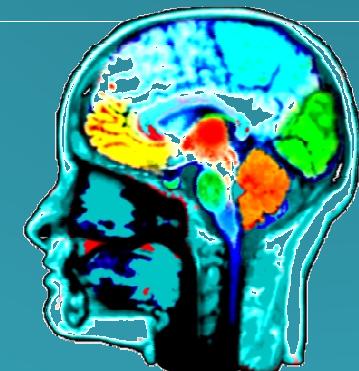




Introduction To Data Mining

Isfahan University of Technology (IUT)
Bahman 1401



Getting to Know Your Data

Dr. Hamidreza Hakim
hamid.hakim.u@gmail.com

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ

Content

Attributes and Objects

Types of Data

Basic Statistical Descriptions of Data

Data Visualization

Similarity and Dissimilarity Measures

ATTRIBUTES AND OBJECTS

What is Data?

- Collection of *data objects* and their *attributes*

The diagram shows a table representing a dataset. The columns are labeled *Name*, *Team*, *Number*, *Position*, and *Age*. The rows are indexed from 0 to 6. A label *Rows* points to the vertical axis of the table, and a label *Columns* points to the horizontal axis. A pink box labeled *Data* encloses the entire table area.

	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

دیتا چیست؟ یک مجموعه از ابجکت ها که هر کدام از این ابجکت ها ما میگیم یکسری مشخصه داره توی جدول:

هر کدام از این ردیف ها مرتبط با یک ابجکت است
ستون ها داره attributes رو مشخص میکنه
دیتا مقادیری است که تک تک این ها پیدا میکنه

What is Data?

- 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, dimension, or feature
- A collection of attributes describe an **object**
 - Object is also known as record, point, case, sample, entity, or instance

Attributes

The diagram illustrates a dataset consisting of 10 objects. Each object is represented by a row in a table with five columns: Tid, Refund, Marital Status, Taxable Income, and Cheat. The objects are grouped by a curly brace labeled "Objects". The "Attributes" are grouped by another curly brace above the table.

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

- attributes چیست؟

هر کدام از اون ابجکت های ما وقتی توی پدیده دارن رفتار می کنن ما یکسری ویژگی برآشون اندازه گیری میکنیم که این ویژگی ها تشکیل شده از همون مشخصه ها مثلًا این جدول مسئله اش این بوده که این فرد و امش رو پرداخت کرده یا نه --> مثل جدول : همچین اطلاعاتی رو از ادم های مختلف گرفتن --> حالا attributes چی میشه ؟ مشخصه هایی هر کدام از این ابجکت های ما --> الان ابجکت های ما اون ادم ها است و attributes ها وضعیت هایی که هر کدام از این کیس های ما دارند

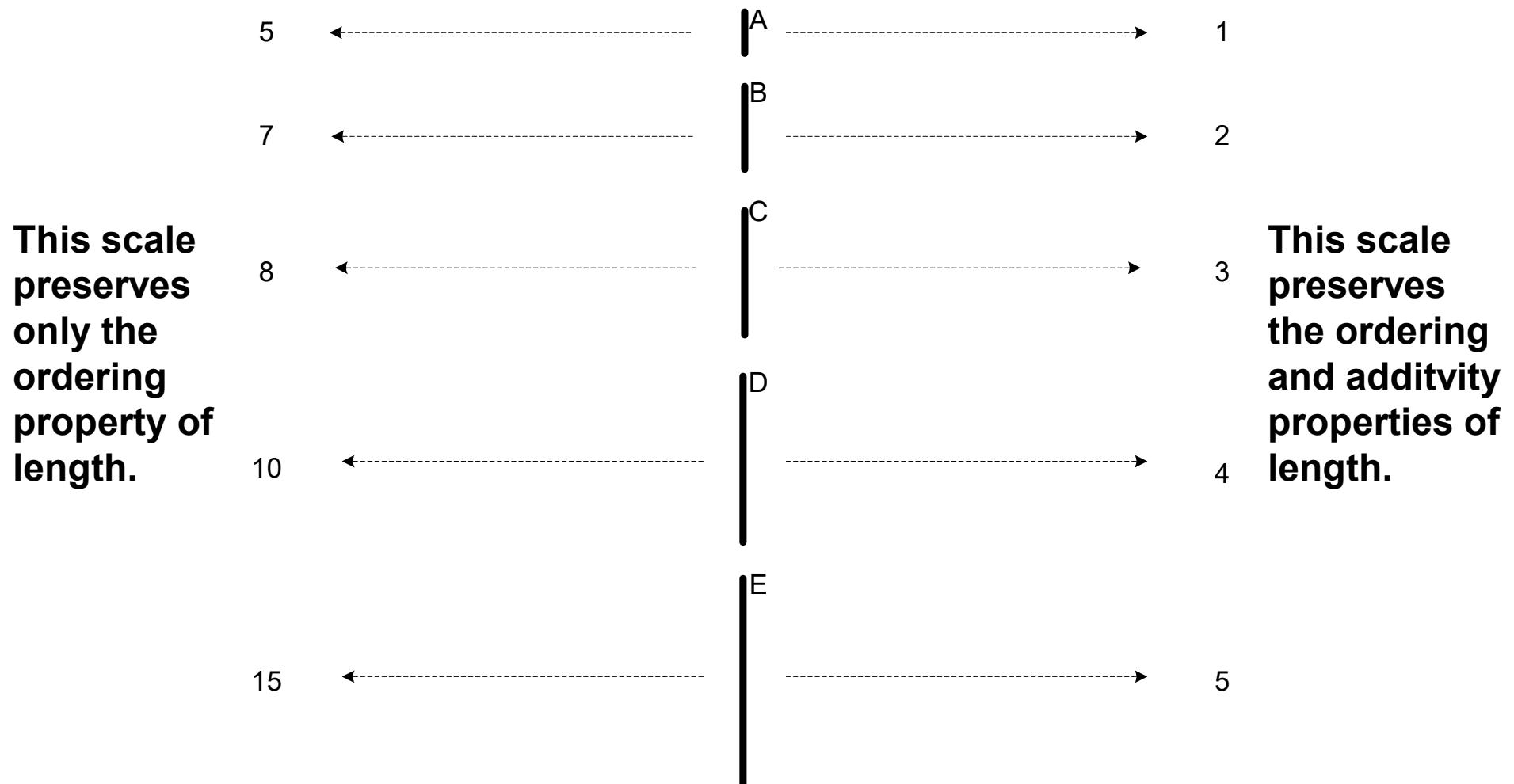
Attribute Values

- **Attribute values** are numbers or symbols assigned to an attribute for a particular object
- 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 can be different than the properties of the values used to represent the attribute

-
: مقداری که اون ابجکت ها توی اون Attribute پیدا کردن

Measurement of Length

- The way you measure an attribute may not match the attribute's properties.



-
مثال:

خیلی مهم است که ما یک مشخصه رو اندازه گیری میکنیم با چه مقدار هایی پر بکنیم و به مقدار هاش
چه سیمبل هایی رو نسبت بدیم
مثلما میخوایم قد دانشجویان رو به عنوان یک مشخصه ذخیره بکنیم --> قد را میتوانیم بگیم اینجا
کم زیاد یا ...

مثلما میخوایم طول رو اندازه گیری بکنیم و طول رو ذخیره بکنیم توی دیتابیس --> چند مدل
می تونیم ذخیره بکنیم مثلا می تونیم بهش نسبت بدیم ینی بگیم A, B, C ... یا بگیم 1 و 2 و ... یا
بگیم 5 و 7 و 8 و 10 و 15 --> هر کدام از این ها باعث میشه توی کارهای بعدی یک محدودیت
هایی داشته باشیم

پس توی مقدار دهی متغیرها باید یک سری چارچوب هایی رو رعایت بکنیم که 4 دسته مسئله داره
که صفحه های بعدی است ...

Types of Attributes

- There are different types of attributes
 - **Nominal**(اسمی):
 - ◆ categories, states, or “names of things”
 - ◆ Hair_color = {auburn, black, blond, brown, grey, red, white}
 - ◆ marital status, occupation, ID numbers, zip codes
 - ◆ Examples: ID numbers, eye color, zip codes
 - **Ordinal**(ترتیبی):
 - ◆ Values have a meaningful order (ranking) but magnitude between successive values is not known
 - ◆ Size = {small, medium, large}, grades, army rankings
 - ◆ Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}

انواع ویژگی ها:

: Nominal

متغیرهای اسمی است

وقت هایی که ما داریم یک صفت رو نگاه می کنیم که یکسری استیت از مقادیر رو داره مثل داریم یک ابجکت رو توصیف می کنیم و حالت هاش رو می خوایم بیان بکنیم مثل رنگ مو --> رنگ مو یک مجموعه مقادیری داره

: Ordinal

فرقشون با قبلی این است که ما مقادیری که به اون متغیر می دیم مفهوم ترتیب دارن تو شون و می تونیم بزرگتر و کوچکتر بر اشون در نظر بگیریم

Types of Attributes(Example)

- There are different types of attributes
 - Nominal
 - Ordinal
 - Interval(بازه ای)
 - ◆ Measured on a scale of equal-sized units q Values have order
 - ◆ Examples: calendar dates, temperatures in Celsius or Fahrenheit.
 - Ratio(نسبتی)
 - ◆ Inherent zero-point
 - ◆ Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Interval : بازه ای یا فاصله ای
مثالش تاریخ است

فرق بازه ای ها با ترتیبی ها چیه؟ ما داریم مقادیر تاریخ رو که ذخیره میکنیم بهش میگیم Interval چون اینجا نه تنها یک ترتیبی توی مقادیر وجود داره بلکه این ترتیب از لحاظ فاصله یک معنی داره مثلاً تاریخ 1 اسفند و بعد 29 بهمن یا ... می تونیم روی این ها یک فاصله هایی حساب بکنیم و بگیم این 2 روز بعد از اون است : Ratio

وقت هایی که ما با متغیر هایی سر و کله می زنیم که صفر ذاتی تو شون رخ میده نکته: عددی که اندازه گیری کردیم اگر بتوانیم عملیات تقسیم روشن انجام بدیم و بعد با عملیات تقسیم به یک رفتاری بررسیم که اون رفتاره یک رفتار معتبر باشه مثلاً دمای اتاق رو اندازه گیری کردیم شده 25 و بیرون رو اندازه گیری کردیم شده 12 ایا درسته که بگیم دمای اتاق 2 برابر دمای بیرون است؟ اگر کلوین باشه درسته و که اگر کلوین باشه دیگه این عدداً نیست پس ذاتاً غلط است این جمله

نکته: پس جایی که ذاتاً داخلش صفر داریم اون متغیر میشه Ratio مثلاً دما رو توی یک مسئله داده کاوی اندازه گیری کردیم الان مقادیری که برای دما ساختیم و عددهایی که به دست او مده کدوم یکی از اون 4 تای قبلی است؟ Interval میشه

Question

- Q1: Is student ID a nominal, ordinal, or interval-scaled data?
- Q2: What about eye color? Or color in the color spectrum of physics?

-
:Q1

شماره دانشجویی یک مقدار ordinal است بخارط سال ورود اولش که یک ترتیبی داره پس یه جاهایی که توی متغیرها باهاش کار میکنیم باید دقت کنیم انتظار چی هست مثلًا اگر با این شماره دانشجویی بخوایم یک کاری انجام بدیم و بگیم این دانشجو بعد از این دانشجو او مده یا قبلش توی دو رقم اولش سال ورود است پس ترتیب داره توش ولی وقتایی که قرار نیست همچین اتفاقاتی توش بیوقته ما دیگه ترتیب هم نداریم و میشه nominal

:Q2

رنگ چشم میشه nominal ولی اگر توی فیزیک باشه و بخوایم از این لحاظ نگاه کنیم میشه interval چون فاصله رو داره

Question

- Q1: Is student ID a nominal, ordinal, or interval-scaled data?

Nominal

- Q2: What about eye color? Or color in the color spectrum of physics? q

Eye color: Nominal (similar to hair color)

Color spectrum of physics: Interval (**RGB** space supports +/-)

سوال

Q1: آیا شناسه دانشجویی یک داده اسمی، ترتیبی یا فاصله ای است؟
اسمی

Q2: در مورد رنگ چشم چطور؟ یا رنگ در طیف رنگی فیزیک؟
رنگ چشم: اسمی (مشابه رنگ مو)
طیف رنگی فیزیک: فاصله (فضای RGB +/- را پشتیبانی می کند)

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
 - Distinctness: $= \neq$
 - Order: $< >$
 - Differences are meaningful $+ -$
 - Ratios are meaningful $* /$
- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & meaningful differences
- Ratio attribute: all 4 properties/operations

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10° is twice that of 5° on
 - the Celsius scale?
 - the Fahrenheit scale?
 - the Kelvin scale?

- Consider measuring the height above average
 - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?
 - Is this situation analogous to that of temperature?

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. ($=$, \neq)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
	Ordinal	Ordinal attribute values also order objects. ($<$, $>$)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

This categorization of attributes is due to S. S. Stevens

Attribute Type	Transformation	Comments
Categorical Qualitative	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}.
Numeric Quantitative	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.

This categorization of attributes is due to S. S. Stevens

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.

-
دسته بندی بعدی:
گسته بودن و پیوسته بودن است

Asymmetric Attributes

- Only presence (a non-zero attribute value) is regarded as important
 - ◆ Words present in documents
 - ◆ Items present in customer transactions
- If we met a friend in the grocery store would we ever say the following?

“I see our purchases are very similar since we didn’t buy most of the same things.”

Critiques of the attribute categorization

- Incomplete
 - Asymmetric binary
 - Cyclical
 - Multivariate
 - Partially ordered
 - Partial membership
 - Relationships between the data
- Real data is approximate and noisy
 - This can complicate recognition of the proper attribute type
 - Treating one attribute type as another may be approximately correct

TYPES OF DATA

Types of data sets

- Record(Tabular)
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

-
معروف ترین نوع پایگاه داده ها، پایگاه داده های Record محور است

Record Data

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

<i>Tid</i>	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

هر کدوم از المان هاش یک ابجکت رو داره بیان میکنه و مشخصه های اون ابجکت توی ستون هاش داره نشون داده میشه

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 a 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

- فرق این دیتا با قبلیه چی هست؟

اینجا موقعیت خود سطرها هم برآمون موضوعیت داره

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

دیتای متدال دیگه ای که داریم Document Data است

برای هر داکیومنت یکسری دیکشنری برآش ایجاد میکنیم و تعداد رخداد کلمات که توی هر داکیومنت هست رو مشخص میکنیم و بعد دیگه داکیومنت ما میشه این رشته یعنی این رشته ای از اینکه چه کلماتی توش رخ داده

Transaction Data

- A special type of data, where
 - Each 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.
 - Can represent transaction data as record data

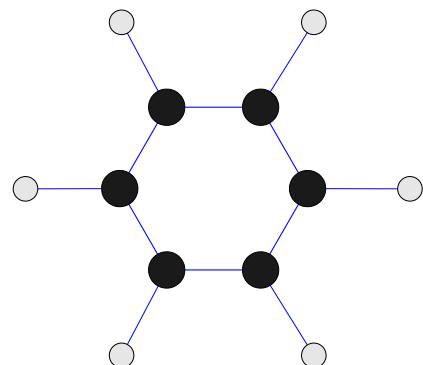
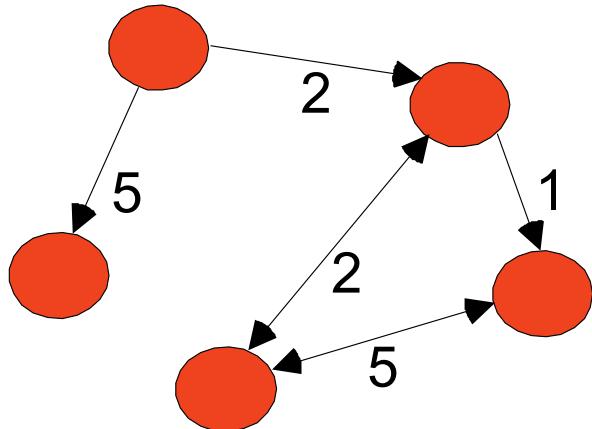
<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

: Transaction Data

یک اتفاقی افتاده مثلا یک سبد خریدی داریم و یک مشتری او مده یک تعداد کالایی خریده و ما نتیجه اون هارو داریم ینی ایتم هایی که مشتری خریده رو داریم

Graph Data

- Examples: Generic graph, a molecule, and webpages



Benzene Molecule: C₆H₆

Useful Links:

- [Bibliography](#)
- Other Useful Web sites
 - [ACM SIGKDD](#)
 - [KDnuggets](#)
 - [The Data Mine](#)

Knowledge Discovery and Data Mining Bibliography
(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 21, no. 1, March 1998.

