92,07,1 Data mining جلسه 3

Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining

by

Tan, Steinbach, Kumar

این جلسه درمورد داده صحبت میکنیم. Data یک کلمه جمع است، Data mining یعنی حفاری داده‌ها. داده‌ها چه نوعی دارند؟ چه کارهایی میشه با این داده‌ها کرد؟ چه طوری میشه این داده‌ها را اندازه گیری کرد؟ چطوری میشه اینها را جمع آوری کرد؟! چه مشکلاتی ممکن است در حین جمع آوری برای این داده‌ها بوجود آید؟ چه محدودیتهایی برای جمع آوری داریم؟!

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 (columns) / Objects (rows)

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

داده هایی که ما باهاش سروکار داریم همه از یک نوع نیستند. همه یک جنس نیستند. اینها اختلافات زیادی با هم دارند. بخاطر اینکه نمی توانند نوع داده ها را مشخص کنند، می گویند ما باید بدانیم که چه نوع داده را شناختیم؟ چه مقیاسی اندازه گیری کرد، براساس مقیاس اندازه گیری ما می آییم داده ها را از هم جدا میکنیم. درواقع ما بخاطر اینکه نتوانیم به خصیصه این نسبت دهیم، آن را به صورت یک عدد یا حتی یک سمبل (یک علامت) میتوانیم بهش نسبت دهیم، براساس این عدد یا علامتی که بهش نسبت می دهیم باید قاعدتاً براساس آن مقیاس اندازه گیری باشد که ما آن خصیصه را بهش

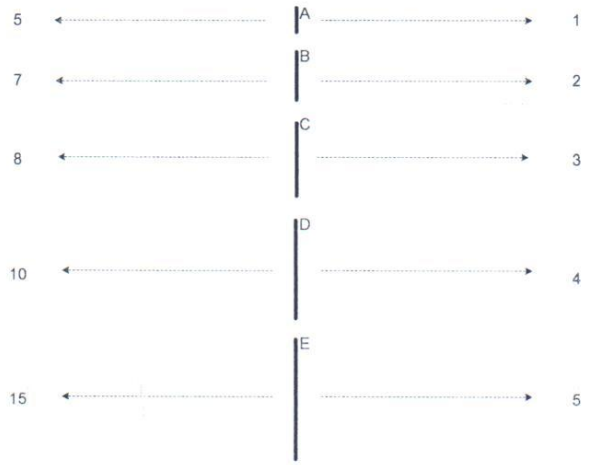
نسبت می دهیم و نوع داده ها مشخص میشود. مثال زده مثلاً طول و ما میتوانیم طول را اندازه گیری کنیم با مقیاس و واحد اندازه گیری
میتوانیم مشخص اش کنیم ولی اگر بخواهیم به صورت رسمی این Attribute ها را صفت ها را مشخص کنیم ما میتوانیم ۴ تا نوع
صفت را اندازه گیری کنیم که می توانیم به آن ها بگوییم: ۴ تا نوع مقیاس اندازه گیری داریم ۱- Nominal ۲- Ordinal
۳- Interval ۴- Ratio

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.



مثال اسمی = مثلاً گروه خونی هیچ برتری نسبت به همدیگر ندارند، یک اسم است به یک گروه خونی نسبت می دهیم به خاطر همین بهش میگویند که Nominal است.

مثال ترتیبی = مثلاً میگوییم رتبه اول، دوم، سوم / رئیس، معاون، کارمند / ... اینها یک ترتیبی بین شان برقرار است که به اینها میگوییم داده های Ordinal.

Types of Attributes

- There are different types of attributes
  - Nominal – اسمی
    - Examples: ID numbers, eye color, zip codes
  - Ordinal – ترتیبی
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval – فاصله ای
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio – نسبتی
    - Examples: temperature in Kelvin, length, time, counts

بعضی از داده ها نسبت به فاصله آن ها مقیاس تغییر می کند مثل: مقیاس سلسیوس / یا اگر بخواهیم سلسیوس را به فارنهایت تبدیل کنیم یا تبدیل فاصله ای بهش می گیم.

