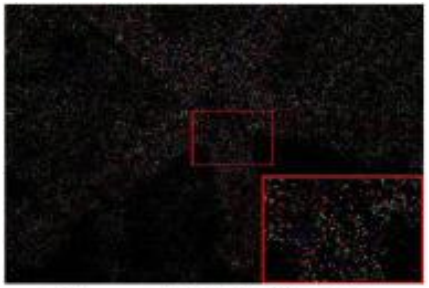
Reasons for missing values

- Information is not collected
  (e.g., people decline to give their age and weight)
- Attributes may not be applicable to all cases
  (e.g., annual income is not applicable to children)

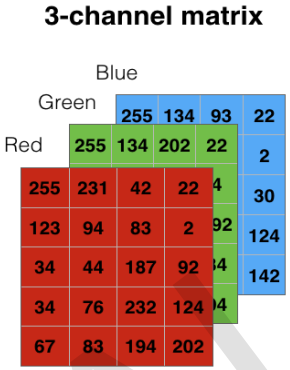
Handling missing values

- Eliminate Data Objects
- Estimate Missing Values
- Ignore the Missing Value During Analysis
- Replace with all possible values (weighted by their probabilities)

Missing Values

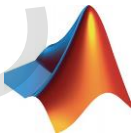


Observed (PSNR 2.03 dB)

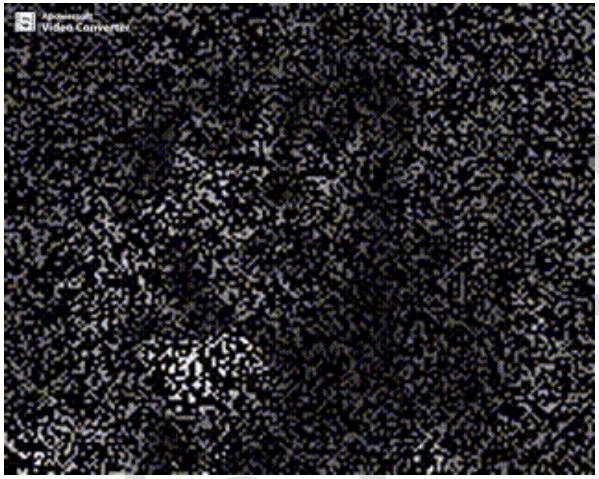


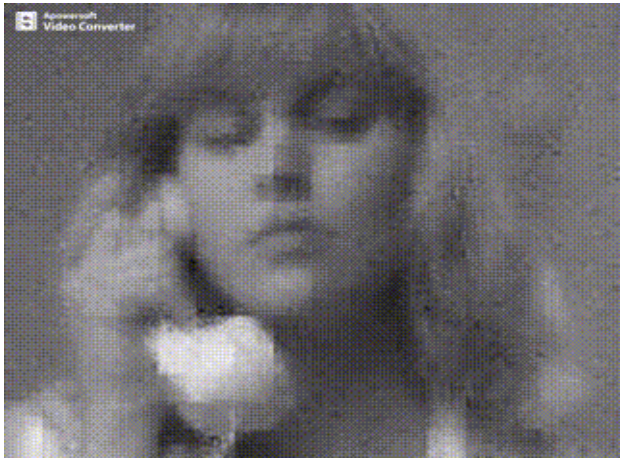
图像/视频数据的矩阵存储

矩阵运算



图像/视频复原





Duplicate Data

Data set may include data objects that are duplicates, or almost duplicates of one another

- Major issue when merging data from heterogeous sources

Examples:

- Same person with multiple email addresses

Data cleaning

- Process of dealing with duplicate data issues

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Discretization
- Attribute Transformation
- Feature creation
- Feature subset selection

Aggregation

Combining two or more attributes (or objects) into a single attribute (or object)

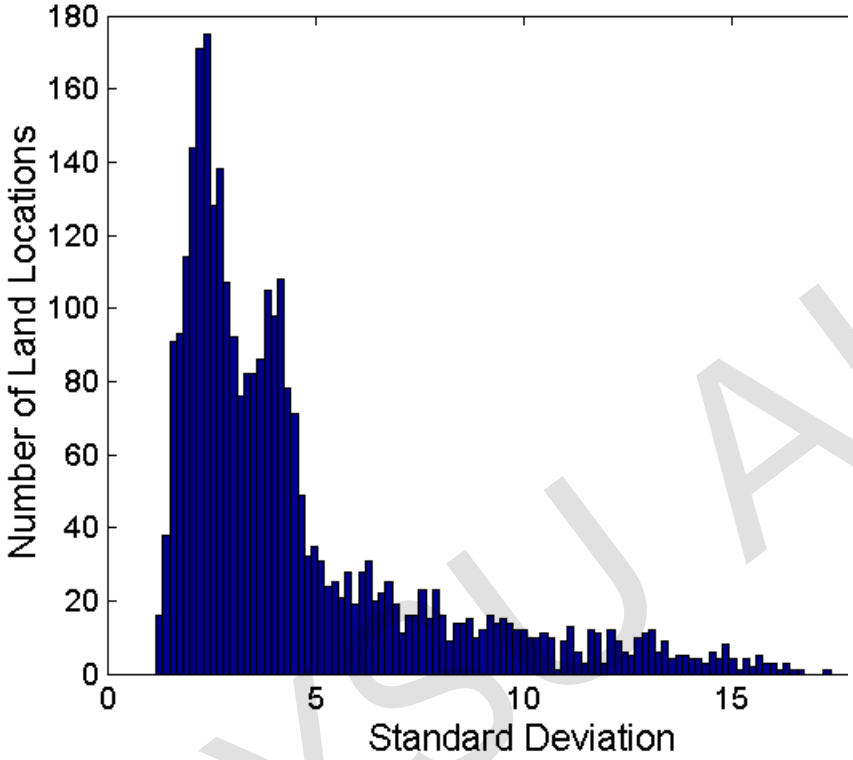
Purpose

- Data reduction
  - Reduce the number of attributes or objects
- Change of scale
  - Cities aggregated into regions, states, countries, etc
- More “stable” data
  - Aggregated data tends to have less variability

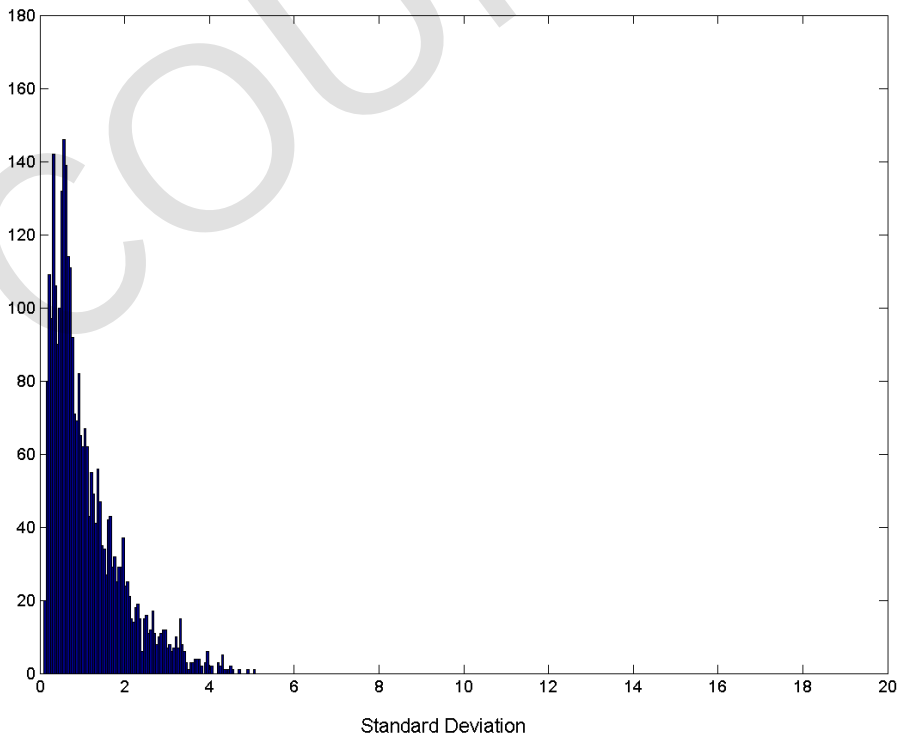
Aggregation

Variation of Precipitation in Australia

1982-1993的降水量，国土按经纬度分成3030个网格



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Sampling

Sampling is the main technique employed for data selection.