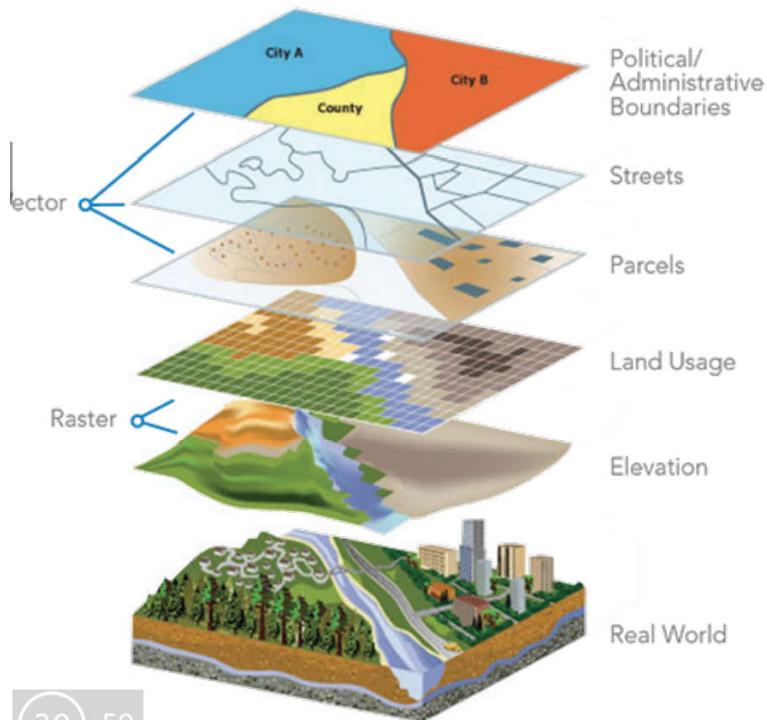
Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for Knowledge Discovery in Databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

-
گراف تشکیل شده از یکسری نود و یال
که یال ها ارتباطات رو بیان میکنه

Spatial/Image Data

● Spatio-Temporal Data

□ Maps



□ Images



- دیتای مکانی:

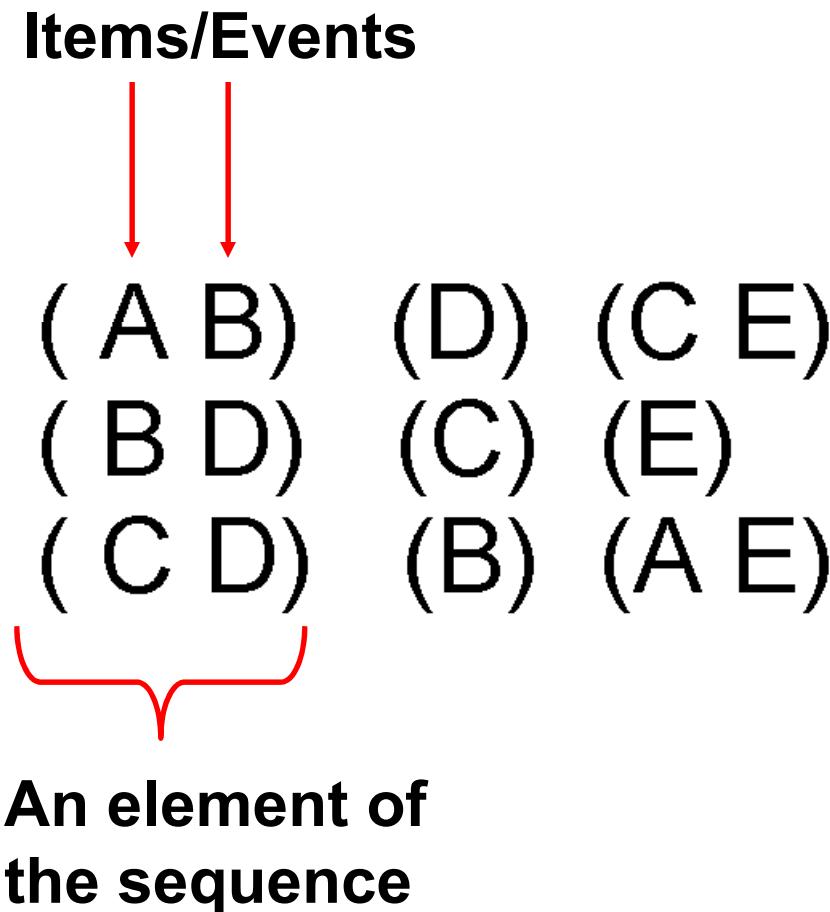
تصویر از جمله دیتای مکانی است چون هر نقطه از تصویر که چه مقداری داره و همسایه هاش چطوری هستن حاوی اطلاعات است

وقتایی که می خوایم یک ابجکت رو شناسایی کنیم رفتار پیکسل های مجاور مهم میشه مثلا اگر ببینیم این پیکسل ها زرد ه و مجاورشون ابی شده می تونیم بگیم یک ابجکت جدیدی شروع شده پس اطلاعات توی مکان ذخیره شده

مثلا توی تصاویر ماهواره ای این مسئله می تونه خیلی پیچیده تر بشه مثلا چندتا لایه داریم مثلا لایه تصویر برداشته از خود زمین --> مثلا شدت رنگش یا وضعیت منابع طبیعیش یا ...

Ordered Data

- Sequences of transactions



داده های ترتیبی:
مثالش صفحه بعدی ...

Ordered Data

- Genomic sequence data

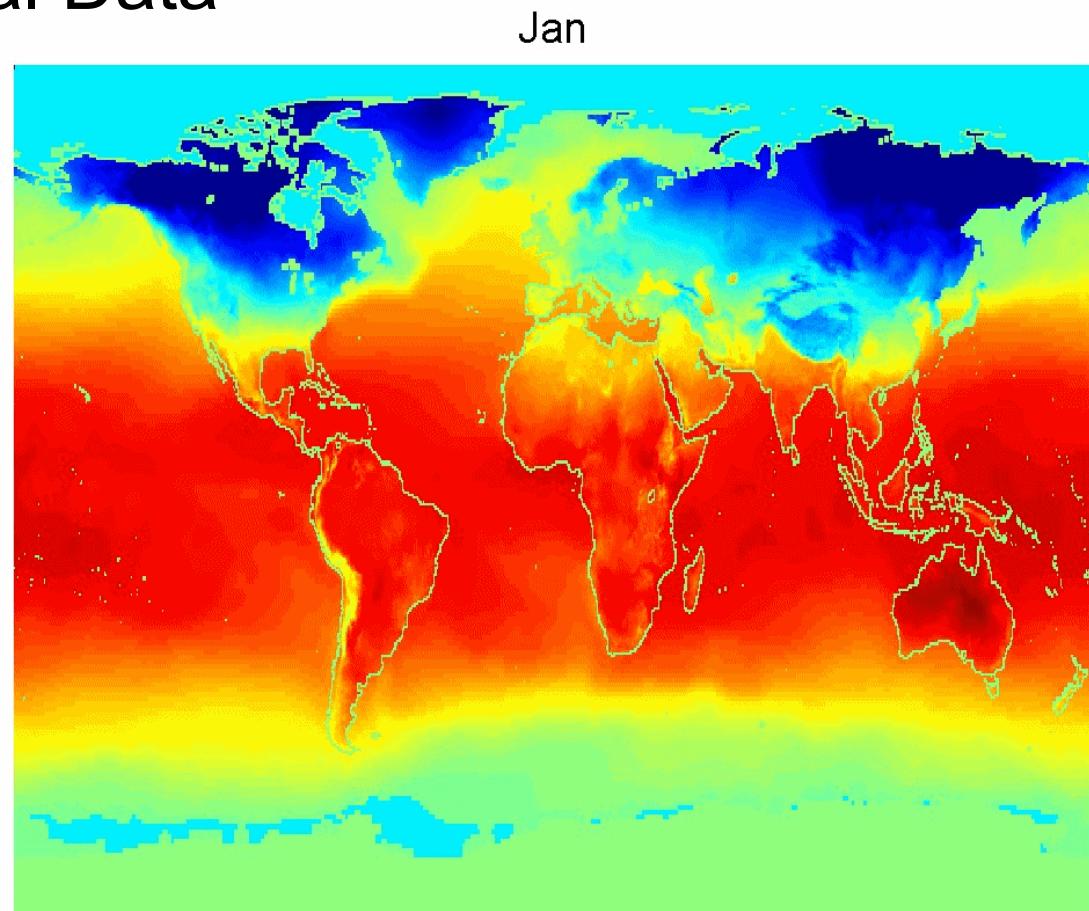
GGTTCCGCCTTCAGCCCCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCAGGGGCCGCCCAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

مثلاً دنباله ژنوم یک فرد --> اطلاعات توی ترتیب این هاست ینی عقب و جلو شدن این هاست که یک فرد رو با یک فرد دیگه متمایز میکنه و مسئله ای که داریم اینه که روی این ترتیب ها باید این اطلاعات استخراج بشه

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



بعضی جاها اطلاعات جدا از اینکه در مکان است اطلاعات در زمان هم رخ می ده مثلًا وضعیت آب و هوا هم یک بخشی از اطلاعات در مکان ذخیره شده و هم یک بخشی از اطلاعات در زمان

مثلًا وضعیت قیمت دلار در کشور اینجا اطلاعات در زمان ذخیره شده

BASIC STATISTICAL DESCRIPTIONS OF DATA

توصیفات آماری پایه داده ها

Basic Statistical Descriptions of Data

- Motivation
 - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
 - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
 - Data dispersion: analyzed with multiple granularities of precision
 - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
 - Folding measures into numerical dimensions
 - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

Note: n is sample size and N is population size.

- Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- Trimmed mean: chopping extreme values?(2%)

-
معیارهای تجمع دیتا:

میانگین معمولی

میانگین وزن دار:

در کجاها استفاده میشه --> مثلا ما یه جاهایی یک دیتایی رو جمع اوری کردیم و راجع به اینکه این دیتا درست است یا نه شک داریم ینی مطمئن نیستیم که این دیتا درسته و در کل می دونیم این دیتا جمع اوری شده ولی به اون کسی که این دیتا رو جمع اوری کرده شک داریم یا مثلا نگرانیم که نکنه توی خود اون دیتا نویز وجود داشته باشه و می خوایم اطلاعاتش رو دور نیندازیم پس میایم سراغ میانگین های وزن دار --> و وزن کمتری به اون داده ها می دیم

حذف داده هایی که مقادیرشون خیلی پرت است چه زمانی توی میانگین گیری سراغ این می ریم؟ خیلی وقتا یکسری دیتاهای هستن که مقادیرشون با بقیه خیلی متفاوت است که اینا توی میانگین گیری کار رو خراب می کن --> بهتره این مقادیر رو حذف بکنیم و بعد بریم سراغ میانگین گیری

Measuring the Central Tendency

- Median:

- Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for grouped data):

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum freq)_l}{freq_{median}} \right) \text{width}$$

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

- میانه:

همون مقداری که وسط قرار داره

مثلا می خوایم توی این جدول بیایم میانه رو پیدا بکنیم --> توی این رویکرد میانه رو تقریب میزنه
اول میاد موقعیت میانه رو مشخص میکنه میگه موقعی که داریم میانه این 6 تا بازه رو مشخص می
کنیم میانه میشه دقیقاً اونجایی که وسط دیتا است پس میاد همه فرکانس هارو جمع می زنه و به همون
تعداد می دونه که رکورد داریم بعد وسط اون مقدار رو حساب میکنه و میگه این رکوردی که الان
داریم توی کدوم بازه سن قرار داره و الان اینجا میشه توی بازه 21 تا 50 حالا عدش رو دقیق چی
بگیم؟ میاد مقدار این عدد رو تقریب می زنه که فرمولش رو پایین نوشته

L1 میشه عرض این بازه که الان میشه 29

فرکانس میانه اینجا میشه 1500 تا

Measuring the Central Tendency

- Mode

- Value that occurs **most** frequently in the data
- Unimodal, bimodal, trimodal
- Estimate Mode
 - ◆ Empirical formula(approximate):

این مساوی نیست بلکه تقریباً مساوی است

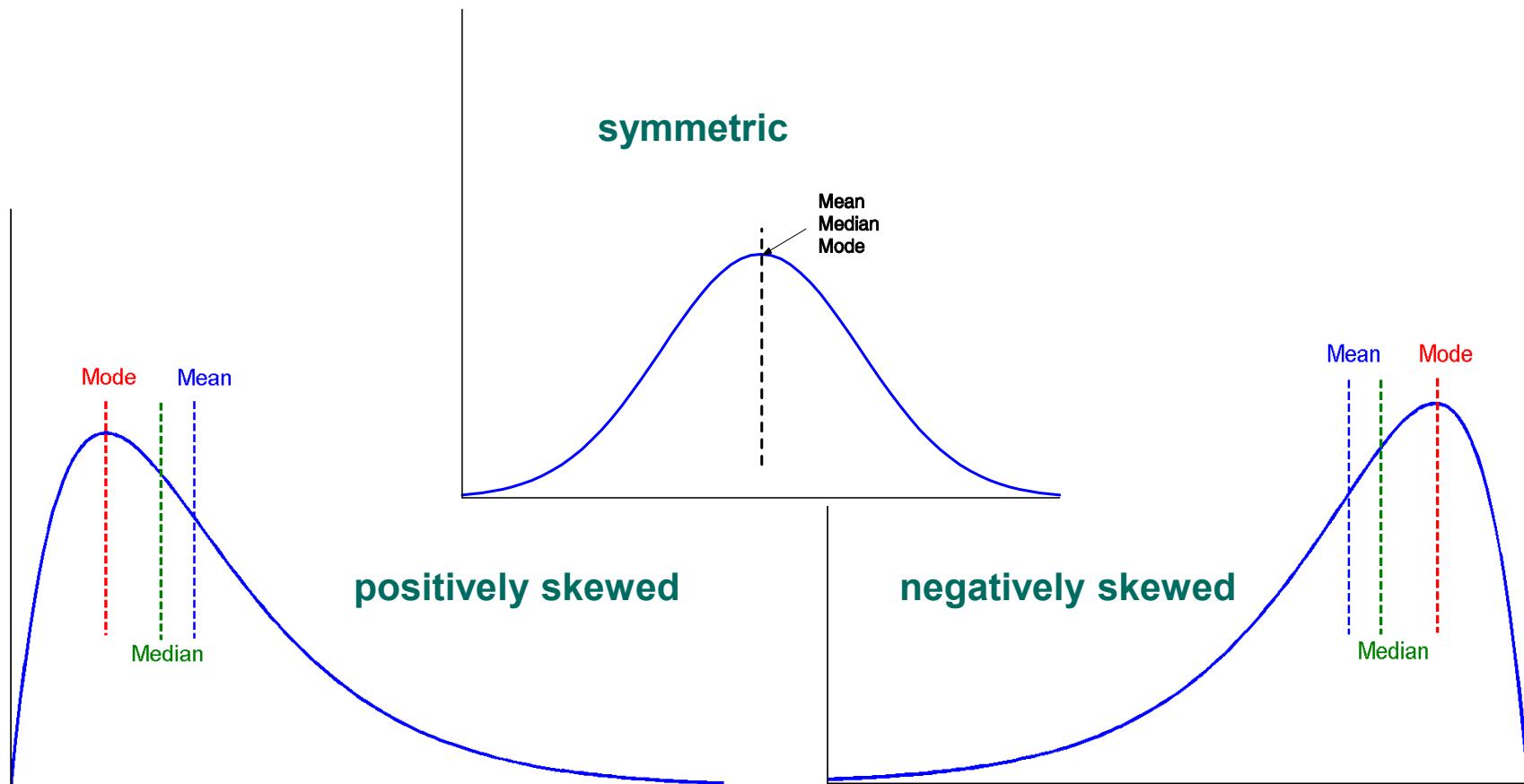
$$mean - mode = 3 \times (mean - median)$$

$$Mode = 3median - 2mean$$

-
مد: ینی پرتکرارترین رکورد یا دیتایی که داریم

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



وقتی فراوانی دیتاها رو رسم میکنیم --> نمودارها رو ببین:
positively skewed : ینی سمت منفیش خم شده
negatively skewed : ینی سمت مثبتش خم شده

چندتا ویژگی داره که قبل از اینکه رسم بکنیم این نمودارها رو می تونیم با کمک میانگین گیری و میانه و مد مشخص بکنیم --> به موقعیت این سه تا ویژگی نگاه می کنیم که نسبت بهم کجا قرار گرفتن وقتی که توزیع نمودار متقارن است ینی هر سه تا از این مشخصه دارن یک مقدار رو گزارش می کنن
مثال:

کی positively skewed داریم --> سن افراد ادمای زنده جامعه: هرچی سن بالاتر میره احتمال اینکه اون فرد زنده باشه خیلی کم است

کی negatively skewed داریم --> ??