نسبتی flexiable ترین نوع مقیاس اندازه گیری است. اگر می خواهیم یک صفت را اندازه گیری کنیم، اگر بتوانیم براساس مقیاس نسبتی اندازه گیری اش کنیم خیلی بهتر است مثلاً: درجه حرارت براساس کلوین برعکس سلسیوس که Ratio نیست ولی کلوین این چنین است یا فرض کنید زمان نسبتی است، طول هم همین طور. اکثر مقیاس هایی که با آن ها سروکار داریم، همه نسبتی هستند. خوبی نسبتی ها این است که خیلی از اعمال ریاضی بر روی آن ها انجام می شود. مقیاس های اسمی محدودترین هستند، فقط میتوانیم برایش بگوییم این دو تا یکی هستند یا خیر.

© Tan,Steinbach, Kumar   Introduction to Data Mining   4/18/2004   5

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

© Tan,Steinbach, Kumar   Introduction to Data Mining   4/18/2004   6

در مقیاس ترتیبی میتوانیم کوچکتر یا بزرگتر بودنشان را تعیین کنیم، فاصله ای میتوانیم با هم جمع یا تفریق کنیم و نسبتی تمام کارهایی که گفتیم را میتوان برایش انجام داد یعنی ضرب و تقسیم هم برقرار است. سوال: اگر داده ها Nominal باشند میتوان بر روی آن ها کارهایی انجام داد که به درد می خورند؟ بله، خیلی از کارهایی را که با داده های فاصله ای یا نسبتی اندازه گیری میکنند میتوانیم آن کارها را روی داده های Nominal انجام دهیم مثلاً ما وقتی یک داده با یک رنگ هر دوشان را مقایسه کنیم (مثلاً = آن دو داده ای است که ... را دارد) تمام داده ها با انواع مختلف

را می‌توان مد را برایش محاسبه کرد، entropy را می‌توان برایش محاسبه کرد، می‌توانیم آنها را در یک جدول توافقی، جداول خویشاوندی (contingency) خلاصه کنیم و correlation ها را یا ارتباط آنها با هم را با هم بسنجیم و این خیلی خوب است بین داده‌های Nominal. همچنین هم داده‌هایی نیستند که نشود روی‌شان کاری انجام داد. خیلی از کارهایی که ما روی داده‌های نسبتی انجام می‌دهیم می‌توانیم روی داده‌ها nominal نیز انجام دهیم.

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 (درصد), rank correlation (ضریب همبستگی), 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

entropy یک معیاری است که می‌آید تشابه و عدم تشابه یا واریانس را به نحوی به ما منتقل می‌کند. خیلی مفهوم مهمی در نظریه اطلاعات است.
می‌توان از entropy در clustering به عنوان یک فاصله در نظر می‌گیریم، با هاش می‌آییم داده‌ها را خوشه‌بندی می‌کنیم برای داده‌های nominal.
در Ordinal علاوه بر مد ما می‌توانیم median حساب کنیم (دهک اول، دهک دوم و... چارک سوم و... هشتک دوم و...) میانه یعنی داده‌ها را به دو قسمت تقسیم کنیم. در مورد Nominal شاید نشه median حساب کرد. اما برای Ordinal هم می‌شه. همچنین می‌توانیم Rank همبستگی داده‌ها را با توجه اصل محاسبه کرد. همچنین می‌توانیم در مورد قرار گرفتن داده‌ها یک اطلاعاتی کسب کنیم به نحوی می‌گویند Run tests. خیلی از تست‌های آماری که در آمار بحث می‌شود برای داده‌های ordinal هم می‌شه در موردشان صحبت کرد.

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	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	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*	Length can be measured in meters or feet.

(17)

داده‌ها اگر ترتیب داشته باشند، ما می‌توانیم آنها را به عنوان یک دنباله sequence در نظر بگیریم و بر اساس این که اولی از دومی بزرگتر است، داده‌هایی که شبیه به هم هستند را نسبت به این‌ها اعداد صفر و یک نسبت بدهیم، به صورت یک sequence binary درش بیاوریم، این sequence binary را می‌توان تست کرد که تصادفی است یا نه.
در Interval می‌توانیم mean بگیریم، standard deviation بگیریم، اینها دیگر می‌توانیم Pearson's correlation حساب کنیم، جایی...

و میتوانیم بصورت آماری خیلی از تستهای مثل F و t را براش به کار ببریم.