- It is often used for both the preliminary investigation of the data and the final data analysis. 数据初步调研与最终分析

Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Sampling …

The key principle for effective sampling is the following:

- Using a sample will work almost as well as using the entire data sets, if the sample is representative
- A sample is representative if it has approximately the same property (of interest) as the original set of data

Types of Sampling

Simple Random Sampling

- There is an equal probability of selecting any particular item

Sampling without replacement （无放回抽样）

- As each item is selected, it is removed from the population

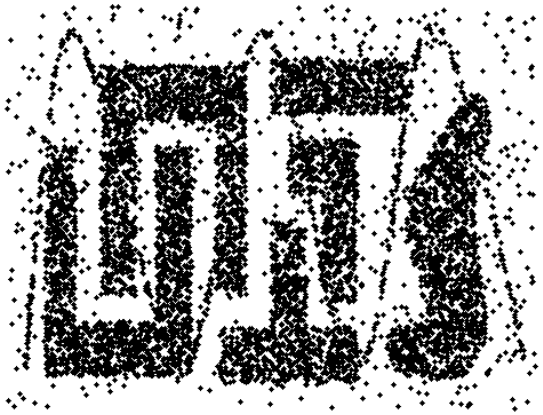
Sampling with replacement （有放回抽样）

- Objects are not removed from the population as they are selected for the sample.
  - In sampling with replacement, the same object can be picked up more than once（每个对象被选中的概率保持不变）

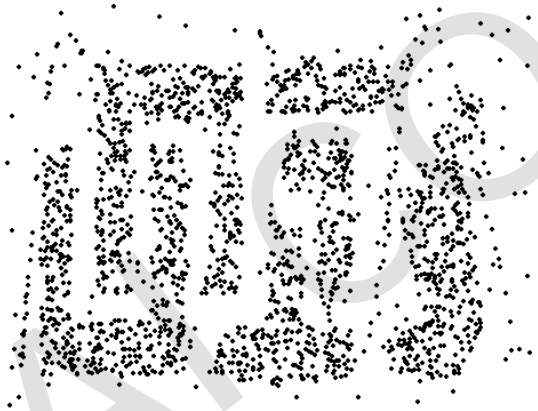
Stratified sampling (分层抽样)

- Split the data into several partitions; then draw random samples from each partition

Sample Size



8000 points



2000 Points

500 Points

Progressive Sampling

Start with a small sample

Increase the sample size

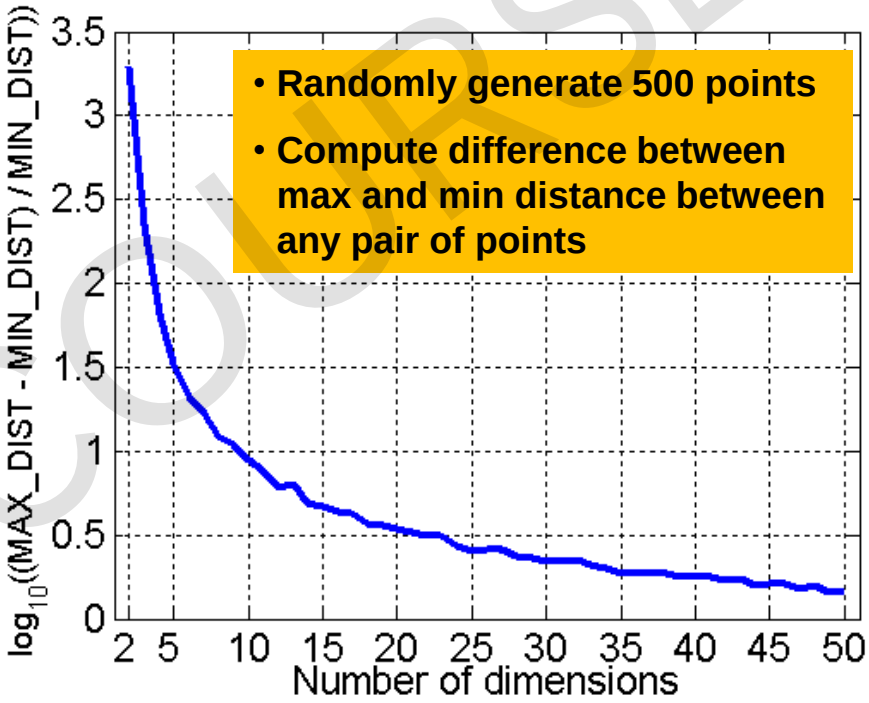
Need to evaluate the sample to judge if it is large enough

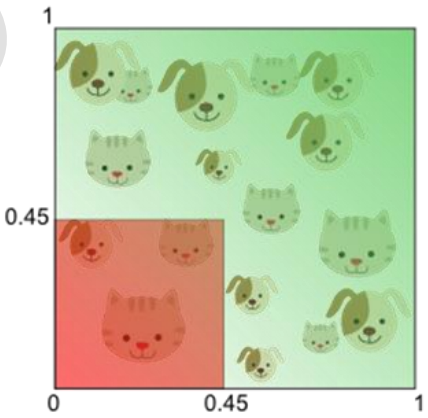
Marginal effect (边际效应)

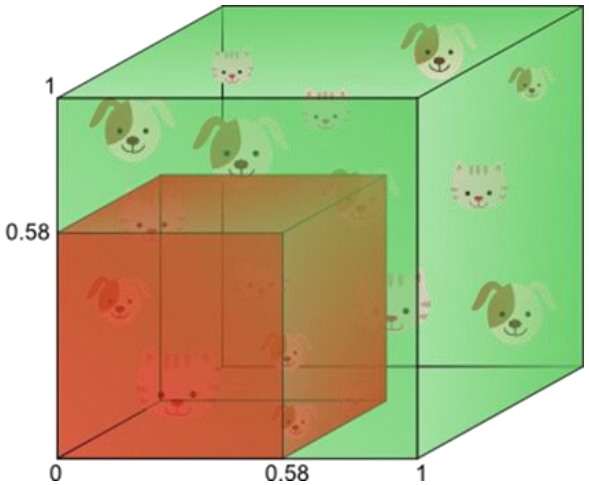
Curse of Dimensionality

When dimensionality increases, data becomes increasingly sparse in the space that it occupies

Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful







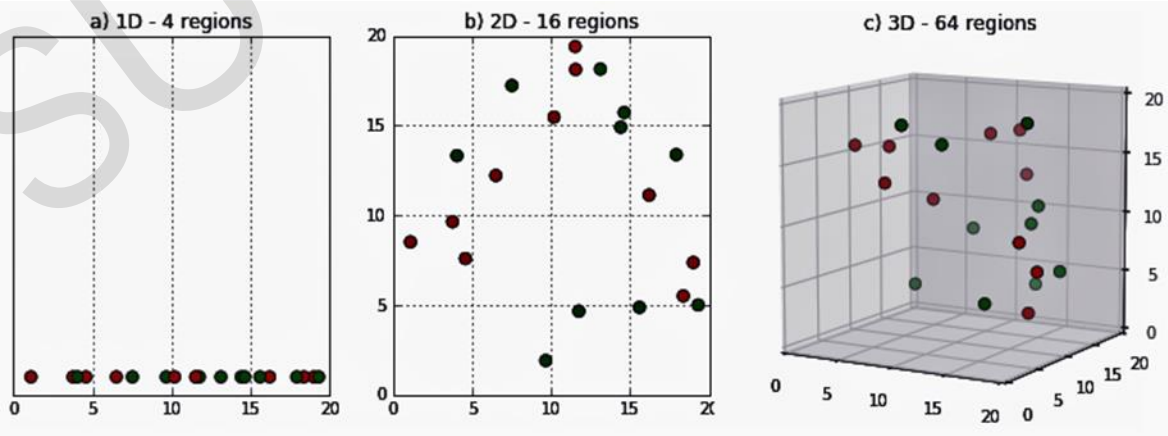
Dimensionality Reduction

Purpose:

- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise

Techniques

- Principle Component Analysis (PCA)
- Singular Value Decomposition (SVD)
- Others: supervised and non-linear techniques



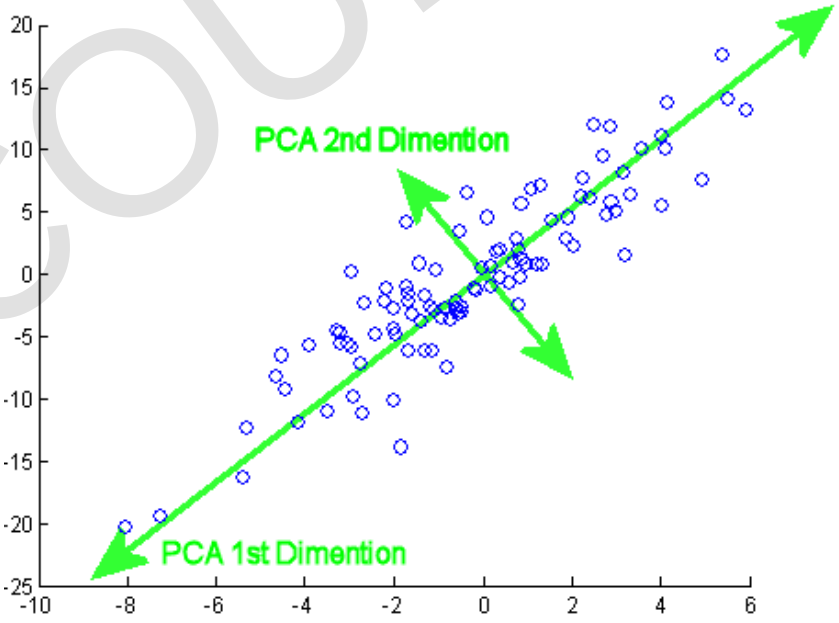
Dimensionality Reduction: PCA

变量之间是有一定的相关关系的

当两个变量之间有一定相关关系时，可以解释为这两个变量的信息有一定的重叠

主成分分析是对于原先提出的所有变量，将重复的变量（关系紧密的变量）删去多余，建立尽可能少的新变量，使得这些新变量是两两不相关的

这些新变量在反映信息方面尽可能保持原有的信息



Discretization

Discretization: Transform a continuous attribute to categorical attribute

The best discretization depends on the algorithm being used

How many categories?