Measuring the Dispersion of Data

- Variance and standard deviation (*sample*: s , *population*: σ)

- **Variance**: (algebraic, scalable computation)

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

- **Standard deviation s (or σ)** is the square root of variance s^2 (or σ^2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 \right]$$

-
معیارهای پراکندگی دیتا:

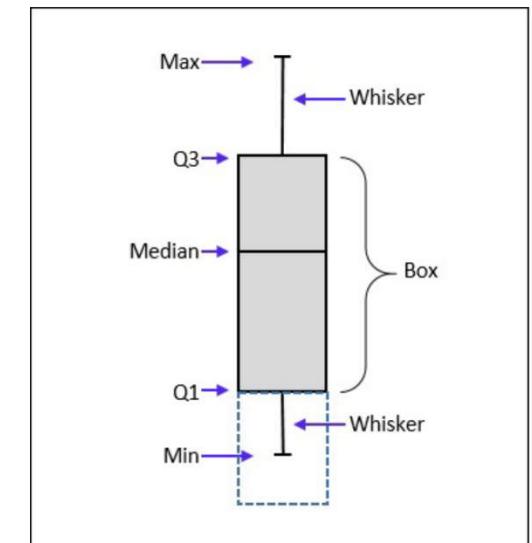
واریانس: فاصله از میانگین رو اندازه گیری میکنه --> توی داده کاوی سراغ این روش نمی‌ریم
چرا سراغ این نمیریم؟ چون توی داده کاوی ما میانگین رو معمولاً نداریم

بیشتر از Standard deviation استفاده میکنیم توی DM

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
 - Quartiles:** Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
 - Inter-quartile range:** $IQR = Q_3 - Q_1$
 - Five number summary:** min, Q_1 , median, Q_3 , max
 - Boxplot:** ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
 - Whiskers: two lines outside the box extended to Minimum and Maximum
 - Outlier:** usually, a value higher/lower than $1.5 \times IQR$

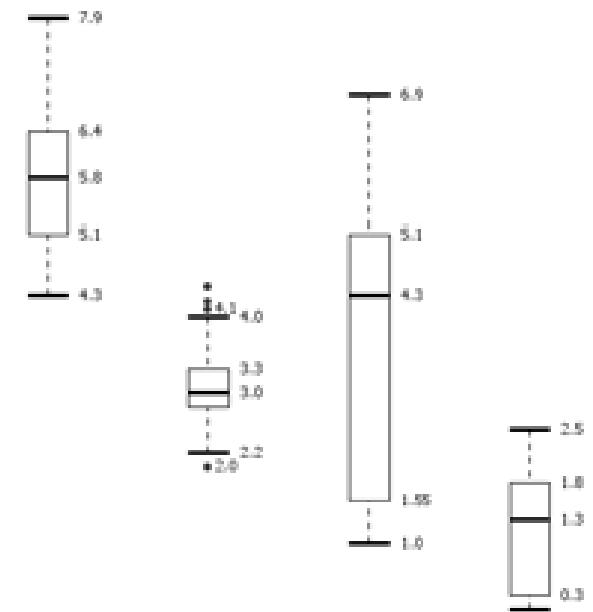
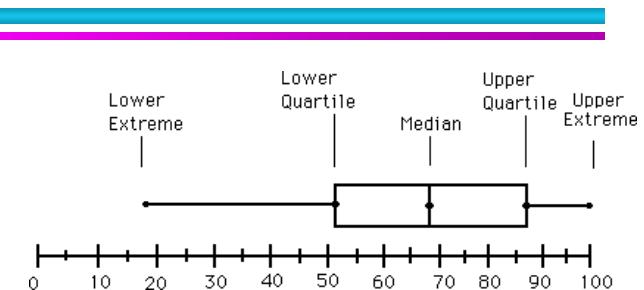
این شک رو بپرس بعضاً! الان IQR میشه کل جبه بیرون اون میشی نمونه پرت؟ و چرا ماکس و مین داریم از کجا اونا اومده؟



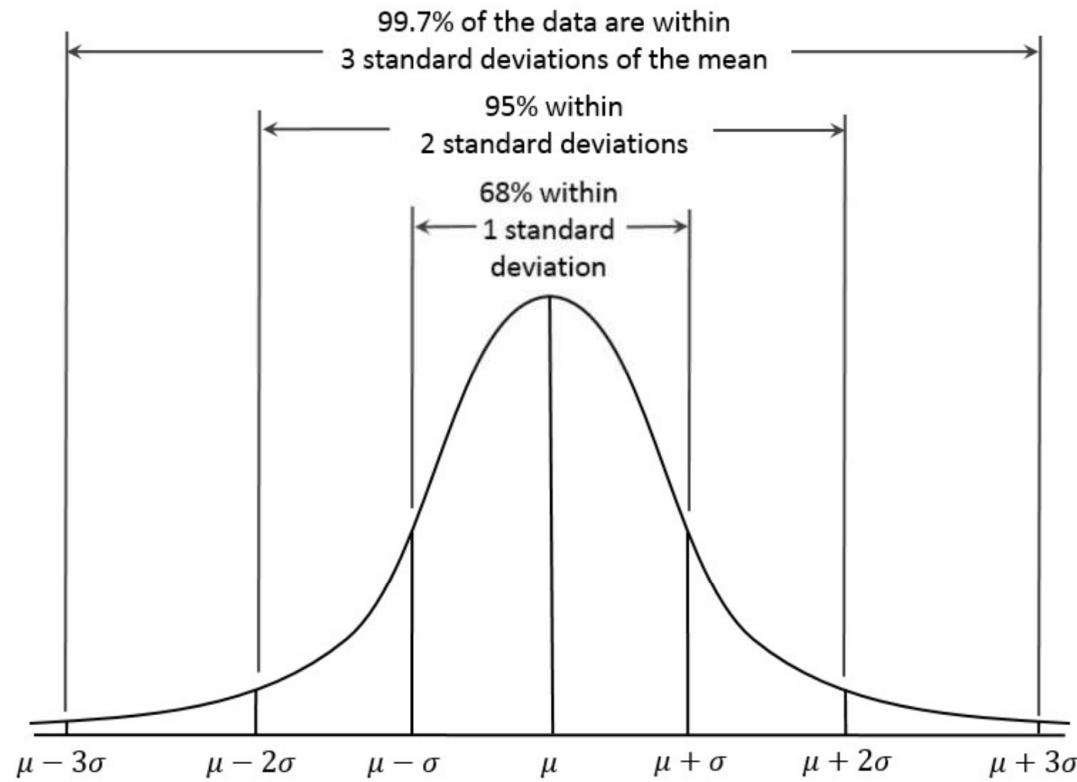
یک نمودار دیگه ای که داریم Boxplot است که میاد از چارک ها کمک می گیره --> اینم یک اطلاعاتی راجع به پراکندگی دیتا بهمون میده که اینجا 4 تا چارک داریم چارک دوم میشه میانه به فاصله بین چارک اول و سوم می گن IQR طبق این تعریف میشه نمونه پرت --> نمونه پرت، نمونه ای میشه که از 1.5 برابر IQR کمتر است یا بیشتر

Boxplot Analysis

- **Five-number summary** of a distribution
 - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
 - Data is represented with a box
 - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
 - The median is marked by a line within the box
 - Whiskers: two lines outside the box extended to Minimum and Maximum
 - Outliers: points beyond a specified outlier threshold, plotted individually



Properties of Normal Distribution Curve



□ Z-score: $z = \frac{x - \mu}{\sigma}$

توزیع نمودار Z یا :z-score

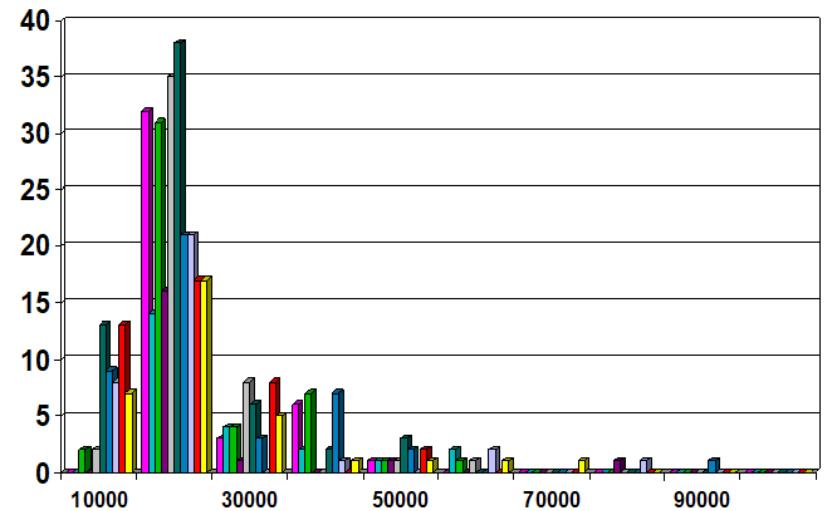
کاری که اینا می کنن اینه که میان اون داده های ما رو می برن توی یک فضای توزیع نرمال استانداردش می کنن و بعد میگن اگر این توزیع داده ها نرمال باشه این نمونه ای که ما داریم راجع بهش حرف می زنیم چقدر از میانگین فاصله داره و اینو با یک سری عدد گزارش میکنن که این عدد رو همه می فهمن

اگر رکوردی دیدیم که مقدارش از میو - 3 سیگما کمتر بود اون دیتا واقعاً دیتای پرتی است

میگه Z های خیلی بزرگ و کوچک چقدر از میانگین فاصله دارن پس با یک عدد می تونیم هم فاصله از میانگین رو بسنجیم و هم یه جوابی انحراف معیار هم در نظر بگیریم

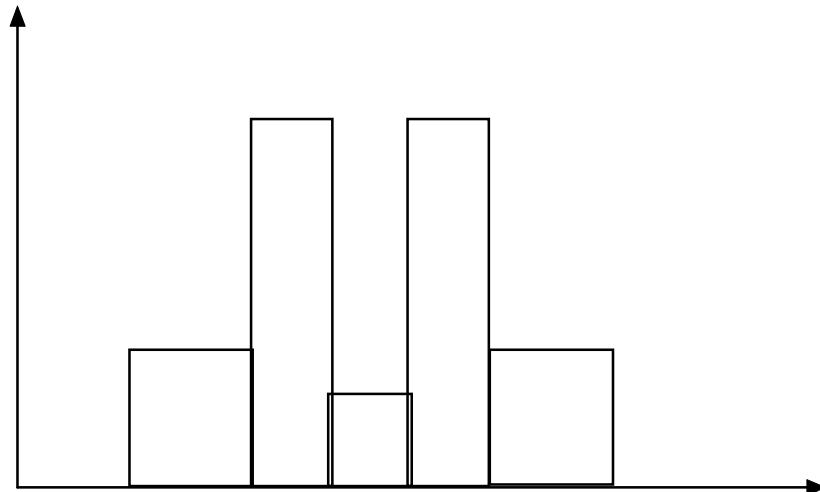
Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

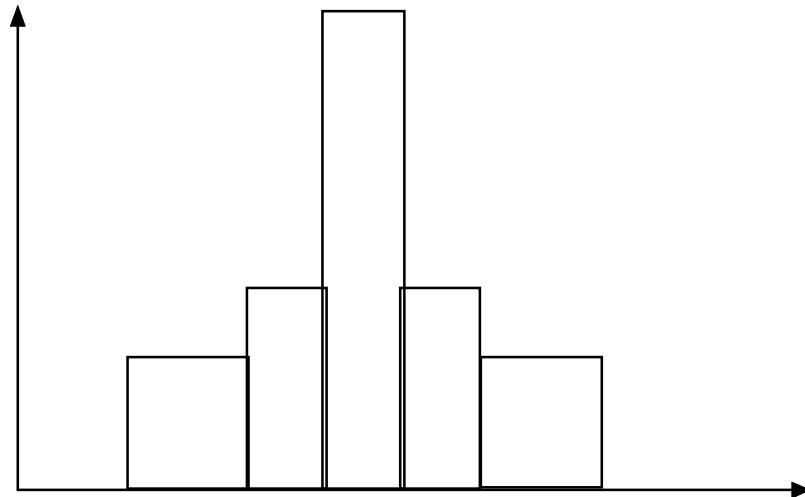


هیستوگرام انالیز: این بهمون فراوانی مقادیر رو میگه
اینجا میان اون پدیده ای که پیوسته هست رو بازه بندی می کن و بعد میان فراوانیش رو می شمرن

Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
 - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions



چرا می ریم سراغ هیستوگرام و هیستوگرام اطلاعات بیشتری نسبت به Boxplots داره؟ بعضی وقتا پیش میاد که Boxplots دو تا نمودار با هم برابره ولی هیستوگرام هاشون با هم متفاوته و این ینی این که هیستوگرام خیلی اطلاعات بیشتری داره بهمون منتقل می کنه --> اطلاعات بیشتر در طول تحلیل های دیگه بهمون کمک میکنه

این دو تا نمودار رو به رو از لحاظ Boxplots شبیه هم هستند ولی هیستوگرام هاشون متفاوته

هیستوگرام خیلی کمک میکنه که توزیع داده ها رو مشخص بکنیم

Graphic Displays of Basic Statistical Descriptions

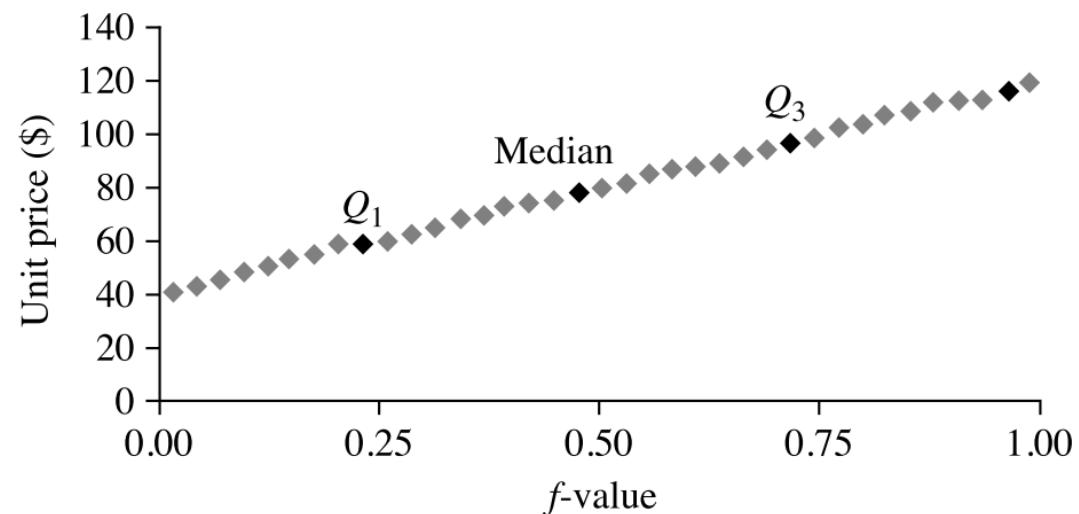
- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis repres. frequencies
- **Quantile plot:** each value x_i is paired with f_i , indicating that approximately $100 f_i \%$ of data are $\leq x_i$
- **Quantile-quantile (q-q) plot:** graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- **Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane

Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **quantile** information
 - For a data x_i data sorted in increasing order, f_i indicates that approximately $100 f\%$ of the data are below or equal to the value x_i

Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price (\$)	Count of items sold
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350



: Quantile Plot

نمایشی که داره به این صورت است که محور X اش داره درصد داده ها رو می گه ینی چند درصد داده ها و محور y داره اون مقداری که ما برآش این نمودار رو رسم کردیم رو می گه مثلا قیمت یک کالا رو توی فروشگاه های مختلف یک شهر اندازه گیری کردیم و کف قیمت کالا 40 بود و ماکزیمم 120 تا بوده

برای تولید این نمودار باید قیمت ها رو سورت بکنیم از کمترین قیمت به بیشترین قیمت هدف این نمودار چیه؟ می خواد بگه تا چند درصد داده ها قیمت روی چنده مثلا 25 درصد داده ها قیمتیشون از 60 کمتر است

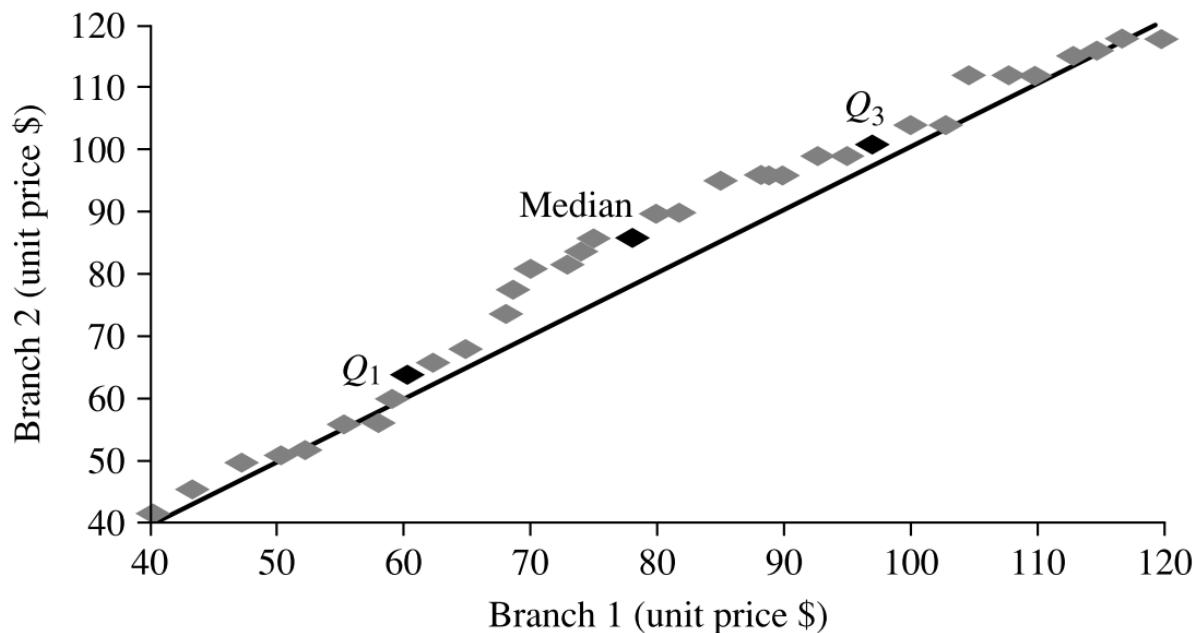
این همون Outlier است ولی به صورت پیوسته --> فقط اینجا با چارک اول و دوم و سوم و چهارم دیگه کار نداریم و یه جور ای اینجا داره پشت سر هم میشه