برای داده های نسبتی میتوانیم پایمان را از این مباحثی که گفتیم فراتر ببریم به جای اینکه میانگین حساب کنیم، geometric mean میگیریم (مثلاً وقتی میخواهند از سرعت دو تا ماشین میانگین بگیرند، نه میانگین معمولی گرفته، geometric mean به درد میخورد)، harmonic mean بگیریم همین طور و... پس دستمان خیلی باز است. دقت کنید که برای ratio تمام کارهای بالا را میشود انجام داد.

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.

* ما میتوانیم روی داده هایی که با مقیاس های مختلف اندازه گیری شده اند یک تبدیل بزنیم، مثلاً برای داده های اسمی یک جابجایی شان را در نظر بگیریم به عنوان یک تبدیل، برای داده های ترتیبی می توانیم یک تابعی روی ایشان به کار ببریم که آن تابع یکنوا باشد (صعودی یا نزولی). برای Interval میتوانیم یک تبدیل مبدأ مقیاس بزنیم و برای ratio میتوانیم کشش کم و زیاد کنیم.

* این صفاتی را که ما گفتیم به دو بخش تقسیم میکنیم یا Discrete هستند یا Continuous.

* پس صفات ۴ تا نوع شدند اما میتوان آنها را به دو بخش عمده تقسیم کرد: Discrete و Continuous

* معمولاً متغیرهای تصادفی اسمی Discrete هستند و به ندرت میتوان continous پیدا کرد و در مورد نسبتی بیشتر continuous هستند.

Types of data sets

- Record
  - Data Matrix (ماتریس)
  - Document Data (داکیومنت)
  - Transaction Data (همان Transaction کارت بانکی و فاکتور باشد)
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered (مثال)
  - Spatial Data ← داده های مکانی
  - Temporal Data ← داده های زمانی (سری های زمانی)
  - Sequential Data ← داده های سری، خط تولید و ...
  - Genetic Sequence Data

* Discrete یعنی چه؟ یعنی تعداد شمارا (اعداد طبیعی N است) یک تناسب را آنقدر برقرار میکنیم تا به عدد ۶ برسیم، ممکن است دفعه اول بیاید یا ممکن است نیاید. اگر کل فضای نمونه همچنین آزمایش را بنویسیم تا بی نهایت ادامه دارد اما تعداد اعضایش شمارا است، پس یک مجموعه شمارا است. پس Discrete گسسته (منظور از آن این است که تعداد اعضای مجموعه را بتوان با انگشتان دست و یا بشماریم یا تعداد اعضایش تا بی نهایت ادامه داشته باشد اما تا بی نهایتی که بتوان یک تناظر یک به یک بین آن مجموعه و اعداد طبیعی برقرار کرد. صفتهای پیوسته خیلی زیاد هستند مثل وزن، قد، درجه حرارت اینها با یک مقیاس نسبتی اندازه گیری اش میکنیم، متغیرهای پیوسته ای اند، مقدار پیوسته است. در واقع یک Continuous Attribut یک subset ای از اعداد حقیقی است.

* نوع های داده به چه شکلی هستند؟ میتوانند به شکل Record باشند یعنی ثبت شده، میتوانند بصورت گراف باشند و Network شبکه ای که در دو است، انواع و اقسام شبکه های اجتماعی را میتوانیم به صورت Data در نظر بگیریم همچنین داده ها میتوانند بصورت ordered باشند یعنی ترتیبی.

Sequential data ← مثل داده های DNA sequence که داده های ژنتیکی به صورت یک دنباله ای هستند.

* این خصوصیت هایی که در داده‌ها وجود دارد خیلی مهم است، یعنی داده‌ها را که ما جمع آوری کردیم به ۳ مسئله اشاره کنیم: یعنی بعد داده‌ها است، چون اگر بعد داده‌ها خیلی زیاد باشد با کامپیوتر هیچ کاری نمی‌توانیم بکنیم روی آن‌ها انجام دهیم مثلاً فرض کنید PCA می‌خواهیم انجام دهیم روی بردار داده‌هایی که ۱۰۰۰ ابعادی هستند، اصلاً با کامپیوتر نمی‌شه در عرض زمان یک ماتریس 1000 x 1000 را به صورت real Time حساب کرد عملاً مثلاً اینجا با مشکل مواجه می‌شویم. با بعد داده‌ها ما دو تا مشکل داریم، یکی اینکه از نظر ریاضی ما مشکلی نداریم ولی از نظر محاسباتی مشکل داریم مثل مثلاً PCA و دیگری اینکه ما تعداد پارامترهایمان از تعداد داده‌هایمان بیشتر است که بهش می‌گن curse of Dimensionality.