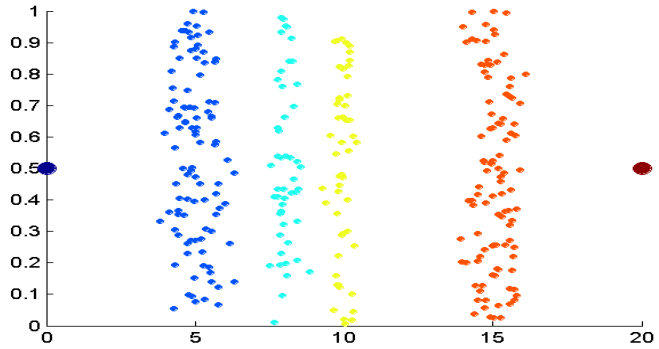
How to map the values of continuous attributes to these categories?

How many split points to choose and where to place them?

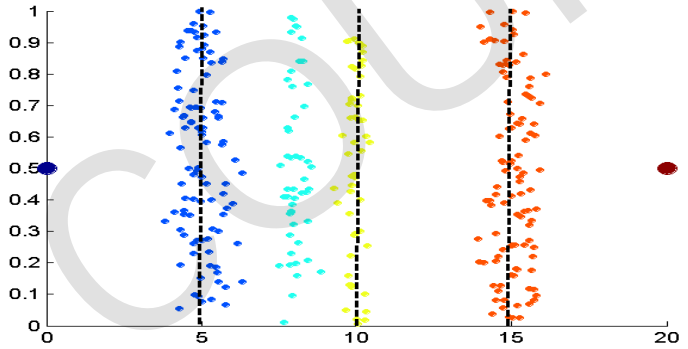
Solutions

- Unsupervised discretization
- Supervised discretization

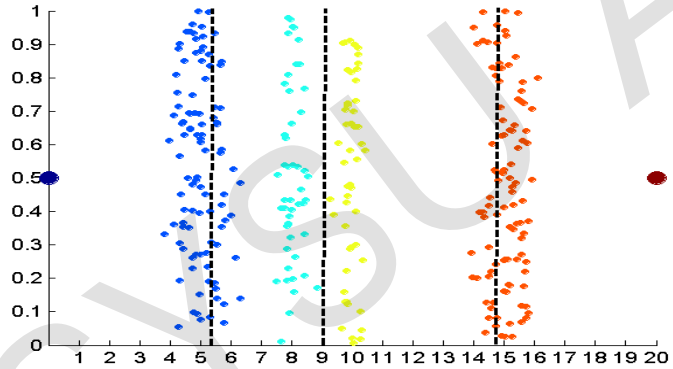
Discretization Without Using Class Labels



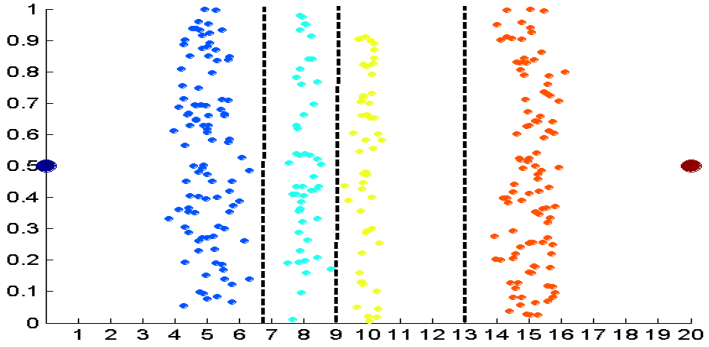
Data



Equal interval width



Equal frequency



K-means

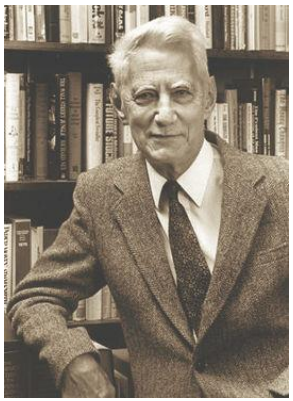
Supervised Discretization：Entropy (熵)

Entropy (熵)

- 熵的概念是由德国物理学家克劳修斯于1865年所提出。熵最初是被用在热力学方面的
- 香农1948年的一篇论文《A Mathematical Theory of Communication》提出了**信息熵**的概念，解决了对信息的量化度量问题，并且以后信息论也被作为一门单独的学科

要搞清楚一件非常不确定的事，就需要了解大量的信息。相反，如果我们对某件事已经有了较多的了解，我们不需要太多的信息就能把它搞清楚。

对于任意一个随机变量X，熵定义如下：“变量的不确定性越大，熵也就越大，把它搞清楚所需要的信息量也就越大。”



香农外传：http://blog.sciencenet.cn/blog-453322-978153.html

Entropy (熵)

- 世界杯谁是冠军?
- 世界杯赛后问一个知道结果的观众“哪支球队是冠军”？ 他不愿意直接告诉我，而要让我猜，并且我每猜一次，他要收一元钱才肯告诉我是否猜对了，那么我需要付给他多少钱才能知道谁是冠军呢?
- 我可以把球队编上号，从 1 到 32， 然后提问： “冠军的球队在 1-16 号中吗?” 假如他告诉我猜对了， 我会接着问： “冠军在 1-8 号中吗?” 假如他告诉我猜错了， 我自然知道冠军队在 9-16 中。 这样最多只需要五次， 我就能知道哪支球队是冠军
- 谁是世界杯冠军这条消息的信息量值五块钱

热情好客的中大学子
正准备迎接远道而来的法国朋友

Entropy (熵)

不需要猜五次就能猜出谁是冠军，巴西、法国、阿根廷这样的球队得冠军的可能性比美国、越南等队大的多。

第一次猜测时不需要把 32 个球队等分成两个组，而可以把少数几个最可能的球队分成一组，把其它队分成另一组。然后我们猜冠军球队是否在那几只热门队中。

重复这样的过程，根据夺冠概率对剩下的候选球队分组，直到找到冠军队。也许三次或四次就猜出结果。

当每个球队夺冠的可能性（概率）不等时，“谁世界杯冠军”的信息量比五比特少。香农指出，它的准确信息量应该是

- “谁是世界杯冠军”的信息量:

$$= -(p1*\log p1 + p2 * \log p2 + \ldots + p32 *\log p32),$$

- p1,…p32是32个球队各自夺冠的概率

课外阅读：《数学之美》第六章“信息的度量与作用”

Supervised Discretization

基于熵的离散化方法

- 最大化区间的纯度

$$e_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$$

首先，需要定义熵（entropy）。设 k 是不同的类标号数，m_i 是某划分的第 i 个区间中值的个数，而 m_{ij} 是区间 i 中类 j 的值的个数。第 i 个区间的熵 e_i 由如下等式给出

$p_{ij} = m_{ij}/m_i$ 是第 i 个区间中类 j 的概率（值的比例）。

类别1=X,		类别2=0	
p11=3/4,	X X	0 0	p12=1/4
p21=1/4,	X 0	0 X	p22=3/4

Supervised Discretization

熵：区间纯度的度量

- 只包含一个类：熵为0
- 包含所有类，并且每类出现的概率相等：熵最大

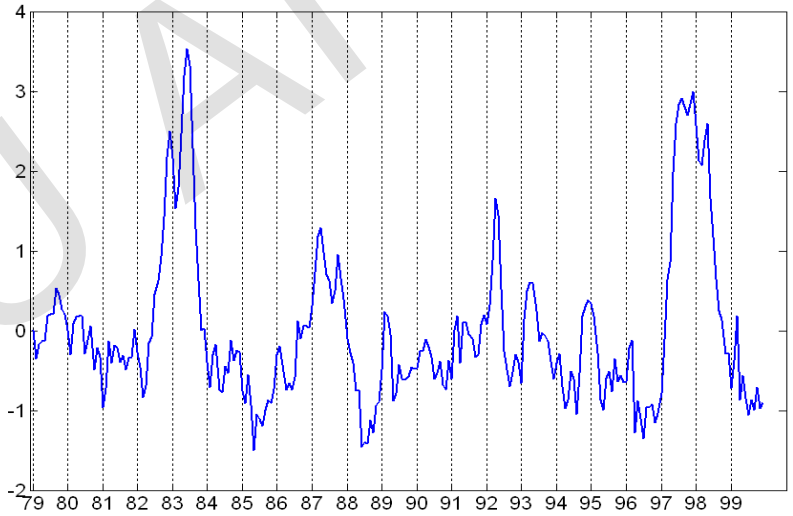
划分连续属性的简单方法：

- 将初始值切分成两部分，让结果区间产生最小的熵
- 然后选取熵最大的区间，重复该过程
- 直到区间数量达到用户指定个数

Attribute Transformation

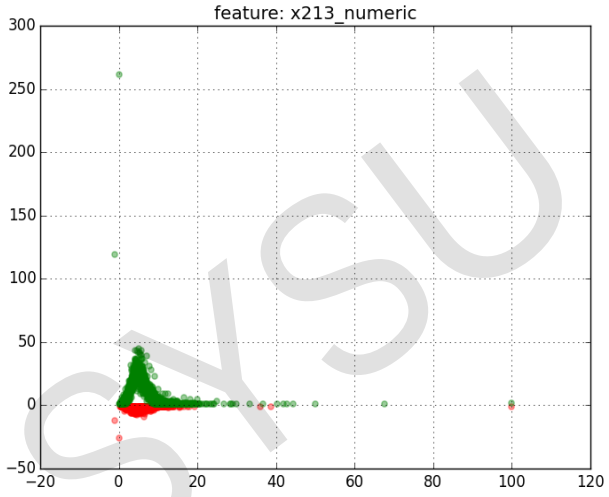
A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values