مثلا اینجا که 25 درصد است میشه موقعیت چارک اول

وقتی این نمودار رو داشته باشیم می تونیم بفهمیم چقدر منظم قیمت ها توزیع شده

Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

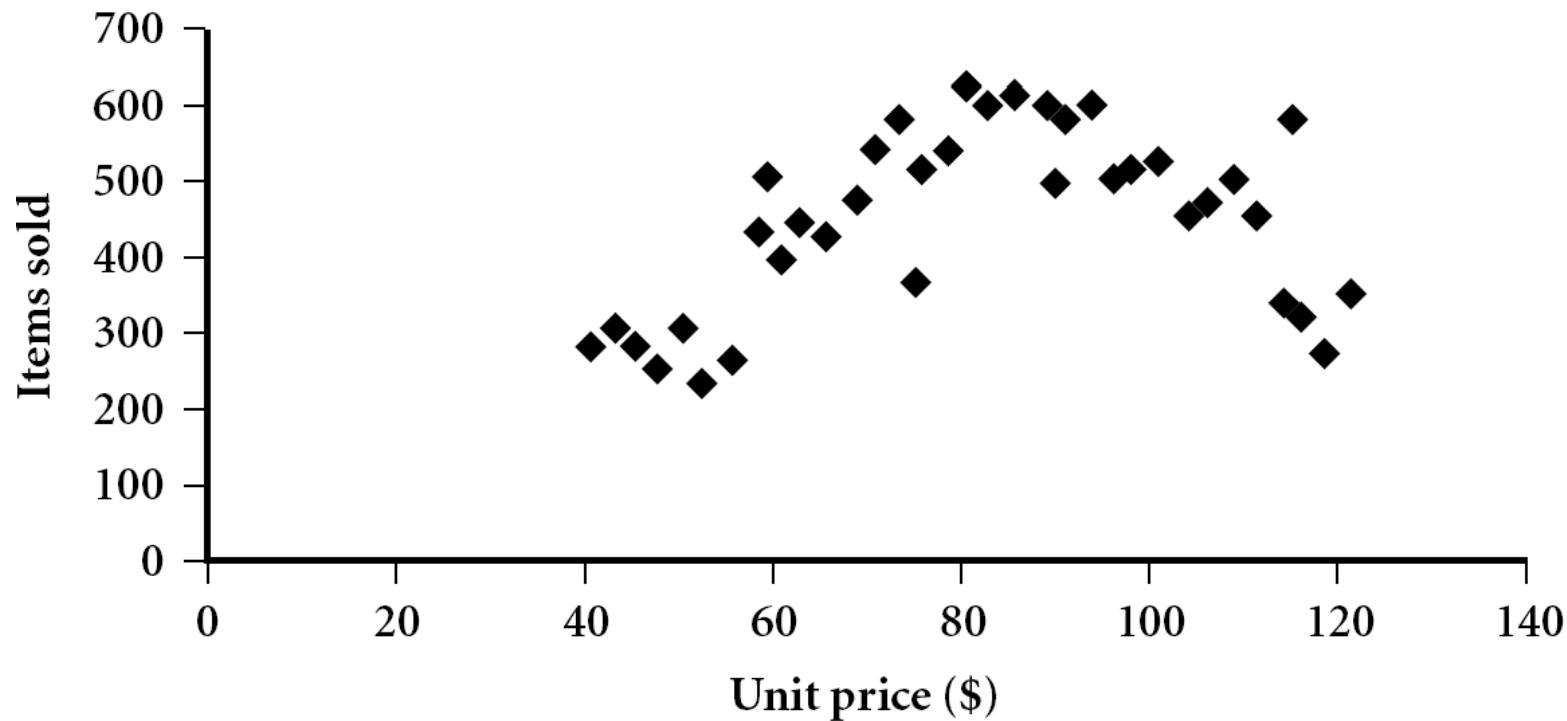


- نمودار صفحه قبل مقدمه‌ی استفاده است توی تحلیل‌های دیگه که بهش می‌گن Q-Q Plot این بیشتر توی انالیز دوتا پدیده با هم است مثلاً نمودار قیمت گوشت توی یک شهر رو گرفتیم و توی یک شهر دیگه هم گرفتیم --> یک نمودار یک شهر رو بهمن میده و یک نمودار دیگه هم یک شهر دیگه رو حالا می‌خوایم این دوتا شهر رو با هم مقایسه بکنیم --> این دوتا نمودار محور X مشترکی دارند و می‌تونیم با کمک محور X این دوتا نمودار رو کنار هم بذاریم و یک نمودار جدیدی ایجاد بکنیم که بهش می‌گن Q-Q Plot --> این نمودار یک گزارشی میده که قیمت‌ها توی چه شهری تراکم بیشتری دارند

نکته: اگر داده‌ها واقعاً توزیعشون نرمال باشه می‌افته روی یک خط ولی اگر نرمال نباشه داده‌ها از اون خط فاصله می‌گیره --> هر چقدر توزیع این دوتا داده اگر مثل هم باشند اینا روی یک خط رفتار می‌کنند --> اونجاهايی که از خط فاصله می‌گيرند يني توزيع يكى نسبت به يكى دیگه متفاوت است

Scatter plot

- Provides a first look at **bivariate data** to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



:Scatter plot

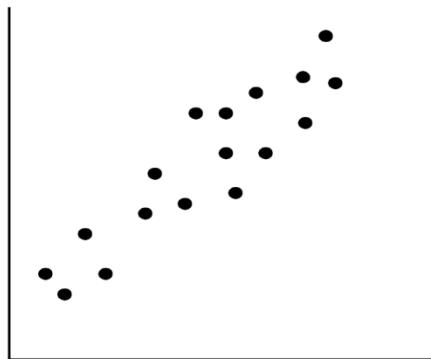
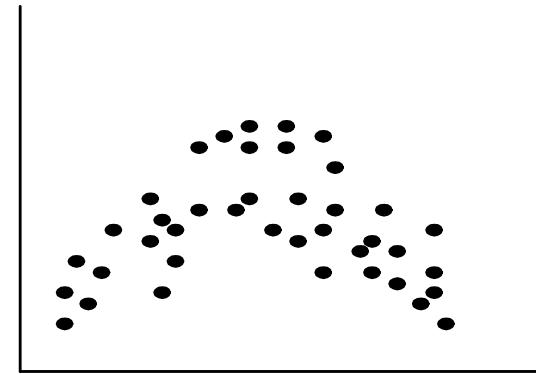
یه جاهایی می خوایم دوتا متغیر رو با هم مقایسه بکنیم مثلا یک ویژگی قیمت یک کالا است و یک ویژگی دیگه تعداد فروش اون کالا است --> مثلا ما او مدیم میزان فروش کالا رو توی مغازه های مختلف و قیمت متوسطی که داشتند مقایسه کردیم --> مثلا متوسط قیمتی که داشتیم روی کالاها 40 دلار است و 300 تا هم فروخته به طور متوسط

محور X اش یک ویژگی است و محور Y اش هم یک ویژگی دیگه است --> و این یک گزارشی به ما میده مثلا الان روی این نمودار اطلاعاتی که کسب می کنیم اینه که قیمت های این وسط فروش بیشتری دارند و قیمت هرچی کم بشه یا زیاد بشه تقاضا برآش کمتر است --> روی نمودار رو ببین

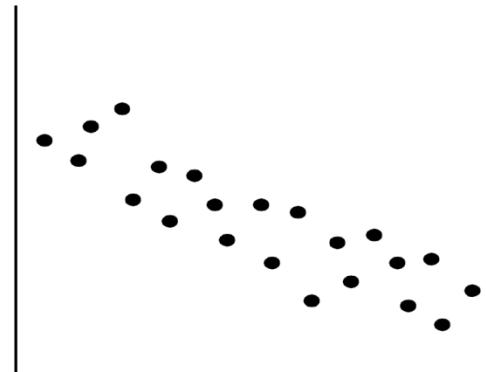
نکته: این نمودار برای زمانی که می خوایم دوتا ویژگی رو با هم دیگه مقایسه بکنیم خیلی بهمن کمک میکنه و اگر کسی داده اش پرست باشه با رسم این نمودار دید پیدا میکنه

Positively and Negatively Correlated Data

- The left half fragment is positively correlated
- The right half is negative correlated



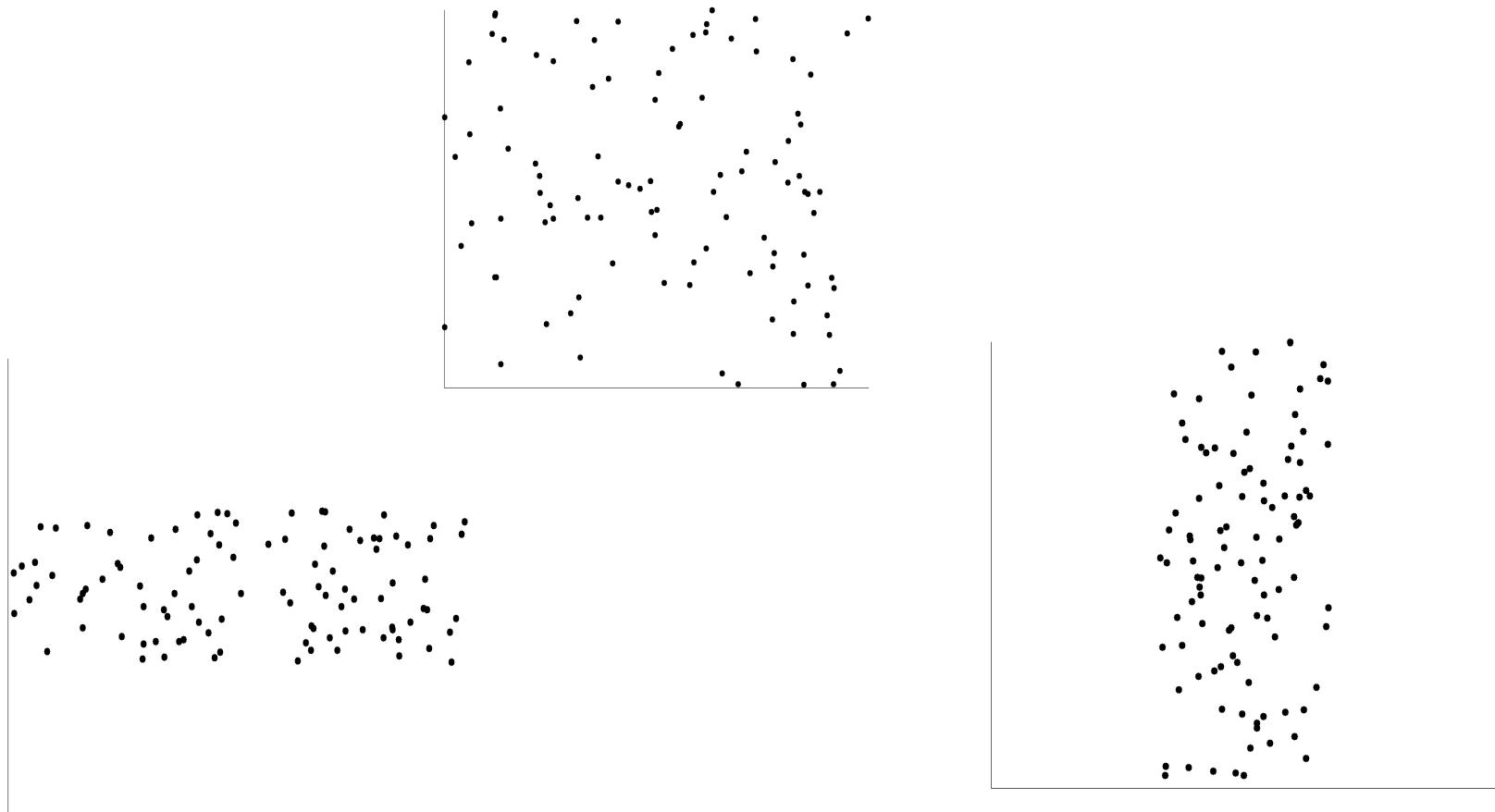
Positively correlated



Negatively correlated

این Scatter plot دو حالت مشخص داره که معروفه به شbahت مثبت و شbahت منفی:
مثلا دوتا مشخصه اینجا داریم --> قد و وزن و این ها رو رسم کردیم --> هر جایی که قد داره افزایش پیدا میکنه وزن هم داره افزایش پیدا میکنه پس میگیم قد و وزن شbahت مثبت دارند چون با افزایش قد، وزن هم افزایش پیدا کرد
ولی یکسری پدیده ها بر عکس این است --> مثلا چگالی و حجم --> هر چه حجم بیشتر باشه چگالی کمتر است --> شbahت منفی میشه

Uncorrelated Data



یک حالتی هم هست که هیچ الگویی ندارند --> توی یک رنجی شباہت مثبت دارند و توی یک رنج دیگه ای شباہت منفی دارند --> این ها رو معمولاً میان دو قسمت می کنن مثلاً میگن از این مقدار کمتر یک رفتار مثبتی داریم و از این مقدار بیشتر رفتار منفی داریم

یه جاهایی هم هیچ ارتباطی با هم ندارند



DATA VISUALIZATION

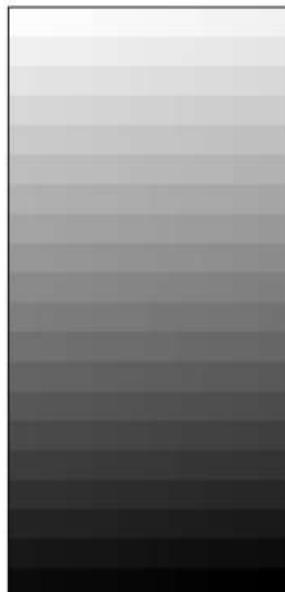
Data Visualization

- Why data visualization?
 - Gain insight into an information space by mapping data onto graphical primitives
 - Provide qualitative overview of large data sets
 - Search for patterns, trends, structure, irregularities, relationships among data
 - Help find interesting regions and suitable parameters for further quantitative analysis
 - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
 - Pixel-oriented visualization techniques
 - Geometric projection visualization techniques
 - Icon-based visualization techniques
 - Hierarchical visualization techniques
 - Visualizing complex data and relations

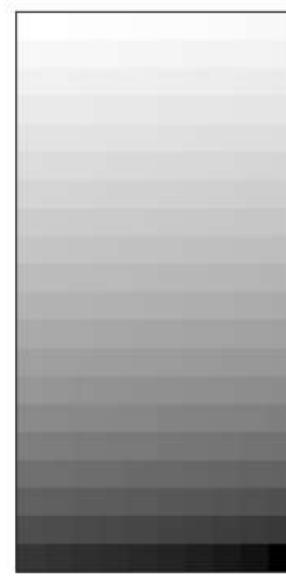
- هدفمون از بصری سازی چیه؟ اینکه بفهمیم چه اطلاعاتی داخلش هست و از نزدیک داده ها رو با هم دیگه ببینیم - کیفی بررسی بکنیم داده ها رو ینی کیفیت داده ها به چه صورت است نکنه داده هامون داده های پر تی باشه یا .. - یه جاهایی دنبال الگو هستیم مثلا می خوایم یک پترنی رو استخراج بکنیم و براساس پترنی که در اوردیم بیایم یه کاری رو راه بندازیم - بعضی جاها برای پیدا کردن پارامترها سراغ Visualization می ریم و می خوایم بعضی وقتا هم تایید بکنیم

Pixel-Oriented Visualization Techniques

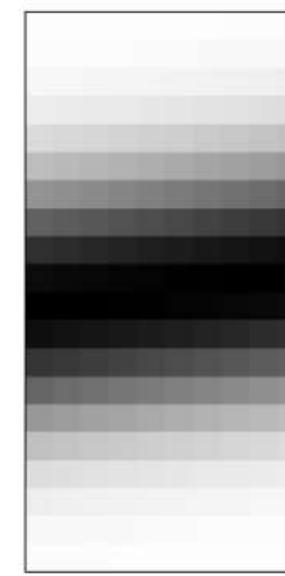
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