Important Characteristics of Structured Data

1 – **Dimensionality**
- **Curse of Dimensionality**

2 – **Sparsity**
- **Only presence counts**

ممکن است داده‌های ما داده‌های اسپارس باشند، داده‌های پراکنده باشند، اینقدر دور از هم باشند که ما نتوانیم در مورد آن‌ها هیچ تصمیمی بگیریم، یعنی ایده و هیچ ... برایشان ارزش دهیم. برای حل مشکل اسپارسیتی دوباره روش‌های زیادی است.

3 – **Resolution**
- **Patterns depend on the scale**

ممکن است داده‌های ما Resolution شان خوب نباشد، تغییر Resolution و مقیاس، یعنی بعضی مواقع کارهای ما را خراب می‌کنند، اطلاعات بدی به ما می‌دهند، داده‌ها را جمع آوری کردن با دقت خوبی جمع آوری شده باشند تا نتیجه غلط از آن‌ها نگیریم.

© Tan,Steinbach, Kumar   Introduction to Data Mining   4/18/2004   11

* دقتی که ما می‌گوییم منظورمان از Record of Data این است که داده‌های ما درون فایلی مثل Excel یا برای squel ضبط شده که معمولاً سطر اول نام متغیرهای ما است و بعد اطلاعاتی در مورد حالا مثلاً ...

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

© Tan,Steinbach, Kumar   Introduction to Data Mining   4/18/2004   12

* ممکن است داده‌هایمان به صورت ماتریسی باشد یعنی از یک بعد بیشتر باشد، ممکن است از دو بعد هم بیشتر باشد.

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

* این مثال همان Document است که برایمان گفته

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

* اسلاید ۱۵: داده های Transaction است که این داده ها را در کتاب از آن ها استفاده میکنیم.

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.

نگهدارنده رفتن خرید چیزی
خرید

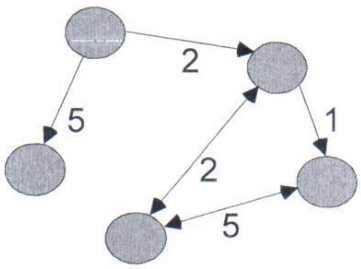
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Transaction