- Simple functions: x^k, log(x), e^x, |x|
- Standardization and Normalization

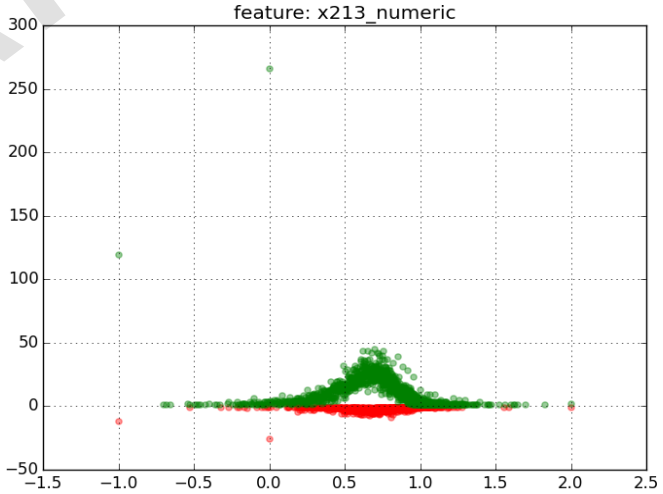


Attribute Transformation

Standardization	**(x-mean)/sd**	
Max-Min	**(x-min)/(max-min)**	**0 - 1**
Sigmod	**1/(1+exp(-x))**	**0 - 1**
Tanh	**(exp(x)-exp(-x))/(exp(x)+exp(-x))**	**-1 - 1**
Log	**log(x)**	
Nesting	**log(Normalization/Max-Min/Sigmod/Tanh)**	

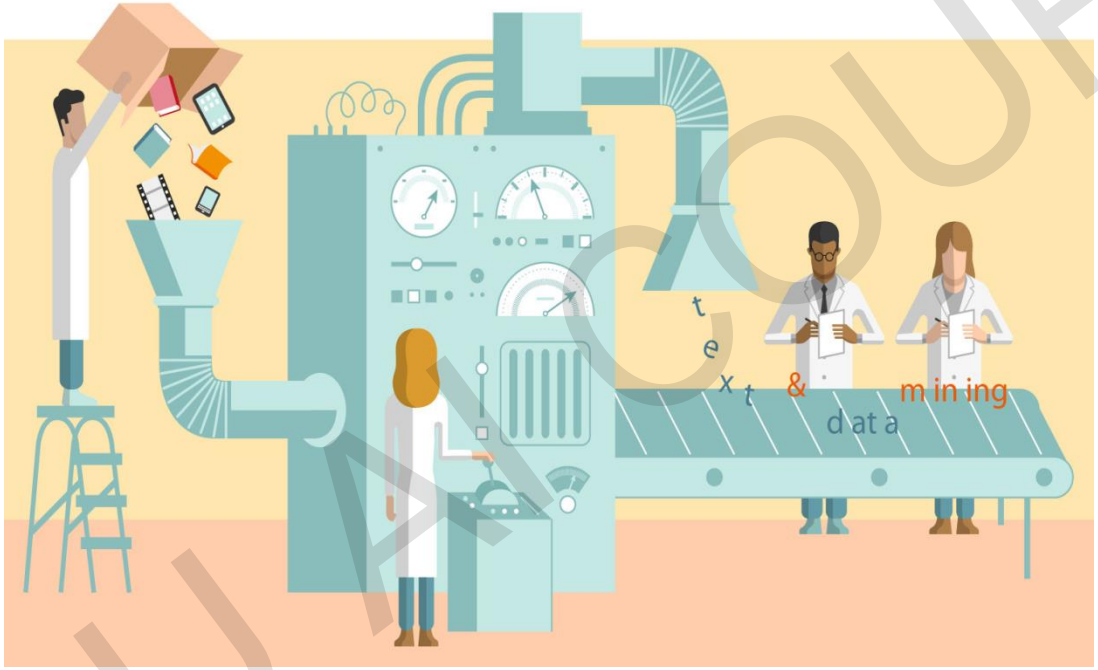


X



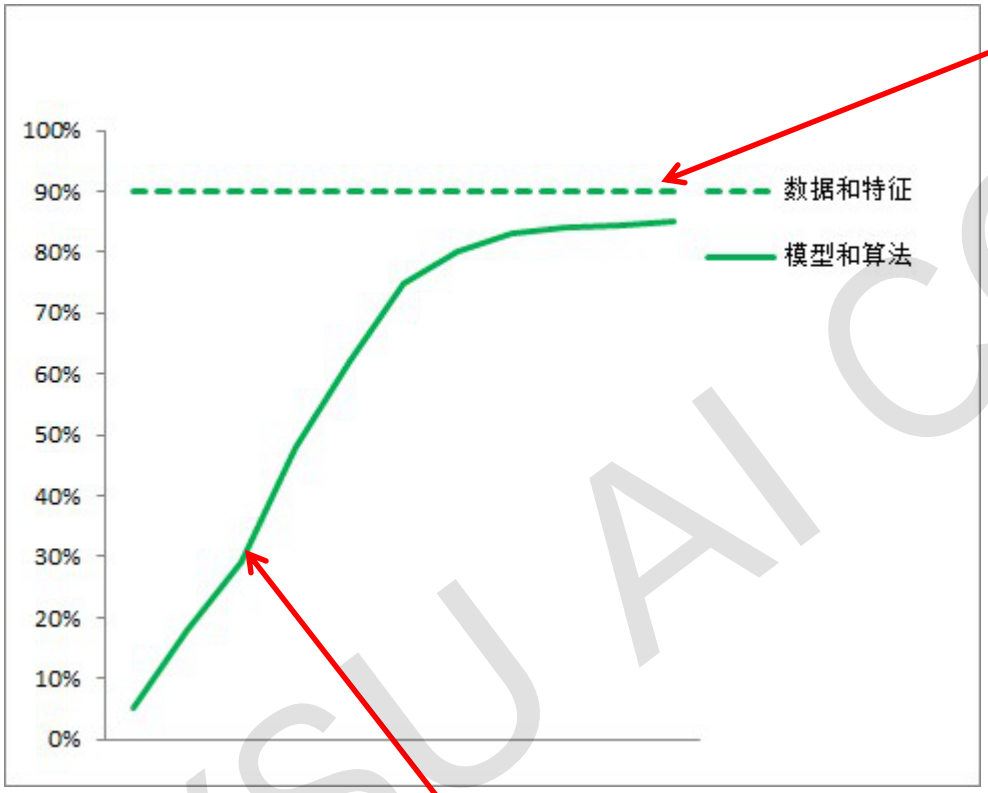
Log(X)

1 引入问题



数据 → 预处理 → **特征工程** → 机器学习算法 → 结果

1 引入问题



数据和**特征**决定了数据挖掘的上限

模型和**算法**只是帮助我们逼近这个上限

2 特征创建（Feature Creation）

由**原始数据**创建新的特征，从而更有效地捕捉原始数据中的重要信息

常用方法

- 特征提取（Feature Extraction）
- 空间映射（Mapping Data to New Space）
- 特征构造（Feature Construction）