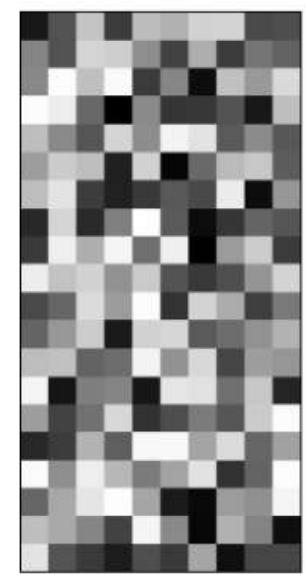
(a) Income



(b) Credit
Limit



(c) transaction
volume



(d) age

یکسری از روش های بصری سازی دیتا روش های مبتنی بر پیکسل هستند:
چطوری؟ ما بایم تک تک سپل ها رو تبدیل بکنیم به یک نقطه رنگی که حالا اینجا گری اسکیل است و هر نقطه معرف یک مقدار است که ما میایم بین سفید تا مشکلی بهش نسبت می دیم
حالا میخوایم ببینیم داده ها چطوری هستن؟ --> فرض کنیم داده ها، داده های یک بانک است که سطح درامد ادم ها و میزان اعتباری که دارن و حجم تراکنشی که دارن و سنشون رو ما گرفتیم و میخوایم یک مسئله روی داده ها تعریف بکنیم --> اول می خوایم ببینیم چخبره توی این داده ها پس میان ادم ها رو براساس سطح درامشون سورت میکنیم مثل سطح درامد کم رو صفر می ذاریم که اینجا ینی سفید میشه و بیشترین سطح درامد یک ینی سیاه

معادل همین رو روی حجم تراکنش ها و سطح اعتبار و سنشون هم انجام میدیم --> ینی کمترین سن سفید و بیشترین سن سیاه ..

a : سطح درامد

b : میزان اعتبار

c : تراکنش

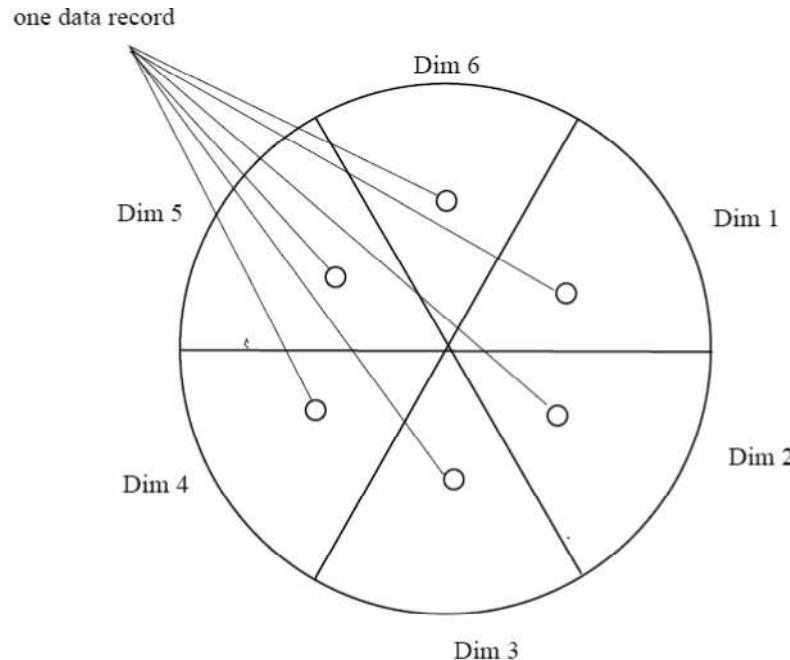
d : سن

کسی که سطح درامش صفر بوده --> اعتبارش صفر بوده --> تراکنش کم بوده و سنش هم خیلی زیاد بوده

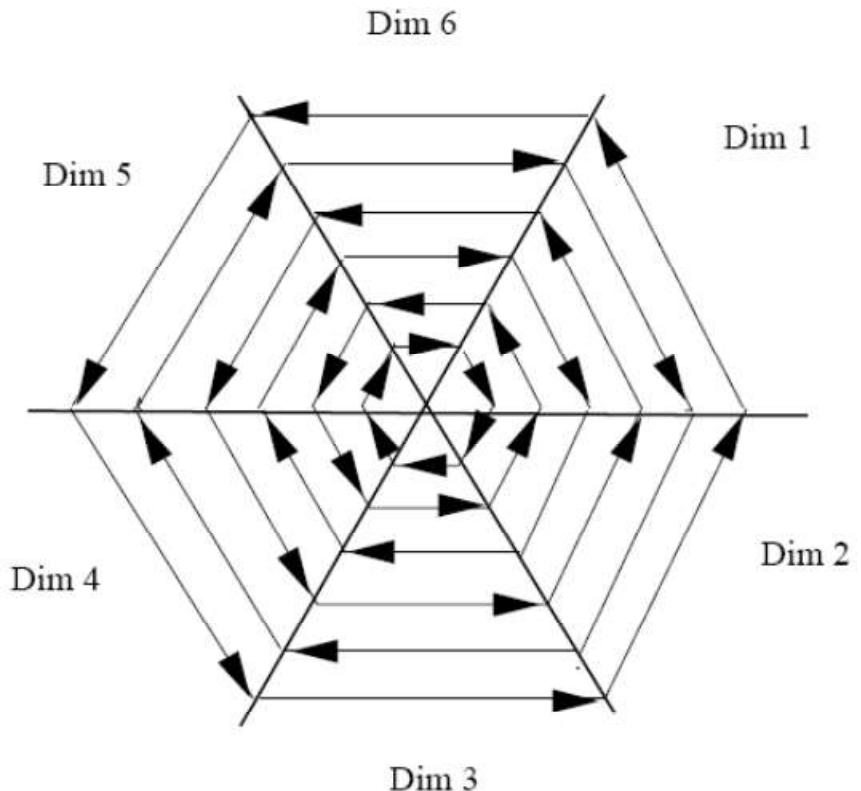
نکته: اگر به اینا پشت سر هم نگاه بکنیم یک الگویی تو ش به دست میاد که بهمون کمک میکنه تغییرات سطح درامد چه ربطی به تغییرات سطح اعتبار داره و تغییرات تراکنش ها و تغییرات سن داره --> با یک نگاه می تونیم بفهمیم تغییرات سطح درامد هیچ ربطی به سن نداره ولی سطح درامد یه ربطی به حجم تراکنش ها و میزان اعتبار داره

Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment



(a) Representing a data record in circle segment



(b) Laying out pixels in circle segment

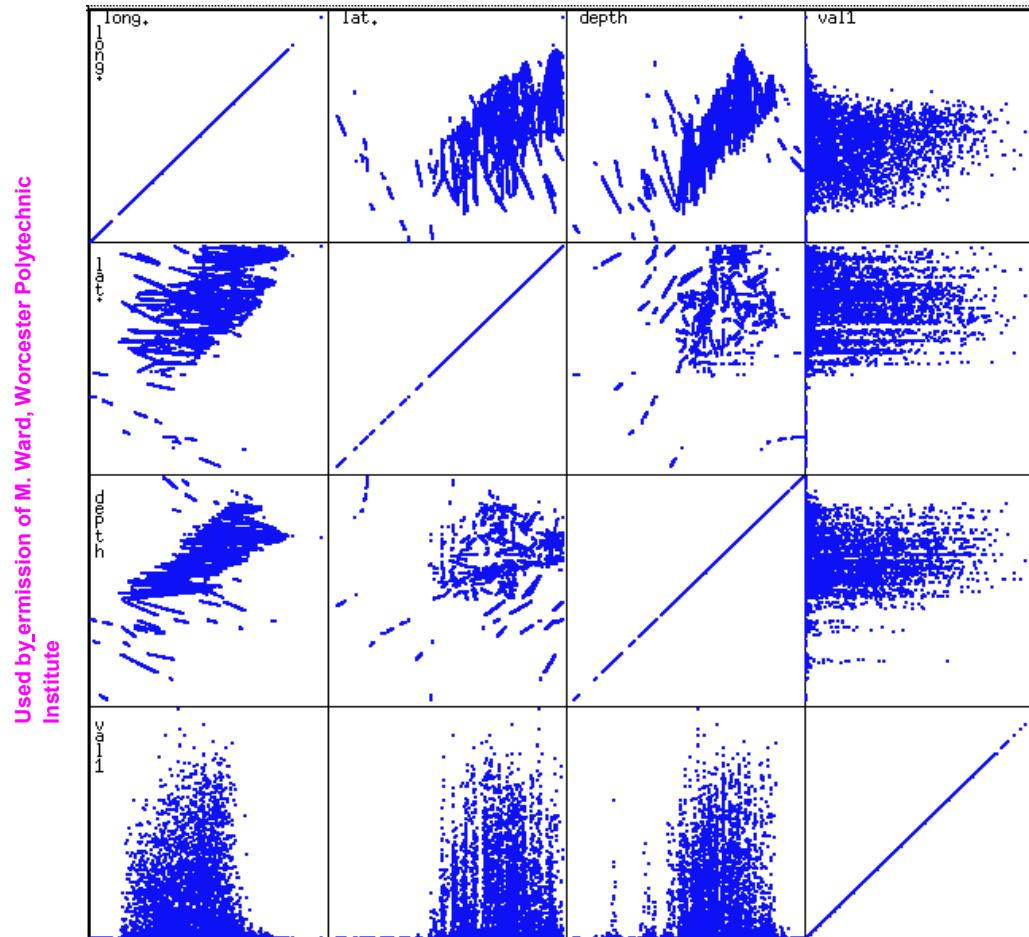
Geometric Projection Visualization Techniques

- Visualization of geometric transformations and projections of the data
- Methods
 - Scatterplot and scatterplot matrices
 - Landscapes
 - Parallel coordinates

-
یک دسته از روش ها، روش هایی هستند که به صورت هندسی با مسئله بخورد میکنن --> یکش Scatterplot است

Scatterplot : ما ببایم ویژگی های مختلفی که داریم رو دو به دو باهاشون نمودار بکشیم
Landscapes : ما میخوایم یک پدیده ای رو که جنس مکان داره رو گزارش بکنیم مثلاً توی بحث های جغرافیایی ما میخوایم عوارض زمین رو ببینیم --> چند صفحه جلوتر...

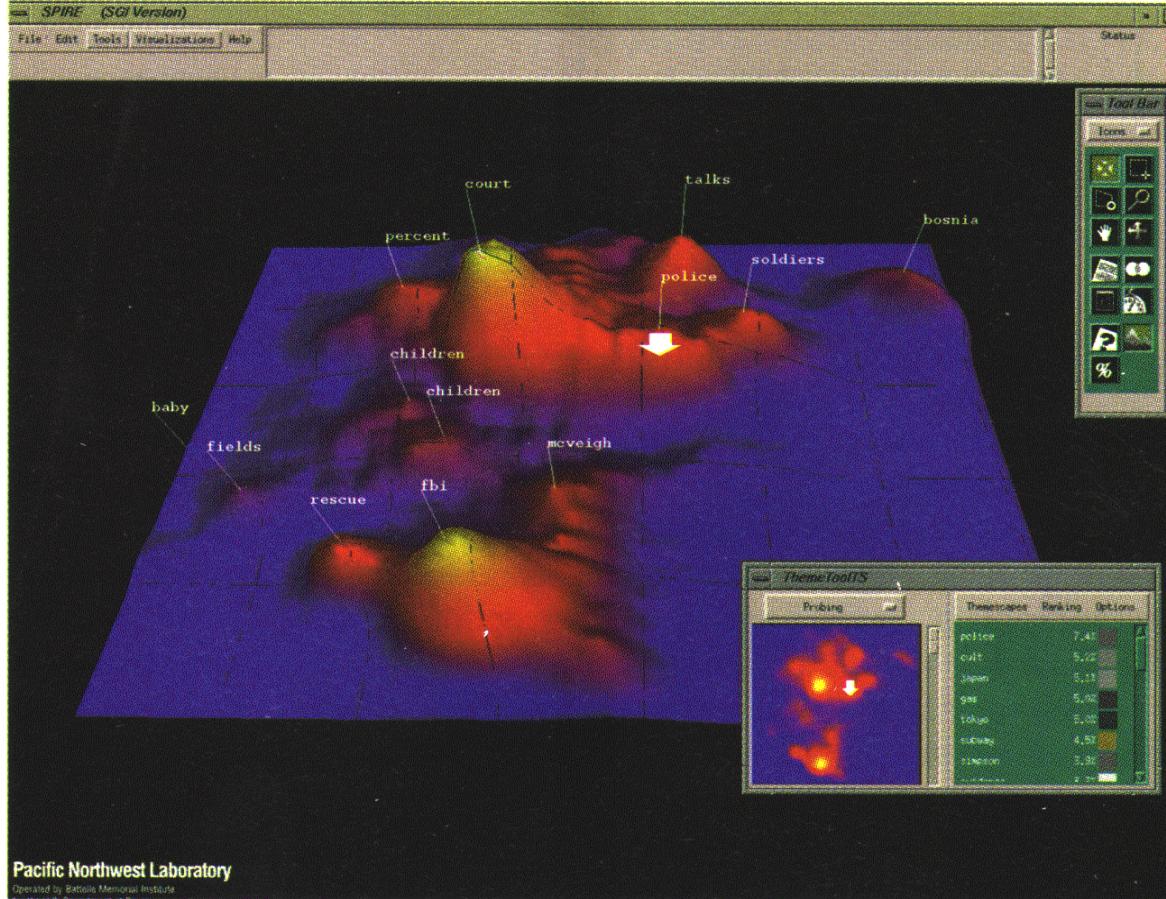
Scatterplot Matrices



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of ($k^2/2-k$) scatterplots]

Landscapes

Used by permission of B. Wright, Visible Decisions Inc.



news articles
visualized as
a landscape

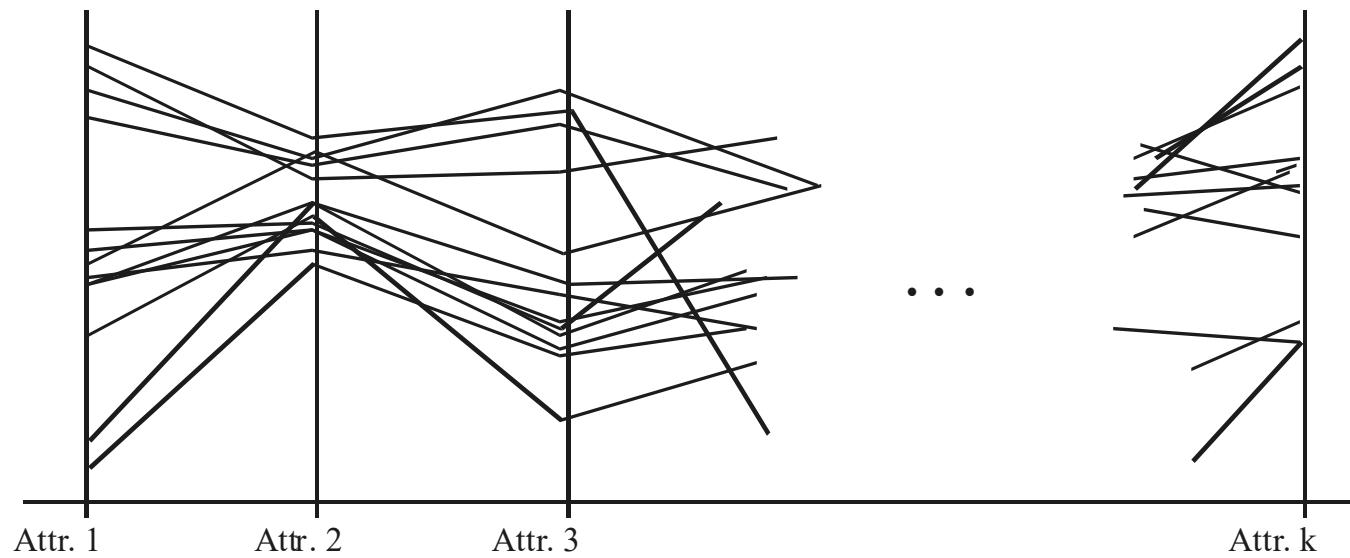
- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

ادامه ...