Graph Data

داده های گراف

- Examples: Generic graph and HTML Links

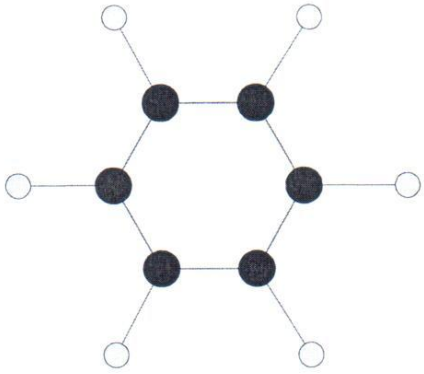



```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System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

Chemical Data

داده ها ممکن است chemical باشند (داده های شیمی)

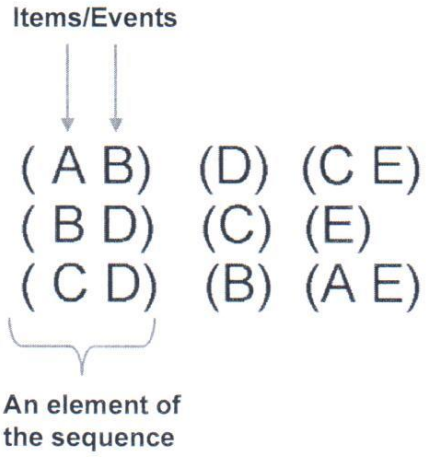
- Benzene Molecule: C_6H_6



Ordered Data

داده ها مرتب باشند

- Sequences of transactions



Ordered Data

DNA

- Genomic sequence data

این Sequence است.

این داده های ما است میخواهیم آینده را Predict کنیم، یعنی آیا این ژن ها با این توالی چه اتفاقی خواهد افتاد و چه خواهد داشت؟!

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```


* داده های مکانی =
* " زمانی = میخواهیم در زمان های مختلف گرم شدن کره زمین را بررسی کنیم.

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean

* کیفیت داده ها برای ما خیلی مهم است یعنی اگر ما پیشرفته ترین روش ها را داشته باشیم، پیشرفته ترین software و hardware داشته باشیم ولی اگر داده هایمان کیفیت لازم را نداشته باشند نمیتوانیم کار زیادی انجام دهیم. در جمع آوری داده ها معمولاً ما با چند نوع مشکل مواجه هستیم: خطاهایی که پیش می آید مثلاً؛ در مرکز آمار میگن آمارگیری زیر درختی یعنی آمارگیر رفته تا آمارگیری کند ولی بعد خسته شده زیر سایه درخت نشسته، برای خودش فرم ها را پر کرده، خیلی روشن و رسوا است.

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

* داده هایی که ما به دست می آوریم، بعداً باید کیفیتشان را چک کنیم ببینیم واقعاً این آمار آمار زیر درختی نبوده. ما انواع و اقسام روش ها داریم برای چک کردن کیفیت داده ها. (32)

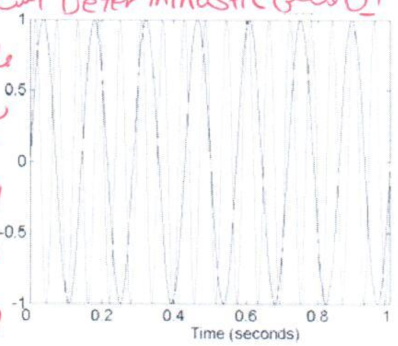
کتاب روش شناسی آمارگیری کتاب خوبی است برای ما دکتری بشیره دکتر صالحی ترجمه اش کرده، اطلاعات بسیار زیادی دارد.

Noise

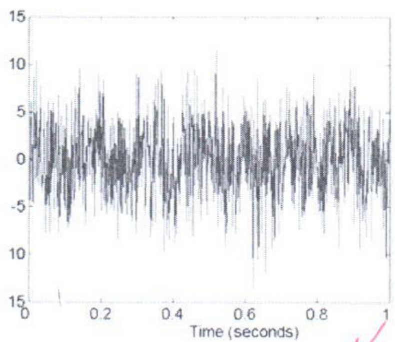
- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen

این منحنی Determinastic است، یعنی Deterministic یعنی تعیینی ما میتوانیم آینده را بصورت دقیق پیش بینی کنیم. الآن را با یک اختلاف فاز مثلاً یک متغیر تصادفی نرمال جمع کنیم یک همچنین شکل (شکل بعد) را داریم، صفرها اند.

این نویز است



Two Sine Waves



Two Sine Waves + Noise

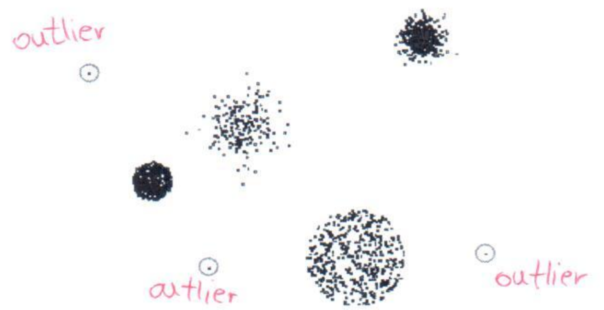
شکلی باشد که بتوانیم با استفاده از تکنیک های Noise Reduction به همچنین شکلی برسیم.

خطاها ترین دو تا است: خطاهای نمونه گیری و خطاهای غیرنمونه گیری. یعنی اینکه در خودش گنجانده شده، خطاهای نمونه گیری یعنی ما یک نمونه از جامعه می گیریم و یک استنتاج میکنیم، چون نمونه ما کل جامعه نیست، اگر یک نمونه دیگر بگیریم ممکن است نتیجه مان با قبلی کمی فرق کند. به این میگن خطای نمونه گیری، یعنی داده هایی که عوض شود نتیجه مان نیز عوض میشود. ولی یک خطاهایی هست خطاهای غیرنمونه گیری، مثلاً وقتی این فرم های سرشماری را پر میکردند، میدیدند که ماشین یا دستگاه هایی آنها را بخواند، اسم دستگاه