2 特征创建：特征提取

特征提取（Feature Extraction）：由原始数据创建新的特征

例子：对图片是否包含人脸进行二分类

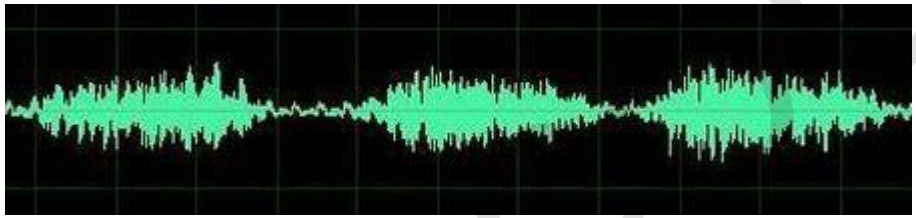
- 原始数据是像素
- 提取人脸相关的边缘特征、区域特征等

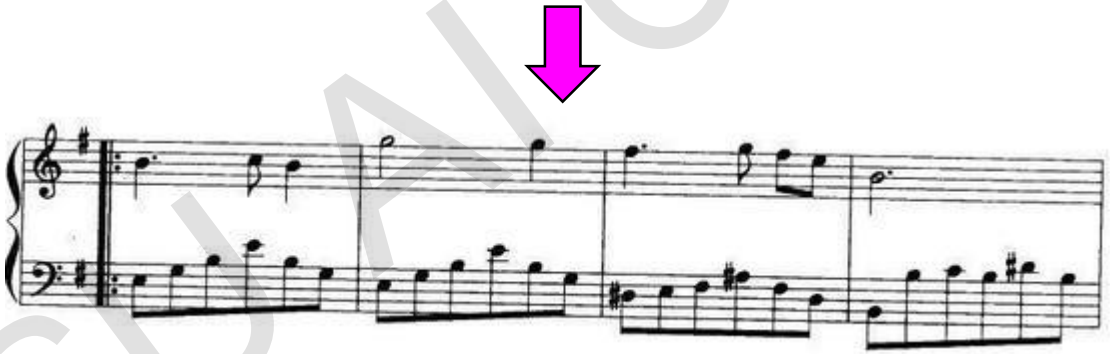
常用的特征提取技术都是针对具体领域的

数据挖掘用于新领域时，需开发新的特征提取方法

2 特征创建：空间映射

将数据进行空间映射，使用不同的视角挖掘数据

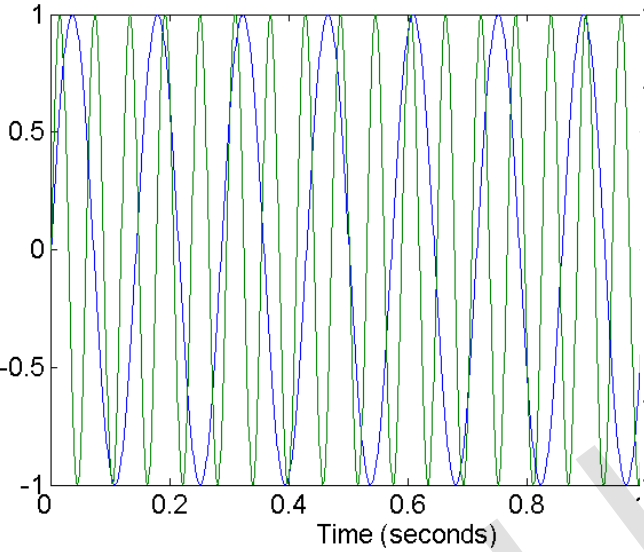




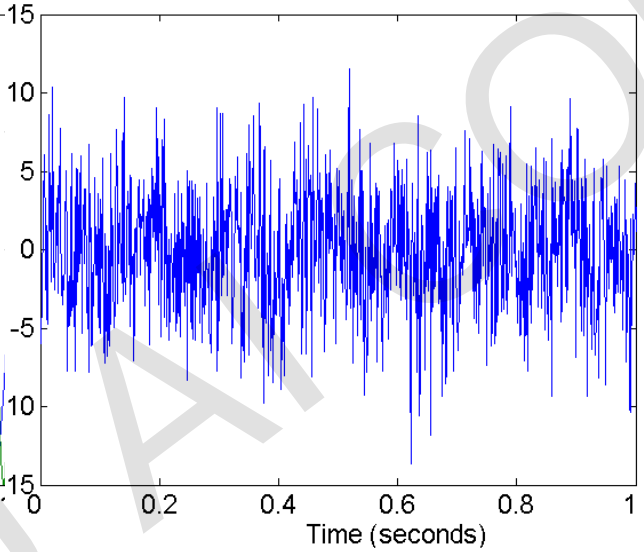
数据空间映射之后，可以更好地提取特征

2 特征创建：空间映射

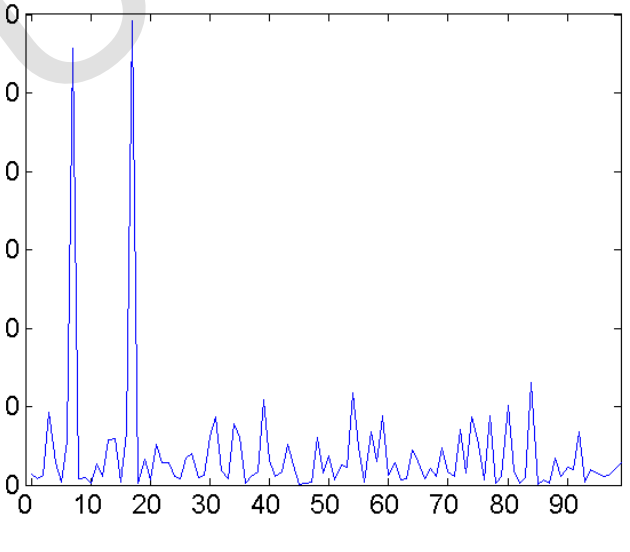
空间映射：傅里叶变换



原始波形（频率为7，17）



原始波形+噪声

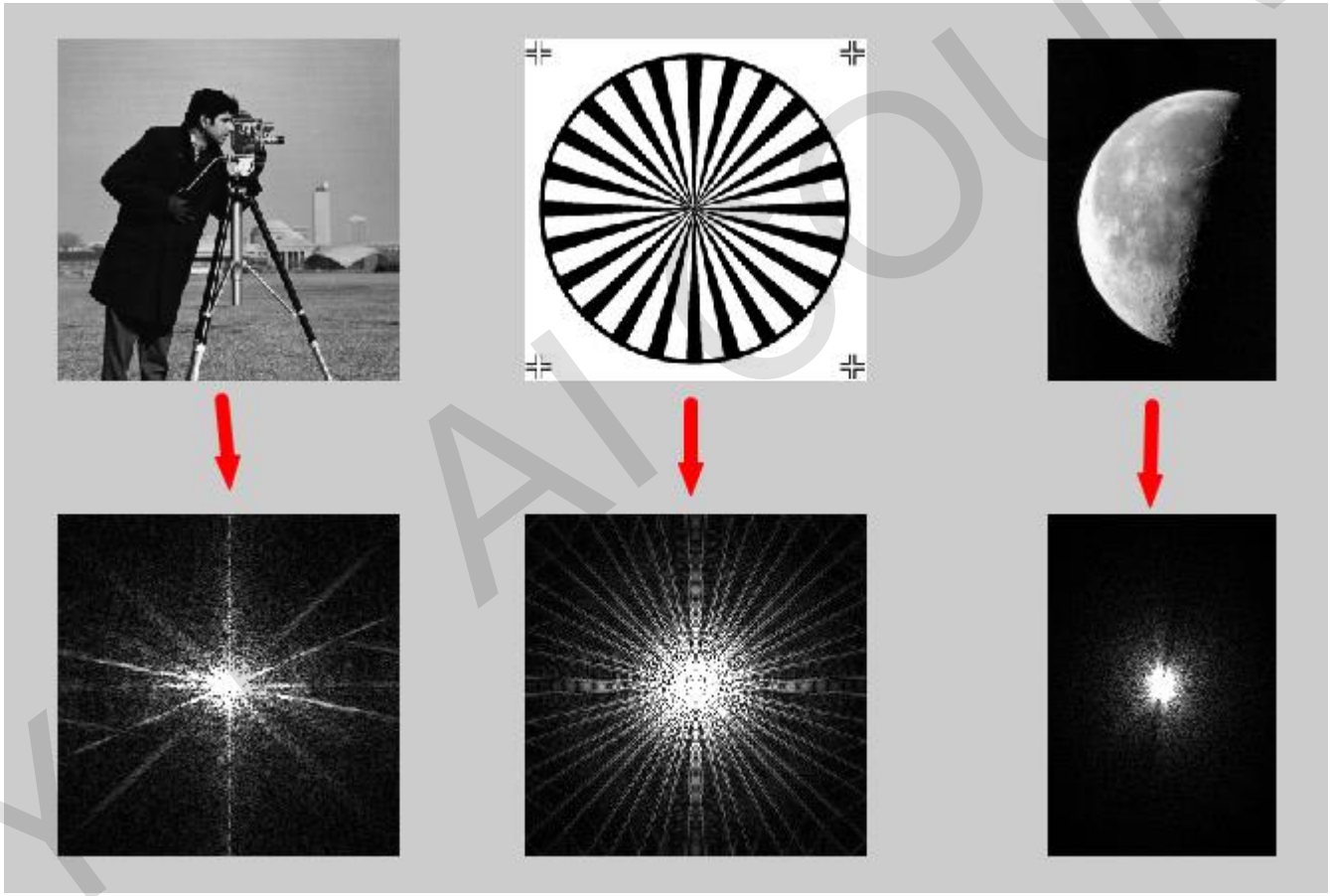


功率频谱

对时间序列实施傅立叶变换，转换成频率信息明显的表示

2 特征创建：空间映射

空间映射：傅里叶变换

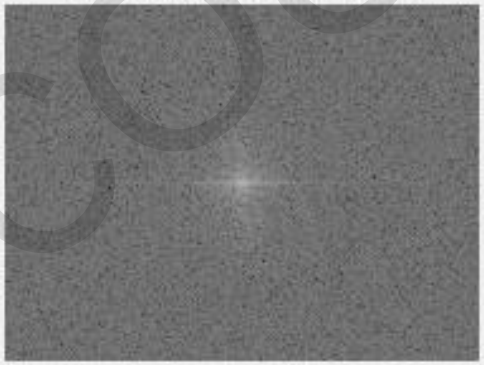


2 特征创建：空间映射

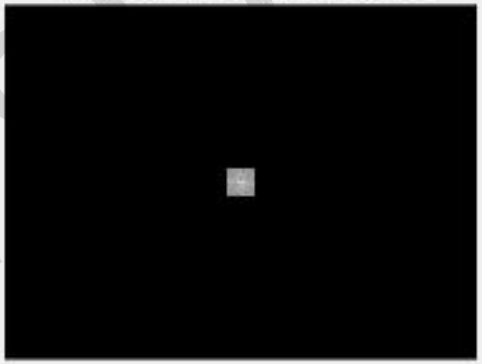
空间映射：傅里叶变换

噪声图像

傅里叶变换后幅度图



去除外围幅值后幅度图



去噪后的图像

2 特征创建：特征构造

特征构造（Feature Construction）：原始特征包含了必要信息，但是形式不适合，因此需由原特征构造新特征

例子：人工制品分类

- 使用不同材料制造：木材、陶土、铜、黄金等
- 希望根据制造材料对它们进行分类
- 原始特征：质量、体积
- 构造的新特征：密度=质量/体积

常用的方法：使用专家的意见构造特征

3 特征二元化（Binarization）

有些数据挖掘算法要求输入是二元属性形式

类别特征包括：

- **无序类别（Categorical）**
- **有序类别（Ordinal）**

3 特征二元化：无序类别

无序类别（Categorical）

分类值	整数值	X_1	X_2	X_3
blue	0	0	0	0
green	1	0	0	1
red	2	0	1	0
black	3	0	1	1
white	4	1	0	0

独热编码(One hot Encoding): 把每个无序特征转化为一个数值向量

分类值	X_1	X_2	X_3	X_4	X_5
blue	1	0	0	0	0
green	0	1	0	0	0
red	0	0	1	0	0
black	0	0	0	1	0
white	0	0	0	0	1

3 特征二元化：有序类别

- 有序类别（Ordinal）

Status
Bad
Normal
Good

→