محور y , X میشه موقعیت مکانی ما و محور Z که یک مشخصه است مثلا سطح ارتفاع است اینو میایم با شدت رنگ یا ارتفاع مشخص میکنیم

Parallel Coordinates

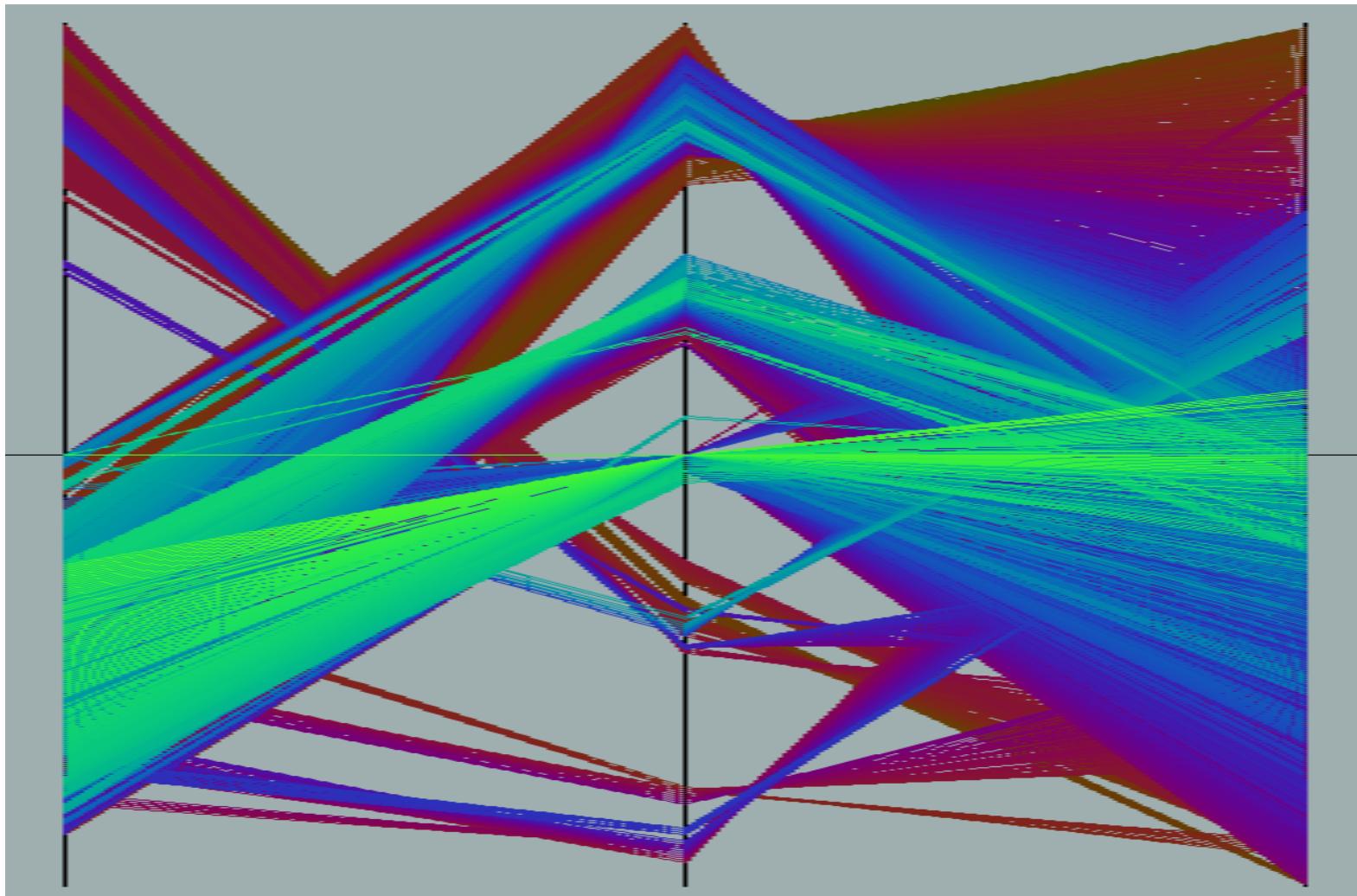
- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute



-
Parallel Coordinates : این نمودار میاد به هر کدام از ویژگی های مختلف یک ستون نسبت میده مثلا k تا مشخصه داریم و هر مشخصه ای که مینیم داره و یک ماکزیممی داره--> مینیم و ماکزیمم هاشو توی این ستون ها مشخص میکنیم
حالا یک نمونه ای داریم که توی ویژگی یک مقدارش کمینه شده و توی ویژگی دوم هم مقدارش کمینه شده و توی سوم هم کمینه شده و... و میایم یک خطی نسبت میدیم بهش و اینجوری توده سمپل ها دستمون میاد

مزیت این نمودار نسبت به بقیه نمودارها اینه که به راحتی می تونیم تعداد ستون ها رو زیاد بکنیم
ینی یک دید چند بعدی پیدا بکنیم

Parallel Coordinates of a Data Set

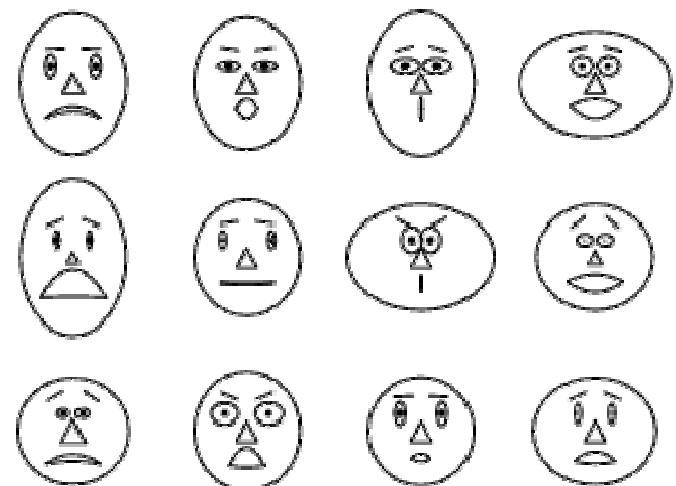


Icon-Based Visualization Techniques

- Visualization of the data values as features of icons
- Typical visualization methods
 - Chernoff Faces
 - Stick Figures
- General techniques
 - Shape coding: Use shape to represent certain information encoding
 - Color icons: Use color icons to encode more information
 - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

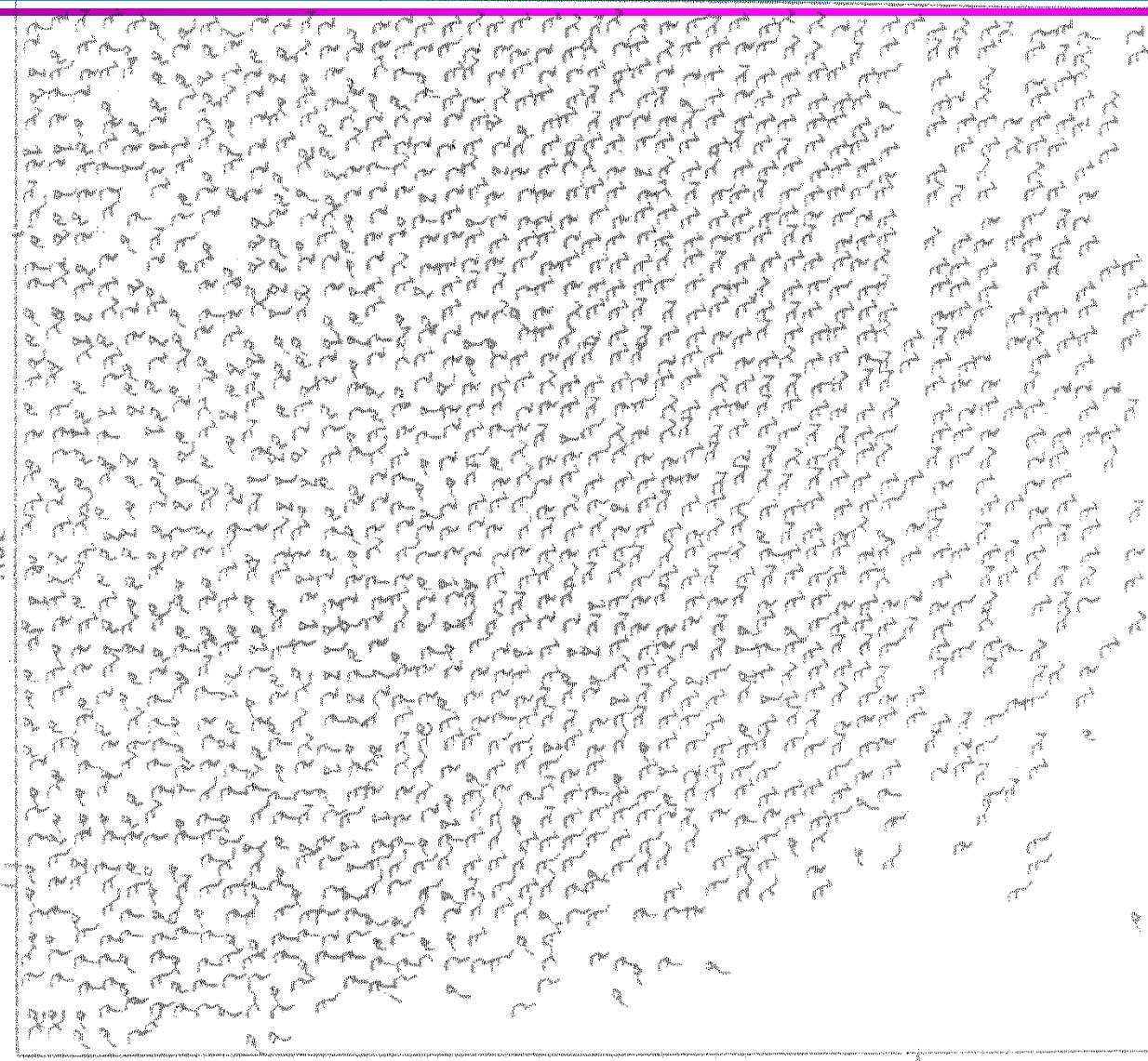
Chernoff Faces

- A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.
- The figure shows faces produced using 10 characteristics--head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using *Mathematica* (S. Dickson)
- REFERENCE: Gonick, L. and Smith, W. *The Cartoon Guide to Statistics*. New York: Harper Perennial, p. 212, 1993
- Weisstein, Eric W. "Chernoff Face." From *MathWorld--A Wolfram Web Resource*. mathworld.wolfram.com/ChernoffFace.html



Stick Figure

used by permission of G. Grinstein, University of Massachusetts at Lowell



INCOME

60

Two attributes mapped to axes, remaining attributes mapped to angle or length of limbs". Look at texture pattern

A census data figure
showing age, income,
gender, education, etc.

A 5-piece stick figure (1
body and 4 limbs w.
different angle/length)

SIMILARITY AND DISSIMILARITY MEASURES

Similarity and Dissimilarity Measures

- Similarity measure
 - Numerical measure of how alike two data objects are.
 - Is higher when objects are more alike.
 - Often falls in the range [0,1]
- Dissimilarity measure
 - Numerical measure of how different two data objects are
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- Proximity refers to a similarity or dissimilarity

خیلی جاها نیاز داریم که اتریبیوت ها و ابجکت ها رو با هم دیگه مقایسه کنیم مثلًا می خوایم یک اتریبیوت با یک اتریبیوت دیگه بسنجدیم که چقدر این ها با هم دیگه شبیه هستند یه جاهایی هم نه می خوایم ابجکت رو با یک ابجکت دیگه مقایسه کنیم --> پس شباخت توی دوتا بعد است یه جاهایی ما توی سطرها می خوایم بسنجدیم یه جاهایی توی ستون ها

دو دسته معیار است که معروف هستند به Proximity یعنی معیار های شباهت و تفاوت

: سعی میکنند این شباهت ها رو کمی بکنن برای ما که ما به صورت کمی بفهمیم این دوتا ابجکت ها چقدر با هم متفاوت است --> هر چقدر بزرگتر باشه ما میگیم این دوتا ابجکت بهم دیگه شبیه نرنده --> عموماً Similarity بین بازه 0 تا 1 هستن Dissimilarity بر عکس بالایی هستن --> تفاوت بین داده یا دوتا ابجکت رو مشخص میکنند: کفشوں 0 است و سطح بالایی ندارند اینا

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, x and y , with respect to a single, simple attribute.

Attribute Type
Nominal
Ordinal
Interval or Ratio

-
انواع اتریبیوت:
ما سه تا اتریبیوت داریم

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, x and y , with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = x - y /(n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = x - y $	$s = -d, s = \frac{1}{1+d}, s = e^{-d}, s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

nominal : مثل یک غذایی یک طعم رو داره و یک غذای دیگه یک طعم دیگه --> مقایسه این دو تا

ordinal : کیفی است

interval , ordinal
Similarity صفر بشه شباخت نداشتن و یک بشه شباخت ماکزیمم داشتن

اختلاف یا d --> کوچکتر اختلاف 0 است --> برای e به توان d - اگر d صفر باشه جواب نهایی میشه یک و بیشترین اختلاف میشه بی نهایت یعنی d شده بی نهایت و برای e به توان d - میشه صفر پس این اختلاف رو بین صفر و یک یه جورایی نرمالش میکنه

Euclidean Distance

- Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

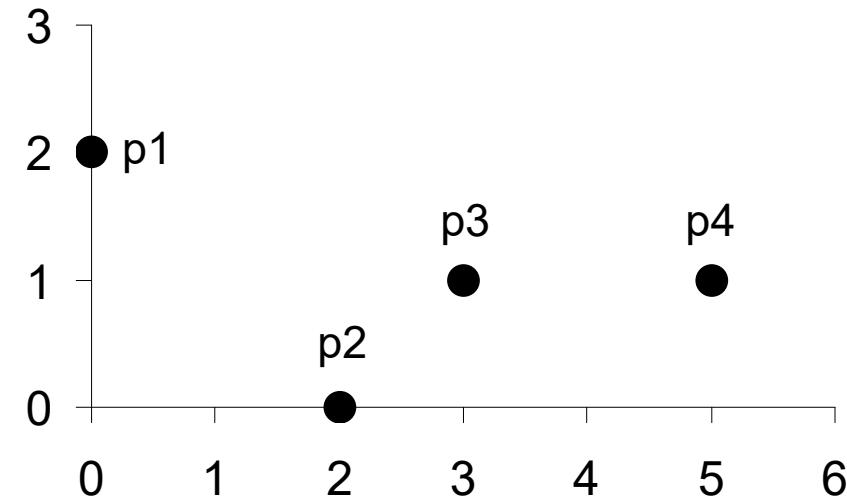
where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects \mathbf{x} and \mathbf{y} .

- Standardization is necessary, if scales differ.

-
یک معیار دیگه برای تفاوت فاصله اقلیدسی است
اینجا میخوایم دو تا ابجکت رو مقایسه بکنیم

نکته: استاندارد سازی لازمه اگر قراره طیف ها متفاوت باشه ینی اتریبیوت ها متفاوت باشه مثلا داریم با اطلاعات یک دانشجو کار میکنیم و قدش بین 100 سانتی متر تا 200 سانتی متر و وزنش بین 50 کیلو تا 150 کیلو و نمرش بین 0 تا 20 باشه --> اینکه بخوایم برای این دو تا دانشجو روی همه مقادیرش کار بکنیم کار درستی نیست مثلا جنس اختلافی که توی معدل ها پیش میاد نهایتا 20 نمره است ولی توی قد ها ممکنه 100 واحد باشه اختلافش --> توصیه میشه وقتی که ابجکت ها همچین شرایطی دارند تک تک این اتریبیوت ها رو استانداردش بکنیم

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

اینجا 4 تا ابجکت داریم که دو تا اتریبیوت دارند : x , y

وقتی فاصله بین ابجکت های مختلف رو می سنجیم یک ماتریسی به دست میاد به نام Distance --> که این میاد فاصله نظیر به نظیر تک تک ابجکت ها رو گزارش میکنه

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^n |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects x and y .

فاصله Minkowski این تعمیم یافته فاصله اقلیدسی است به این صورت که توى فاصله اقلیدسی همین فرم رو داشتیم فقط به جای 2 توان 2 داشتیم

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
 - A common example of this for binary vectors is the Hamming distance, 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.