ICR است، قبلاً OCR بود (Intelligence Character Recognition) یعنی ما فارسی در این فرم ها می نویسیم، آن دستگاه دست خط ما را می خواند و دستگاه ها رقمی هستند چون در هر یک دقیقه می خواهد چند هزار برگ را به نحوی اطلاعاتش را به صورت دیجیتال دربیاورد. این دستگاه ICR بی خطا نیست، ممکن یک کلمه را بد نوشته باشیم و آن یک کلمه دیگر خوانده شود. پس هم در سرشماری عوض شود. به این می گوییم خطاهای غیرنفوذ گیری. این دیگر خطای دستی است.

Outliers

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



چطوری می شه این کارها را انجام داد؟ معمولاً خیلی هاش را نرم افزار به صورت اتوماتیک برای ما انجام می دهد یا گزارش می دهد مثل:
Anomaly Detection یعنی کشف داده های نابهنجار مثلاً در فرآیند بانکداری به وجود می آید.
حالا در مورد داده ها وقتی ما جمع آوری می کنیم همین خطای نفوذ گیری و خطای غیرنفوذ گیری باعث می شه که بعضی وقتها درون داده های ما نویز پیش آید. نویز یعنی اغتشاش یا outlier باشد یعنی داده ی پرت.

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)

clementin یک روش هایی برای این موضوع دارد، داده های گمشده را input می کند، بهترین راه این است که داده را حذف کنیم، standard این است که ستاره بگذاریم، در نرم افزارها می توانیم خودمان missing value را تعریف کنیم که نقطه باشد یا ستاره یا خط، به صورت default ستاره است.

ممکن است ما missing value داشته باشیم یعنی داده ی گمشده، یعنی ما روی ۱۰ تا کودک داشتیم بررسی انجام می دادیم وسط کار یکی اش می میرد. یک سری از مشخصاتش را اندازه گیری کردیم ولی وقتی که مرد دیگر نمی توانیم ادامه مشخصاتش را اندازه گیری کنیم، به خاطر همین ممکن است missing value. نحوه برخورد با outlier یک چیزی است، نحوه برخورد با نویز یک طوری است و نحوه برخورد با missing value یک طور دیگر است. با داده ی گمشده چه کار کنیم؟ جاهای خالی یعنی جای داده را باید یک چیزی بگذاریم یک عددی بگذاریم چه عددی بگذاریم؟ نرم افزار

Ridandanci یا Duplicate — اسلاید 25، ما وقتی داده‌ها را جمع آوری میکنیم، داده‌ی تکراری بوجود می‌آید، این است، اینها کیفیت داده‌های ما را پایین می‌آورد. مثلاً یک نفر عضوی می‌آید در یک جای چند بار رأیی می‌دهد در این سایت‌های اینترنتی. همین باعث میشه داده‌های ما را orib کند و ایراد دارد.

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

جلسه ۴ Data mining 92/07/8

اسلاید 26 = Data Preprocessing چند بخش دارد که در زیر آمده است.

Data Preprocessing

1- Aggregation
2- Sampling
3- Dimensionality Reduction
4- Feature subset selection
5- Feature creation
6- Discretization and Binarization
7- Attribute Transformation

Aggregation همان طور که از اسمش مشخص است، هدف از آن Data Reduction است یا تغییر اسکیل (Scale) است یا پایداری یا ثابت بودن در داده‌ها است. در واقع Aggregation کاری که انجام می‌دهد می‌آید دو یا چند تا از این Attributeها را با هم جمع می‌کند. مثال: ما می‌خواهیم میزان برق مصرفی را در نیروگاه‌ها پیش‌بینی کنیم، یک راه این است که اطلاعاتی که به ما می‌دهند چگونه است؟! مصرف برق خانگی، مصرف برق معابر، مصرف برق صنعتی، چند تا از این مصارف برق را به ما می‌دهند. اگر اینها را جدا جدا هم بگیریم ...