Vectorization
[1, 0, 0]
[0, 1, 0]
[0, 0, 1]

- 向量表示方法（Multi-hots Encoding）：值之间有顺序的含义

当status特征向量输入模型时，对于status这个类别特征模型会学习出w1,w2,w3三个权重，如果是good的话将会是w1w2w3的叠加，如果取值为bad的话只有w1，从而体现出有序性

Status
Bad
Normal
Good

→

Vectorization
[1, 0, 0]
[1, 1, 0]
[1, 1, 1]

$$\sum w_i x_i$$

3 特征二元化：特征组合

基本特征仅仅是真实特征分布在低维空间的映射，不足以描述真实分布，加入组合特征是为了在更高维空间拟合真实分布，使得预测更准确

线性模型对于非线性关系缺乏准确刻画，特征组合正好可以加入非线性表达，增强模型的表达能力

基本特征可以认为是用于全局建模，组合特征更加精细，是个性化建模，所以基本特征+组合特征兼顾了全局和个性化

可以通过笛卡尔乘积的方式来组合2个或多个特征

例如有两个类别特征color和light，分别取值red, green, blue和on, off。两个特征可以分别转化为3维和2维的向量，对他们做笛卡尔乘积转化后可以组合出6维的向量

X	on	off
red		
green		
blue		

样本	color=red& light=on	color=red& light=off	color=green& light=on	color=green& light=off	color=blue& light=on	color=blue& light=off
1	1	0	0	0	0	0
2	0	1	0	0	0	0
3	...	...	...	...	...	...

Feature Creation

连续特征（continuous features）

Student ID	Age	Weight(kg)	Height(cm)
0	18	56	174
1	21	61	176
2	25	58	168

连续特征处理：

- MinMax Scale: $z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$
- Log transform: $z_i = \log(1 + x_i)$
- 描述统计特征：min max mean median mode std var …

3 例子

一学生属性如下，如何转化为能够输入模型的特征向量?

Student ID	age	weight	height	gender	status
0	18	65	178	M	Bad

连续特征age、weight、height 进行MinMax scale：[0，0.29，1]

无序特征gender进行One hot Encoding: [1, 0]

有序特征status向量化：[1, 0, 0]

最终表示学生0的特征向量为：[0, 0.29, 1, 1, 0, 1, 0, 0]