- اگر $r = 1$ باشه نرم یک به دست میاد --> معروف به Manhattan, taxicab, L1 norm مثال: ما اطلاعات دوتا دانشجو رو به صورت باینری داریم نشون میدیم یا محتوای یک داکیومنت رو که چه کلماتی داخلش هست او مدیم با یک دنباله باینری مشخص کردیم که ایا توی دیکشنری ما کلمه اول توی داکیومنت او مده یا کلمه دوم او مده --> که این یک دنباله میشه --> می تونیم این داکیومنت رو با یک دنباله صفر و یک نشونش بدیم (از بودن یا نبودن این کلمات) برای مقایسه این 2 تا داکیومنت یک معیار برای فاصله سنجی این دوتا داکیومنت این است که ببینیم نظیر به نظیر ایا کلمات با هم برابر است یا نه

$r = 2$ میشه فاصله اقلیدسی

$r = \infty$ بی نهایت -->

$r = 0$ --> (به شرطی یه عددی به توان صفر میشه یک که اون عدد صفر نباشه و اگر اون عدد صفر باشه میشه صفر)

نقشه اول

$$f_{ix} = \left(\sum_{\neq 0} ((x_i - y_i) + (x_c - y_r)) \right)$$

مثلاً توی این مثال ما x ها رو گذاشتیم صفر و y های مختلف بهش میدیم به صورت تصادفی -->
در نهایت شکل رو رسم میکنیم

Minkowski Distance

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

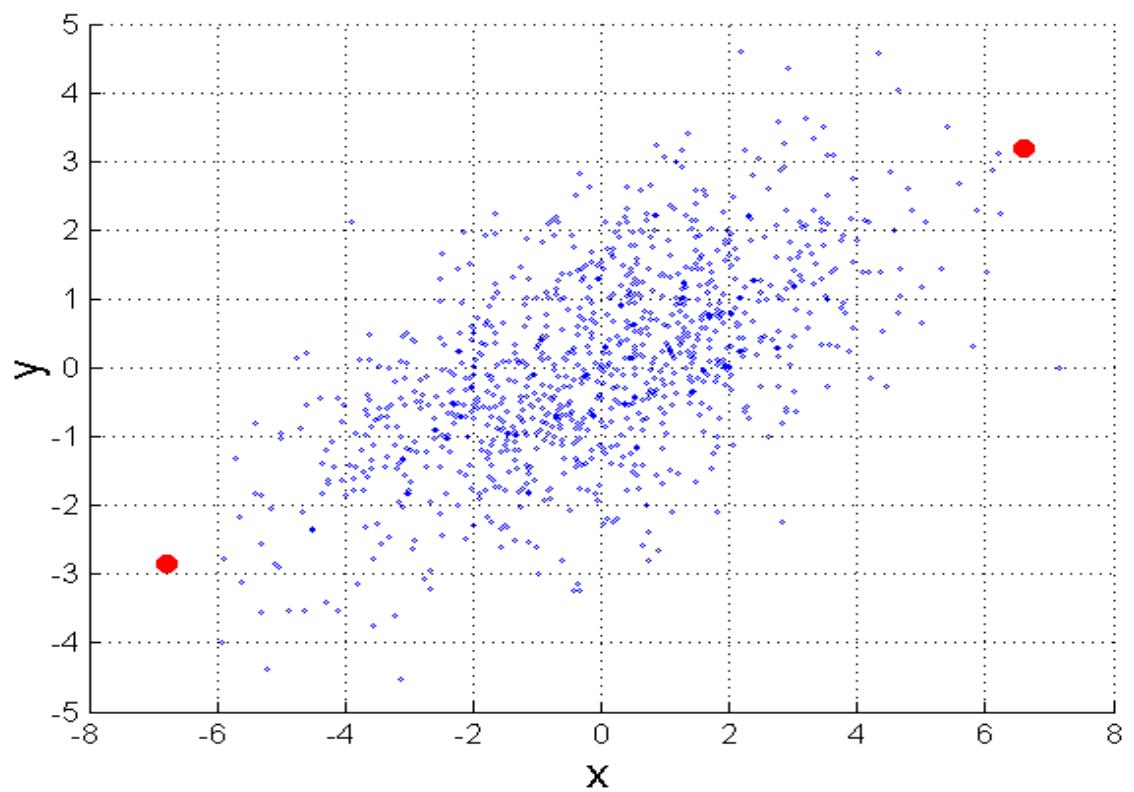
L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

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

Mahalanobis Distance



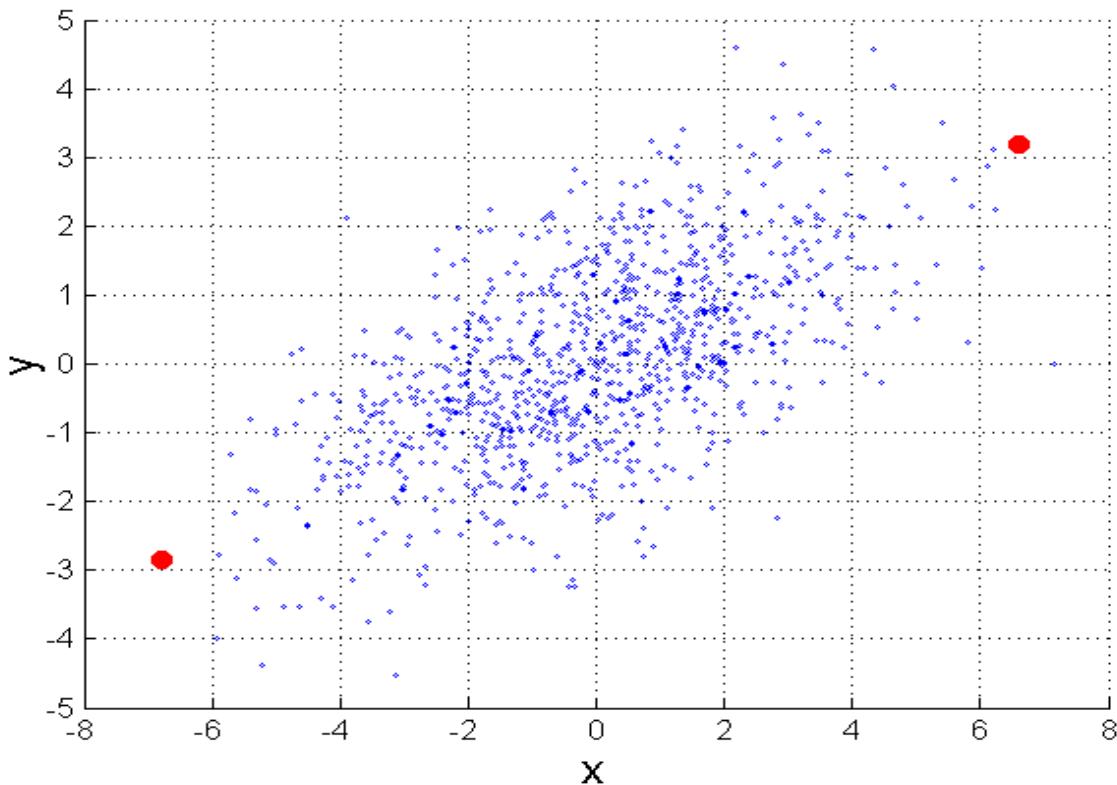
فاصله Mahalanobis:

یه جاهایی که می خوایم اختلاف بین ابجکت ها رو بسنジم برآمون مهم میشه که بقیه داده ها چجوری پراکنده شدن توی فضای نمونه --> مثلا میخوایم دوتا ادم رو با هم بسنジم برآمون مهمه که بقیه ادم ها چجوری هستن و نسبت به بقیه ادم ها این دوتا ادم چقدر به هم شبیه هستند مثال عینیش --> ینی داریم ادم هایی که مقایسه میکنیم توی عرف خودشون نگاهشون میکنیم

Mahalanobis میاد توزیع داده ها رو نگاه میکنه و Mahalanobis میگه که این اختلافی که داری حساب میکنی دقیقا توی یک مسیری است که توده های داده های ما توی همین مسیر پراکنده شده اند و میخواد این اطلاعات رو در نظر بگیره ینی اطلاعاتی که توده جمعیت قرار گرفته اند رو برای فاصله سنجی در نظر میگیره و میاد اینو یه جوابایی وزن دهی میکنه ینی ابعاد مختلف رو وزن دهی میکنیم

برای اینکه بخوایم توده داده ها و پراکنده داده ها حول میانگین رو بسنジم نیاز داریم به یک ماتریسی به نام covariance matrix

Mahalanobis Distance



Σ is the covariance matrix

$$\begin{bmatrix} \text{Var}(x_1) & \dots & \text{Cov}(x_n, x_1) \\ \vdots & \ddots & \vdots \\ \text{Cov}(x_n, x_1) & \dots & \text{Var}(x_n) \end{bmatrix}$$

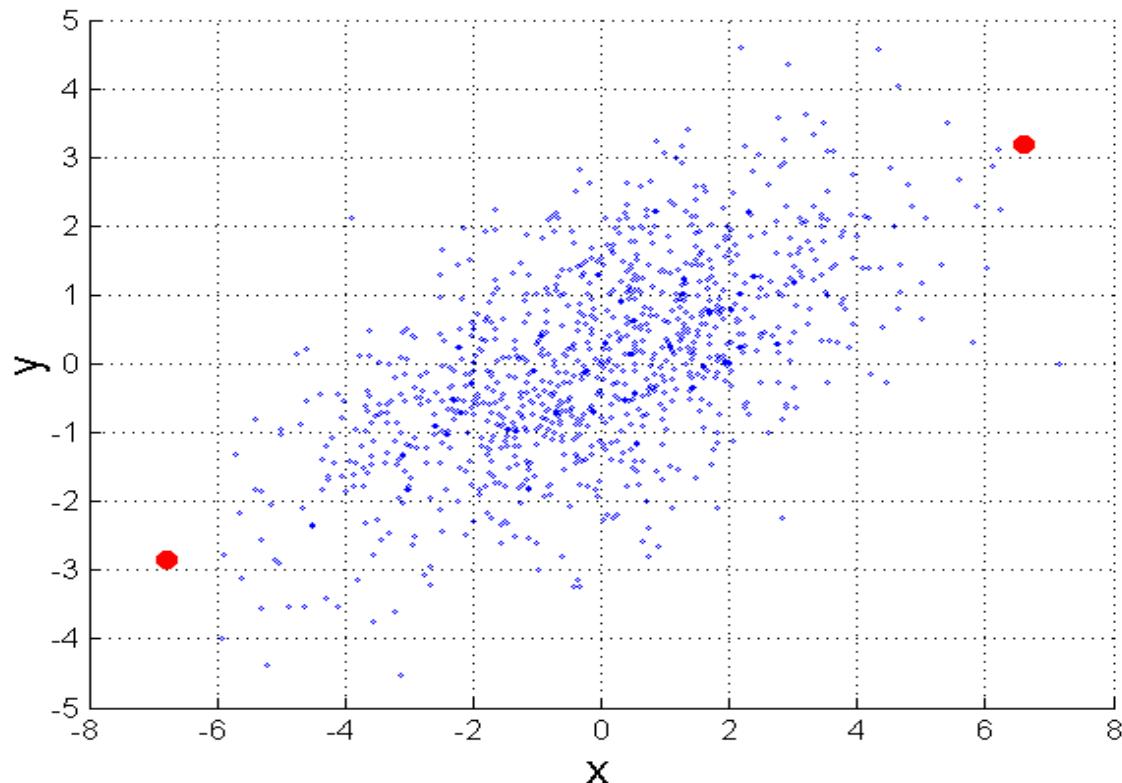
$$\text{var}(x) = \frac{\sum_1^n (x_i - \mu)^2}{n}$$

$$\text{cov}(x, y) = \frac{\sum_1^n (x_i - \mu_x)(y_i - \mu_y)}{n}$$

که این covariance عناصر روی قطر اصلیش است ینی واریانس توی اون بعد است

Mahalanobis Distance

$$\text{mahalanobis}(\mathbf{x}, \mathbf{y}) = ((\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y}))^{-0.5}$$

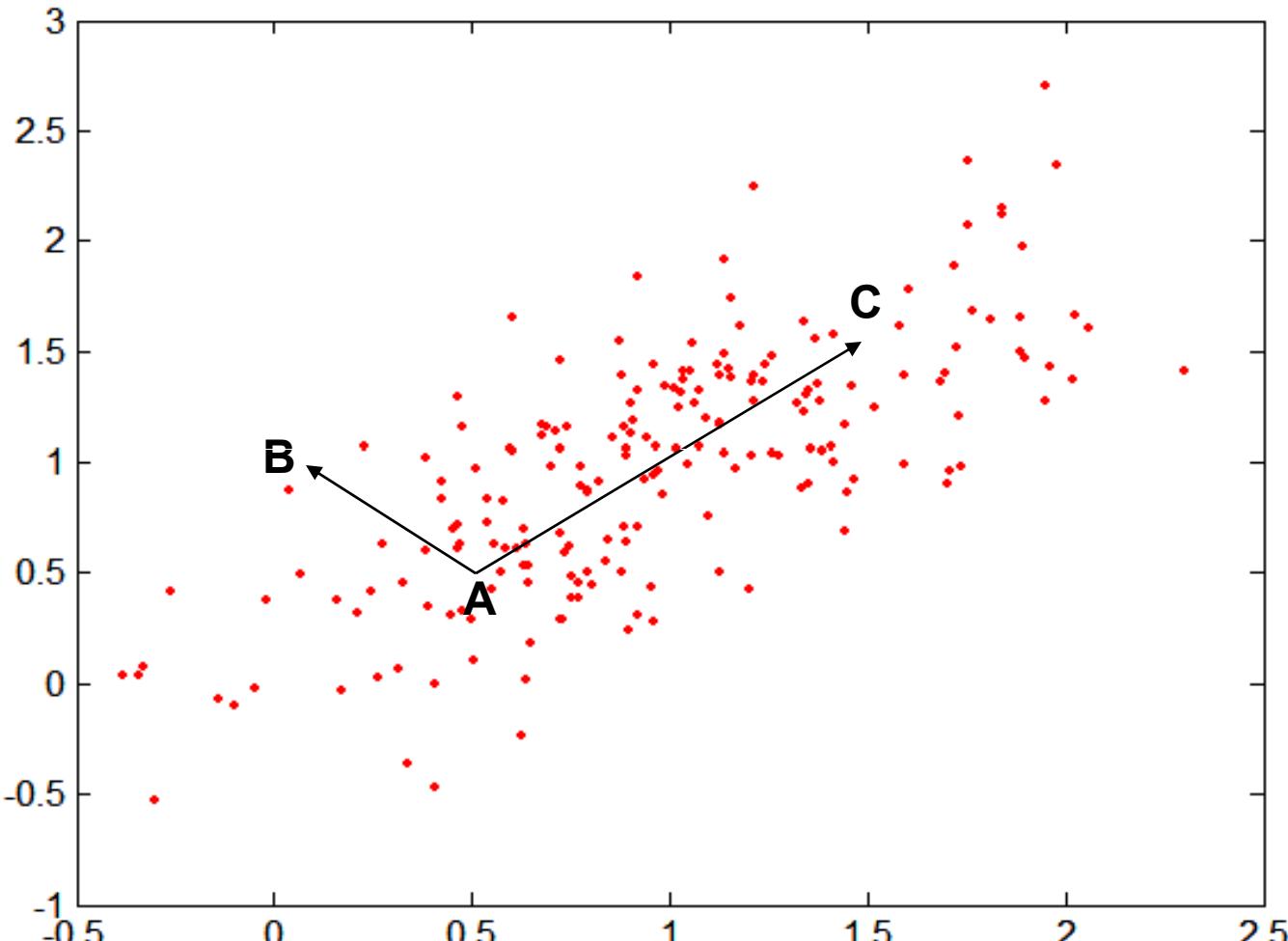


Σ is the covariance matrix

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

-
T ترانهاده است

Mahalanobis Distance



Covariance
Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

$\text{Mahal}(A,B) = 5$

$\text{Mahal}(A,C) = 4$

نکته:

فاصله A, B شده 5 ولی فاصله A, C شده 4 : این ینی این که چون C در جهت پراکندگی دیتا است مقدارش کمتره تا A, B که در خلاف پراکندگی دیتا است پس به توزیع داده ها دقیق میکنه و اگر قراره داده های ما توانی توزیع داده ها بگیرند فاصله های دور اقلیدسی رو کوتاه میکنه

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$


$$A^T = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$


Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.
 1. $d(x, y) \geq 0$ for all x and y and $d(x, y) = 0$ if and only if $x = y$.
 2. $d(x, y) = d(y, x)$ for all x and y . (Symmetry)
 3. $d(x, z) \leq d(x, y) + d(y, z)$ for all points x, y , and z .
(Triangle Inequality)

where $d(x, y)$ is the distance (dissimilarity) between points (data objects), x and y .

- A distance that satisfies these properties is a **metric**

- 1- فاصله ها همیشه مقدارهای بزرگتر از صفر دارند و فاصله صفر زمانی است که دو تا مقدار ورودی با هم دیگه برابر باشند
- 2- فاصله ها حالت Symmetry دارند
- 3- نامساوی مثلثی توشون صدق میکنه

نکته: اگر فاصله ای توی این سه تا مشخصه بخونه ما به اون میگیم یک metric

Common Properties of a Similarity

- Similarities, also have some well known properties.
 1. $s(x, y) = 1$ (or maximum similarity) only if $x = y$.
(does not always hold, e.g., cosine)
 2. $s(x, y) = s(y, x)$ for all x and y . (Symmetry)

where $s(x, y)$ is the similarity between points (data objects), x and y .

Similarity Between Binary Vectors

- Common situation is that objects, x and y , have only binary attributes
- Compute similarities using the following quantities

f_{01} = the number of attributes where x was 0 and y was 1

f_{10} = the number of attributes where x was 1 and y was 0

f_{00} = the number of attributes where x was 0 and y was 0

f_{11} = the number of attributes where x was 1 and y was 1

- Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes

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

J = number of 11 matches / number of non-zero attributes

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

اين معيار برای کدهای باينری خیلی استفاده میشه:

مثلا دوتا دنباله باينری داریم --> مثلا یک دنباله باينری داریم برای این خبر و یک دنباله دیگر هم برای خبر بعدی حالا میخوایم ببینیم این دوتا خبر چقدر بهم شبیه هستند --> یک معیاری که خیلی استفاده میشه L و SMC است

SMC versus Jaccard: Example

x = 1 0 0 0 0 0 0 0 0 0

y = 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$ (the number of attributes where **x** was 0 and **y** was 1)

$f_{10} = 1$ (the number of attributes where **x** was 1 and **y** was 0)

$f_{00} = 7$ (the number of attributes where **x** was 0 and **y** was 0)

$f_{11} = 0$ (the number of attributes where **x** was 1 and **y** was 1)

$$\begin{aligned}\text{SMC} &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00}) \\ &= (0+7) / (2+1+0+7) = 0.7\end{aligned}$$

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

از لحاظ SMC میگه 0.7 این دوتا داکیومنت بهم شبیه هستند و از لحاظ L میگه نیستن

Cosine Similarity

- If \mathbf{d}_1 and \mathbf{d}_2 are two document vectors, then

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\|,$$

where $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, \mathbf{d}_1 and \mathbf{d}_2 , and $\|\mathbf{d}\|$ is the length of vector \mathbf{d} .