3 时间信息

时间信息包含有丰富的数据意义

- 例如：2017-10-01 16:38:43

Year	Month	Day of Month	Week	Holiday
2017	10	1	Sunday	Yes

Season	Hour Type	Day of Holiday	Hour of Day
Autumn	Afternoon	1	16/24

3 地理信息

地理位置信息包含有丰富的数据意义

- 例如：广东省广州市番禺区大学城

Province	City	Area	Distribution	City-level
广东	广州	番禺区	东南	1

Longitude	Latitude	Area Type	Temperature type
113.23	23.16	学校	Hot

4 案例分析：客户用电异常行为分析

国家电网提供了88436名用户，2014~2016年每天的用电数据

基于用户的用电数据， 挖掘窃电用户行为特征，识别窃电用户，这是一个二分类问题

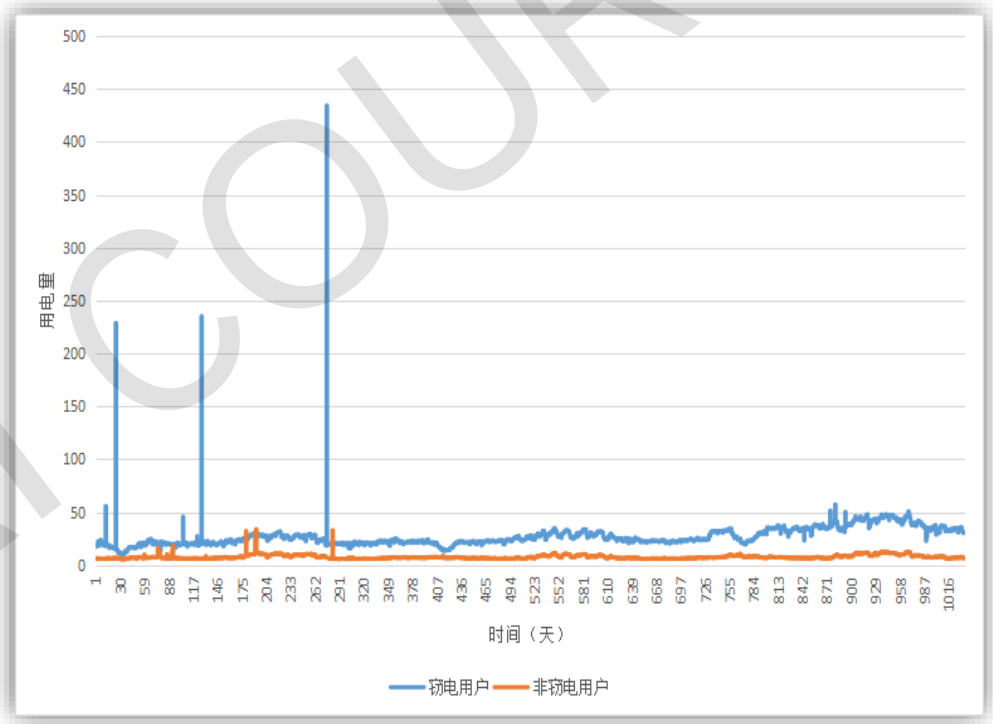
4 案例分析：客户用电异常行为分析

观察问题

- 窃电用户平均用电度数偏高
- 窃电用户用电量瞬时波幅偏高

特征工程

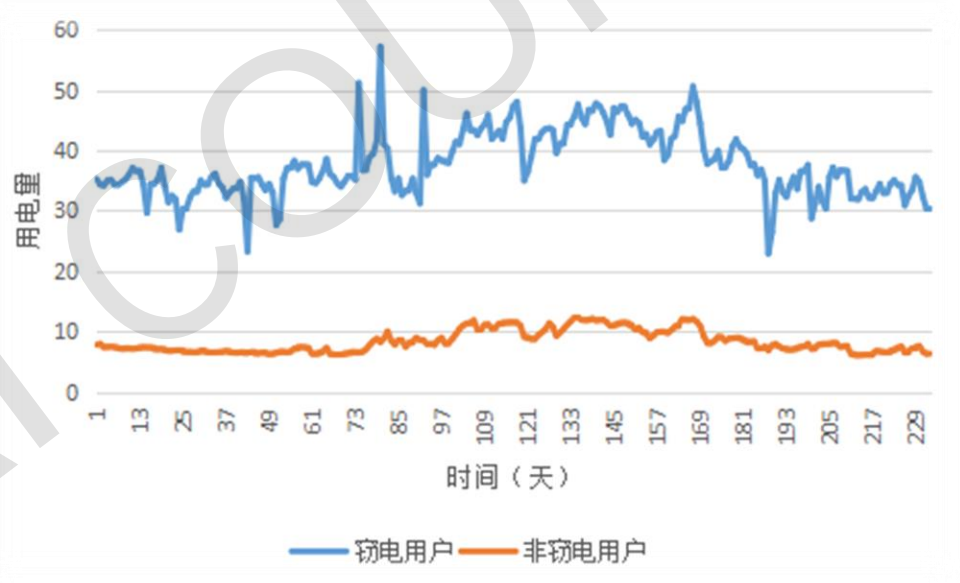
- 用户用电度数**最大值、均值、中位数**



4 案例分析：客户用电异常行为分析

观察问题

- 窃电用户用电量的波动性比较大
- 非窃电用户用电的稳定性比较强



特征工程

- 用户用电度数**标准差、四分位数、异常值的个数**
- 稳定性衡量由前后等长一段时间的数据**相似度**计算

4 案例分析：客户用电异常行为分析

空间映射

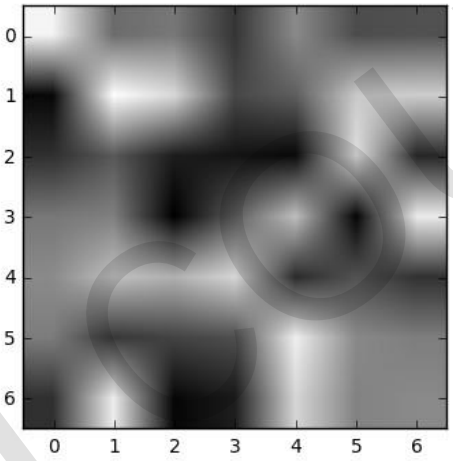
- 用户用电数据由一维向量转换成二维矩阵
- 二维矩阵转换成灰度图

观察问题

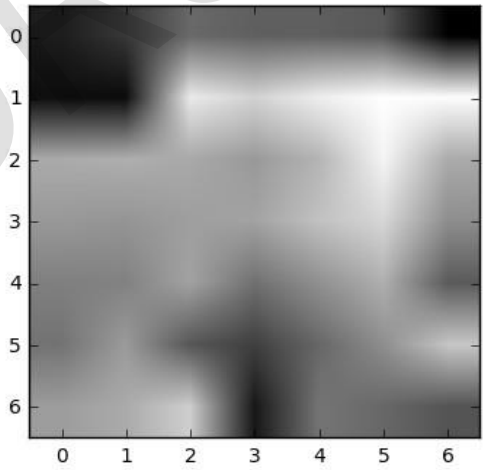
- 窃电用户的图形黑白相间
- 非窃电用户的图形相对空旷

特征工程

- 用卷积神经网络对图像进行自动化特征提取



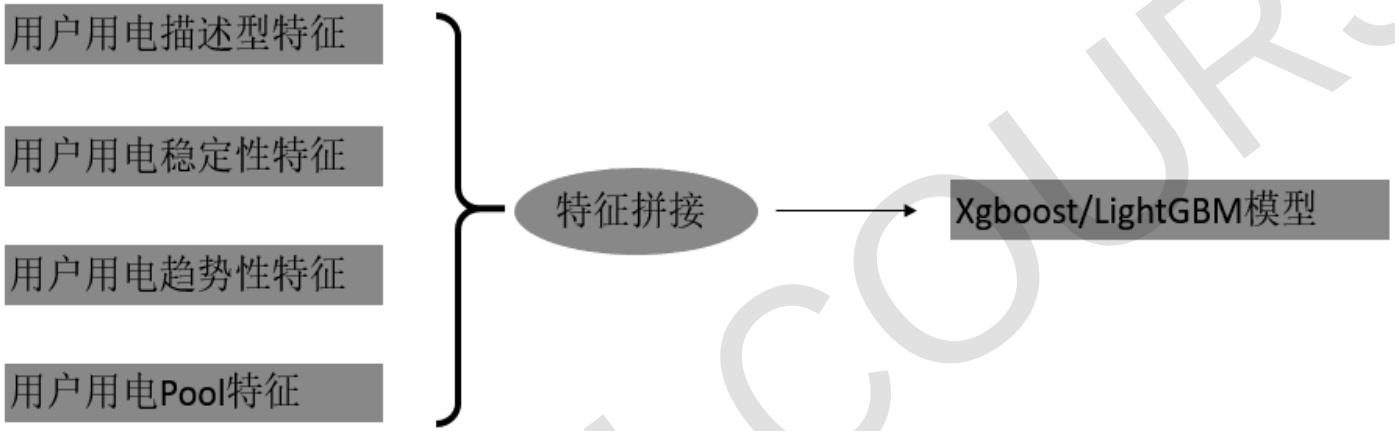
窃电用户



非窃电用户

将7周用电量数据转换成7*7灰度图矩阵

4 案例分析：客户用电异常行为分析



初赛

A榜 B榜

排名	队伍名称	最高得分(B)
1	TNT_000_	0.95459
2	我们又回来了-美林数据	0.95187
3	Top	0.95164

复赛

A榜 B榜

排名	队伍名称	最高得分(B)
1	我们又回来了-美林数据	0.94274
2	隐马尔可夫联盟	0.93373
3	打酱油'拎壶冲	0.92871
4	TNT_000_	0.92564

Feature Subset Selection

Another way to reduce dimensionality of data

Redundant features

- duplicate much or all of the information contained in one or more other attributes
- Example: purchase price of a product and the amount of sales tax paid

Irrelevant features

- contain no information that is useful for the data mining task at hand
- Example: students' ID is often irrelevant to the task of predicting students' GPA

Feature Subset Selection

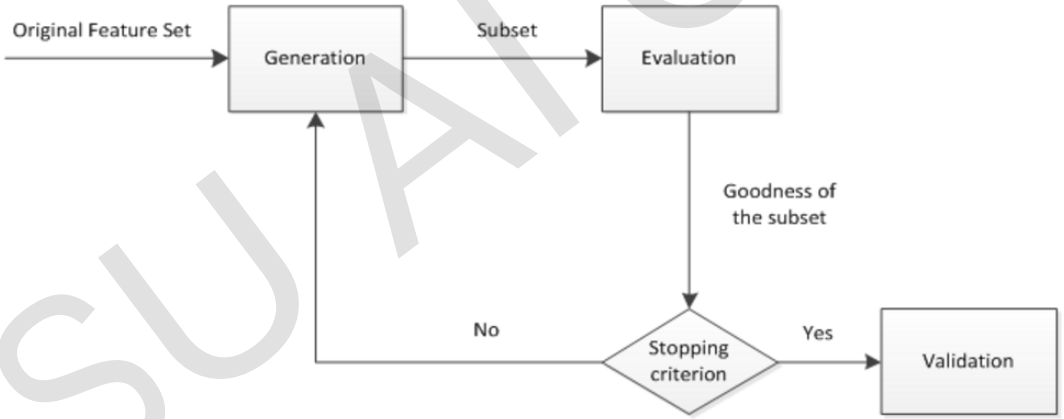
Techniques:

- Brute-force approaches（暴力）:
  - Try all possible feature subsets as input to data mining algorithm
- Embedded approaches (嵌入):
  - Feature selection occurs naturally as part of the data mining algorithm
- Filter approaches（过滤）:
  - Features are selected before data mining algorithm is run
- Wrapper approaches（包装）:
  - Use the data mining algorithm as a black box to find best subset of attributes

Feature Subset Selection

Brute-force approaches:

(1) 产生过程
(2) 评价函数
(3) 停止准则
(4) 验证过程



Feature Subset Selection

- Embedded approaches
  - 学习器自动选择特征
  - 正则化 （L1、L2）：正则化主要是将L1/L2范数作为惩罚项添加到损失函数上，由于正则项非零，这就迫使那些弱的特征所对应的系数变成0。
  - 决策树 （熵、信息增益）：决策树算法在树增长过程的每个递归步都必须选择一个特征，将样本集划分成较小的子集，选择特征的依据通常是划分后子节点的纯度，划分后子节点越纯，则说明划分效果越好，可见决策树生成的过程也就是特征选择的过程
  - 深度学习：从神经网络的中间层的某一层输出可作为特征

Feature Subset Selection

- Filter approaches：
- 思路：特征和目标变量之间的关联
  - 统计检验，如卡方检验、t检验
  - 相关系数，如皮尔森相关系数、
  - 互信息和最大信息系数（MIC）

	适用范围	是否标准化	计算复杂度	鲁棒性
Pearson	线性数据	是	低	低
spearman	线性、简单单调非线性数据	是	低	中等
Kendall	线性、简单单调非线性数据	是	低	中等
阈值相关	线性、非线性数据	是	高	高
最大相关系数	线性、非线性数据	是	高	中等
相位同步相关	时变序列	是	中等	中等
距离相关	线性、非线性数据	是	中等	高
核密度估计(KDE)	线性、非线性数据	否	高	高
k-最邻近距离(KNN)	线性、非线性数据	否	高	高
MIC	线性、非线性数据	是	低	高

Feature Subset Selection

- Wrapper approaches：
- 思路：通过模型选择特征
  - 构建单个特征的模型，通过模型的准确性为特征排序
  - 训练能够对特征打分的预选模型,如RandomForest、Logistic Regression

Feature Subset Selection

医学数据集:

Leukemia	7129×72
Colon	2000×62

特征样例:

sp. / Gene	Sample 1 (Cancer)	Sample 2 (Normal)	……	Sample k
Gene 1	29	19	……	16
Gene 2	5	17	……	40
……	……	……	……	……
Gene n	13	8	……	2

Feature Subset Selection

特征选择结果：

Leukemia（SVM)

Number of genes	Train accuracy	Test accuracy
100	100	99.31
50	100	98.276
34	100	99.31
20	100	98.621
10	100	98.621
8	100	96.552
5	100	95.172
3	100	92.759
1	92.093	78.966

Feature Subset Selection

特征选择结果：

Colon（SVM)

Number of genes	Train accuracy	Test accuracy
100	100	80.4
50	100	80.8
33	100	82
20	100	79.2
10	100	78.8
8	100	77.6
5	99.189	75.6
3	95.405	77.6
1	80	71.6

选择有效的特征能提高预测准确性！

Similarity and Dissimilarity

Similarity

- Numerical measure of how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range [0,1]

Dissimilarity

- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

Similarity/Dissimilarity for Simple Attributes

***p* and *q* are the attribute values for two data objects.**

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance

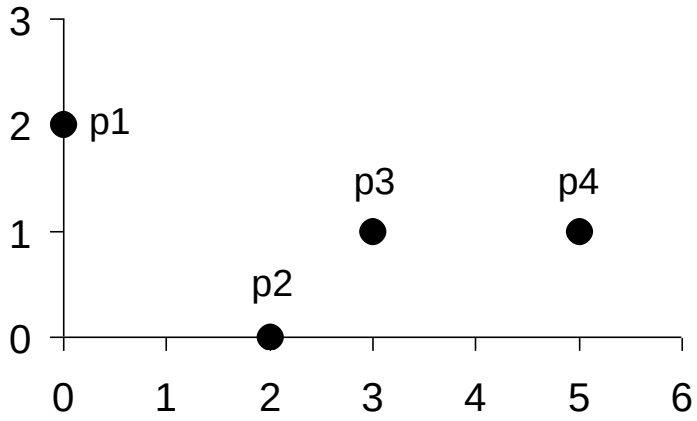
Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance（闵可夫斯基距离）

Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Minkowski Distance: Examples

r = 1. City block (Manhattan, taxicab, L_1 norm) distance.

- A common example of this is the Hamming distance（汉明距离）, which is just the number of bits that are different between two binary vectors

r = 2. Euclidean distance

$r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.

- This is the maximum difference between any component of the vectors

Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Similarity Between Binary Vectors

Common situation is that objects, *p* and *q*, have only binary attributes

Compute similarities using the following quantities

M_{01} = the number of attributes where p was 0 and q was 1

M_{10} = the number of attributes where p was 1 and q was 0

M_{00} = the number of attributes where p was 0 and q was 0

M_{11} = the number of attributes where p was 1 and q was 1

Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes

$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

J = number of 11 matches / number of not-both-zero attributes values

$= (M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)

M_{10} = 1 (the number of attributes where p was 1 and q was 0)

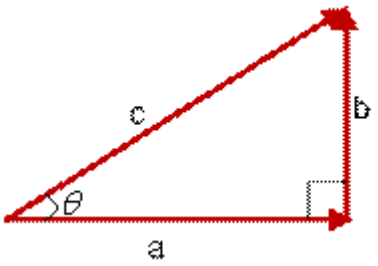
M_{00} = 7 (the number of attributes where p was 0 and q was 0)

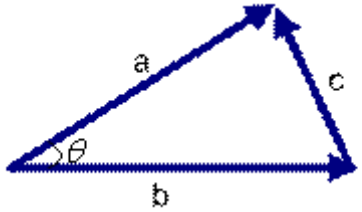
M_{11} = 0 (the number of attributes where p was 1 and q was 1)

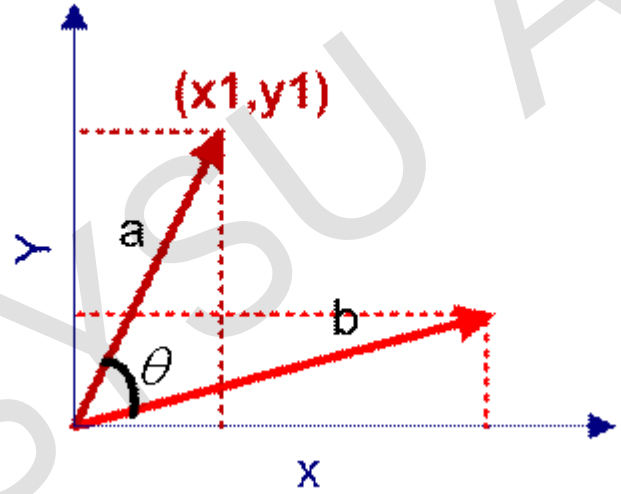
$SMC = (M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ **= (0+7) / (2+1+0+7) = 0.7**

$J = (M_{11}) / (M_{01} + M_{10} + M_{11})$ **= 0 / (2 + 1 + 0) = 0**

Cosine Similarity







$$\cos(\theta) = \frac{a^2 + b^2 - c^2}{2ab}$$

$$= \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

$$= \frac{(x_1, y_1) \bullet (x_2, y_2)}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

$$\cos(\theta) = \frac{\mathrm{a} \bullet \mathrm{b}}{||\,\mathrm{a}\,|| \times ||\,\mathrm{b}\,||}$$

Cosine Similarity

If d_1 and d_2 are two document vectors, then

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

where $\bullet$ indicates vector dot product and $\| d \|$ is the length of vector d.

Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$\|d_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\|d_2\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$$\cos(d_1, d_2) = .3150$$

- 将向量根据坐标值，绘制到向量空间中。
- 求得夹角，得出夹角的余弦值，用于代表两个向量的相似性
- 夹角越小，余弦值越接近于1，它们的**方向**更加吻合，则越相似

Correlation（PCC皮尔森相关性）

Correlation measures the linear relationship between objects

To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = \frac{1}{n} p' \bullet q'$$

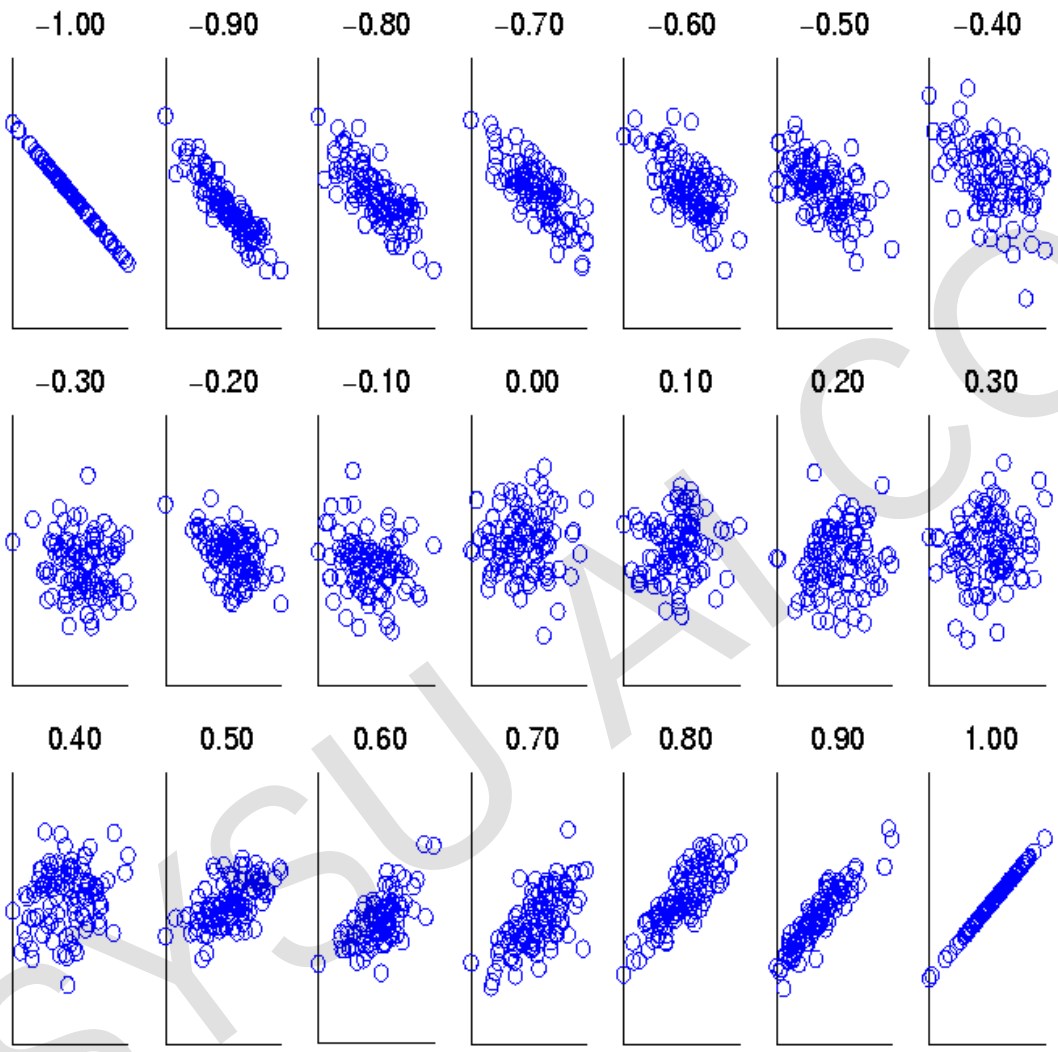
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

先标准化，再内积；

Cosine vs PCC：标准化的过程不同

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Thanks