- Example:

$$\mathbf{d}_1 = \begin{matrix} 3 & 2 & 0 & 5 & 0 & 0 & 0 & 2 & 0 & 0 \end{matrix}$$

$$\mathbf{d}_2 = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 \end{matrix}$$

این وقوع تعداد کلمات است مثلاً وقوع 10 تا کلمه رو
چک کردیم و مثلاً یکی از اون کلمات 3 بار تکرار شده

ضرب داخلی $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$$\|\mathbf{d}_1\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$\|\mathbf{d}_2\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.449$$

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$$

: Cosine Similarity

برای پردازش متن خیلی کاربرد دارد

اینجا ممکن است دو داده داشت که دو داده بین آنها $d_1, d_2 \dots$ هستند

وقتی میخوایم با پردازش متن کار بکنیم ساده ترین کار این است که یکسری کلمات کلیدی انتخاب بکنیم و بینیم این کلمات کلیدی توی این داده هست یا نه

این کلمات کلیدی هم جوری برداریم که این کلمات، خوب بتوانیم این داده ها را بر اساس توصیف بکنیم

یک روش بازگشتی بود که بودن یا نبودن کلمات را می سنجد

تعداد تکرار کلمات هم می تواند موثر باشد برای اینکه بفهمیم چه داده هایی بیشتر راجع به چه موضوعی است

Correlation measures the linear relationship between objects

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard_deviation}(\mathbf{x}) * \text{standard_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y}, \quad (2.11)$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \quad (2.12)$$

$$\text{standard_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

$$\text{standard_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (y_k - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k \text{ is the mean of } \mathbf{x}$$

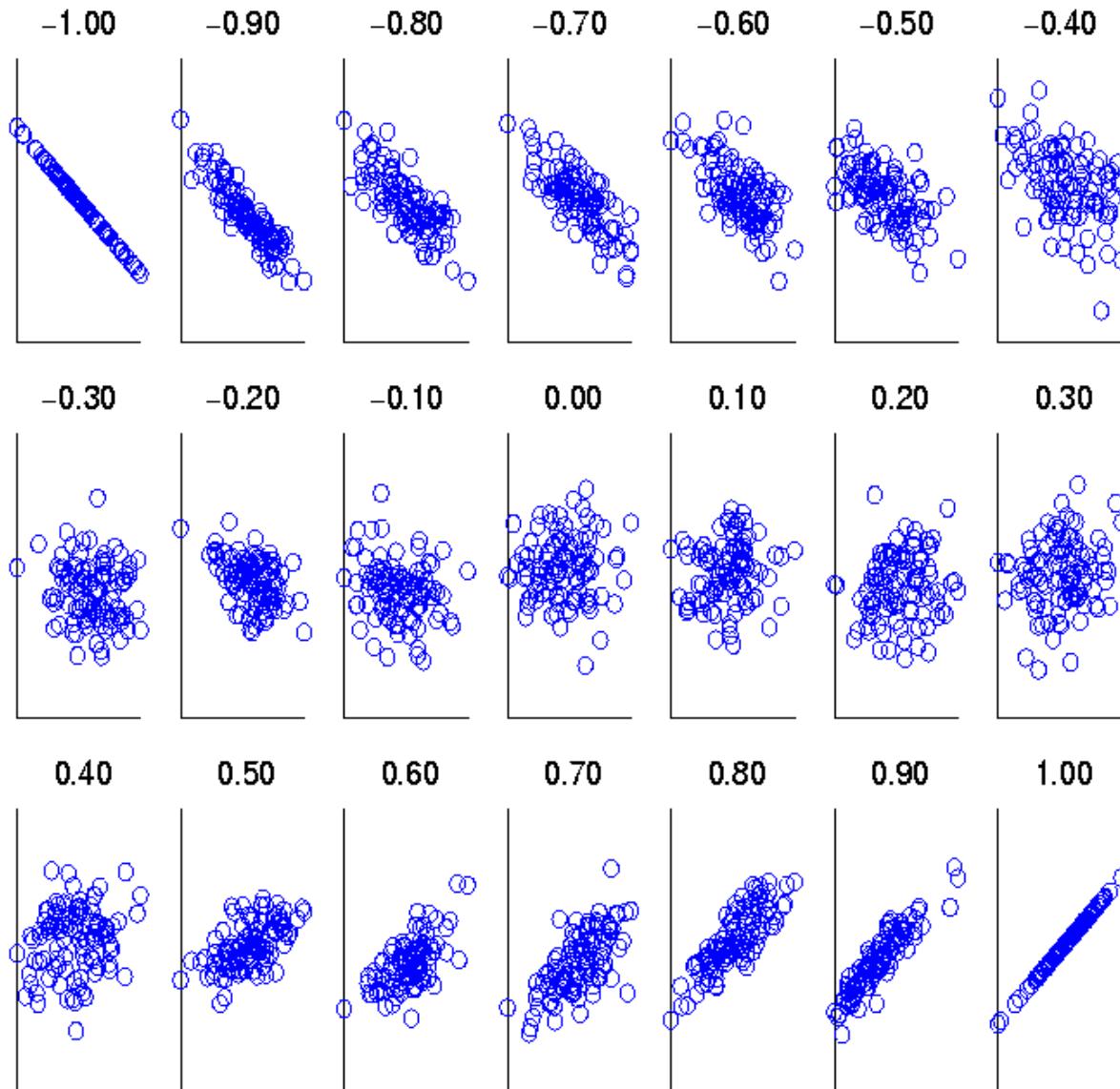
$$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k \text{ is the mean of } \mathbf{y}$$

معیار :Correlation

یه جاهایی پیش میاد که ما میخوایم بین دوتا اتریبیوت یا دوتا ستون مثلًا قدم ادم ها یا وزن ادم ها می خوایم ببینیم ارتباطی بین این ها وجود داره یا نه یکی از معیارهای معروف Correlation است که این Correlation هم تایپ های مختلفی داره تعريفش براساس کوواریانس به دست میاد Correlation

از لحاظ اندازه مقدارش بین صفر و یک است --> یک ینی بیشترین ارتباط بینشون وجود داره و صفر ینی ارتباط خطی بینشون وجود نداره

Visually Evaluating Correlation



**Scatter plots
showing the
similarity from
–1 to 1.**

یکسری ابجکت داریم که این دو تا متغیر را او مدیم گزارش کردیم روی صفحه Scatter plots رسماً شون کردیم

توی این تصویر او مده شرایط مختلف Correlation رو مشخص کرده که کی یک میشه یا اگر داده ها چه شکلی بودن Correlation یک میشه یا Correlation صفر یا منفی یک --> اگر مقادیری که دو تا بعد دارند کاملاً ضد هم عمل بکنند یعنی افزایش یک بعد منجر به کاهش یک بعد دیگه بشه این Correlation ما منفی تر میشه ولی اگر همنوا با هم باشند یعنی با افزایش یک بعد، بعد دیگر هم افزایش پیدا بکنه ما میگیم Correlation یک داریم یعنی بیشترین سطح Correlation رو داریم و وسط اون ها هم Correlation صفر است یعنی افزایش یا کاهش شون خیلی معنا دار نیست

هر چقدر این Correlation بزرگتر باشه همنوایی چشم گیرتر است

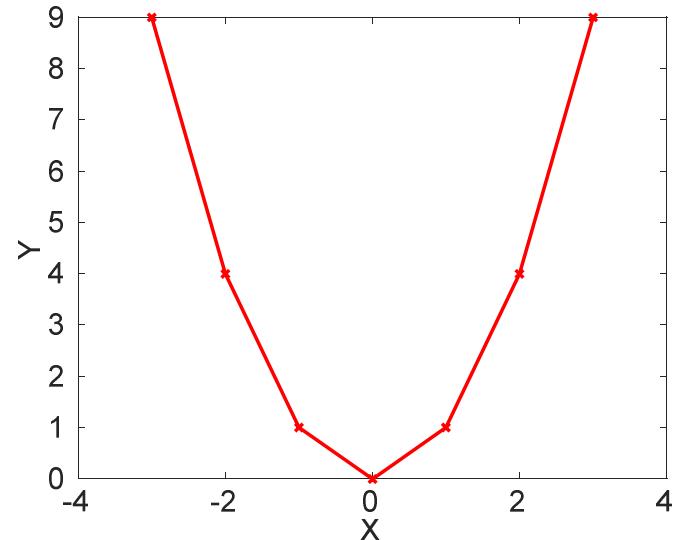
Drawback of Correlation

- $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$
- $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$

$$y_i = x_i^2$$

- $\text{mean}(\mathbf{x}) = 0$, $\text{mean}(\mathbf{y}) = 4$ میانگین
- $\text{std}(\mathbf{x}) = 2.16$, $\text{std}(\mathbf{y}) = 3.74$

$$\begin{aligned}\bullet \text{corr} &= (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74) \\ &= 0\end{aligned}$$



این Correlation یک مشکلی دارد:

فرض کنیم داده های ما یک همچین فرمی را داشته باشند --> فرض کنیم یک اتربیبیوت را از داده ها اندازه گیری کردیم و اسمش X است و یک اتربیبیوت دیگه هم که یک رابطه توان دویی داره با X به اسم y

توی این داده ای که جمع اوری کردیم متغیر X ارتباط داره با متغیر y یا نه؟

(توی صفحه قبل گفتم Correlation می تونه بهمون بگه که ارتباط داریم یا نه --> اگر ارتباط نبود Correlation شون صفر میشد و اگر ارتباط بود Correlation یا یک میشد یا منفی یک) اینجا Correlation با هم ندارند

نکته: اینجا اتربیبیوت X با اتربیبیوت y در ارتباط است پس چرا Correlation اش صفر شد؟ اینجا نکته ای که وجود داره این است که Correlation میاد ارتباط های خطی را پیدا میکنه و اگر اتربیبیوت های ما ارتباطشون غیر خطی باشه یا توانی باشه لزوما Correlation معیار خوبی نیست چون می تونه ما رو بیراه بیره در حالی که این دوتا متغیر با هم ارتباط دارند Correlation صفر شد و اینجا ممکنه ما قضاوت اشتباه بکنیم و بگیم این دوتا متغیر بهم ارتباطی ندارند

Correlation vs Cosine vs Euclidean Distance

- Compare the three proximity measures according to their behavior under variable transformation
 - scaling: multiplication by a value
 - translation: adding a constant

Property	Cosine	Correlation	Euclidean Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

- Consider the example
 - $\mathbf{x} = (1, 2, 4, 3, 0, 0, 0)$, $\mathbf{y} = (1, 2, 3, 4, 0, 0, 0)$
 - $\mathbf{y}_s = \mathbf{y} * 2$ (scaled version of y), $\mathbf{y}_t = \mathbf{y} + 5$ (translated version)

Measure	(\mathbf{x}, \mathbf{y})	$(\mathbf{x}, \mathbf{y}_s)$	$(\mathbf{x}, \mathbf{y}_t)$
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

تفاوت معیارهای Correlation vs Cosine vs Euclidean :

ممکنه ما بخوايم يکسری توابع رو اعمال بکنيم روی اtribut ها و اين مهم ميشه که اين معیار فاصله اي که داريم می سنجيم نسبت به اون توابع بسته است یا نه وقت هايي که میخوايم فاصله Cosine بسنجيم بين دوتا اtribut اگر اون اtribut ها رو اسکيل بکنيم ينی يک مقداری توشن ضرب بکنيم مشکلی برای فاصله Cosine پيش نمیاد پس میگیم نسبت به ضرب بسته است ينی اتفاقی براش نمی افته ولی اگر يک مقداری بهش اضافه بکنيم يا ازش کم بکنيم ديگه فاصله Cosine که به دست میاد فاصله Cosine قبلش نیست پس نسبت به جابه جایی و جمع کردن بسته نیست توی Correlation این اتفاق می افته مثال:

فرض کنيد که يک دیتاست داریم که اtribut اول این دیتاست x است و اtribut دومش y است بعدش يک دیتاست جدید ازش ساختیم به این صورت که y هاش رو دوبرابر کردیم يکبار و ديگه هم يک مقداری با y ها جمع کردیم حالا فاصله های Correlation , Cosine , Euclidean چه تغييری میکنه؟

Correlation vs cosine vs Euclidean distance

- Choice of the right proximity measure depends on the domain
- What is the correct choice of proximity measure for the following situations?
 - Comparing documents using the frequencies of words
 - ◆ Documents are considered similar if the word frequencies are similar
 - Comparing the temperature in Celsius of two locations
 - ◆ Two locations are considered similar if the temperatures are similar in magnitude
 - Comparing two time series of temperature measured in Celsius
 - ◆ Two time series are considered similar if their “shape” is similar, i.e., they vary in the same way over time, achieving minimums and maximums at similar times, etc.

برای موقوعی که میخوایم داکیومنت ها رو با هم مقایسه بکنیم و تکرارهاشون رو داشته باشیم ینی تعداد تکرار کلمات رو داشته باشیم بیشتر می ریم سراغ معیار cosine

یه جاهایی میخوایم دمای بین دوتا منطقه رو مقایسه بکنیم از لحاظ Celsius که برای این می ریم سراغ معیار Euclidean چون اصلا زمان نداریم

نکته: وقتایی که می خوایم یک الگوی زمانی پیدا بکنیم Correlation پیشنهاد خوبی است اخریش هم Correlation میشه

Comparison of Proximity Measures

- Domain of application
 - Similarity measures tend to be specific to the type of attribute and data
 - Record data, images, graphs, sequences, 3D-protein structure, etc. tend to have different measures
- However, one can talk about various properties that you would like a proximity measure to have
 - Symmetry is a common one
 - Tolerance to noise and outliers is another
 - Ability to find more types of patterns?
 - Many others possible
- The measure must be applicable to the data and produce results that agree with domain knowledge

Information Based Measures

- Information theory is a well-developed and fundamental discipline with broad applications
- Some similarity measures are based on information theory
 - Mutual information in various versions
 - Maximal Information Coefficient (MIC) and related measures
 - General and can handle non-linear relationships
 - Can be complicated and time intensive to compute

- تئوری اطلاعات:

بعضی جاها ما رو به اشتباه میندازه مثلاً دوتا متغیر بهم ربط دارند ولی Correlation شون صفر شده Correlation

یک متریک دیگه ای که استفاده میشه برای شباخت سنجی روش هایی که مبتنی بر تئوری اطلاعات کار میکنه ینی Information Based Measures دارند

ویژگی که این تکنیک های مبتنی بر اطلاعات دارند اینه که ارتباطات های غیر خطی هم می تونیم باهاشون بفهمیم --> اگر دوتا متغیر ارتباطشون غیر خطی بودن توی روش های مبتنی بر اطلاعات اینارو مشخص میکنه

Information and Probability

- Information relates to possible outcomes of an event
 - transmission of a message, flip of a coin, or measurement of a piece of data
- The more certain an outcome, the less information that it contains and vice-versa
 - For example, if a coin has two heads, then an outcome of heads provides no information
 - More quantitatively, the information is related to the probability of an outcome
 - ◆ The smaller the probability of an outcome, the more information it provides and vice-versa
 - Entropy is the commonly used measure



وقتی داریم راجع به یک متغیر تصادفی صحبت میکنیم متغیرهای تصادفی یکسری پیشامد دارند و هر کدام از این پیشامدها رو ما می تونیم یک احتمال رخدادی برآشون در نظر بگیریم مثلاً راجع به پرتاب یک سکه --> سکه ممکنه رو بیاد یا پشت بیاد

اینکه هر کدام از این پیشامدها با چه احتمالی رخ میدن ما بهشون میگفتیم تابع توزیع اونها

اینجا یک رابطه ای وجود داره بین احتمال و اطلاعات --> مثلاً یک سکه رو میندازیم و این بیشتر اوقات رو میاد پس ما با یک احتمالی میگیم سری بعدی هم رو میاد --> هر چقدر اون پدیده ما تصادفی تر رفتار بکنه ما میگیم اطلاعات بیشتری توش هست ینی یک رابطه معکوسی وجود داره

Entropy

- For

- a variable (event), X ,
- with n possible values (outcomes), $x_1, x_2 \dots, x_n$
- each outcome having probability, $p_1, p_2 \dots, p_n$
- the entropy of X , $H(X)$, is given by

$$H(X) = - \sum_{i=1}^n p_i \log_2 p_i$$

- Entropy is between 0 and $\log_2 n$ and is measured in bits
 - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Entropy را چطوری اندازه گیری میکنند؟

مثلا یک متغیری داریم و این متغیر یکسری حالت داره مثلا متغیر ما رنگ موی ادم هاست و سه حالت برای رنگ مو وجود داره مثلا روشن - تیره و قهوه ای --> حالا هر کدام از این حالت های رنگ مو یک احتمال رخدادی داره
حالا میخوایم بگیم که توی وضعیت رنگ مو ادم ها چه میزان Entropy وجود داره ینی چه میزان اطلاعاتی وجود داره --> فرمول توی صفحه

تهش این فرمول یک عددی به ما میده که این عدد بین ۰ تا $\log n$ است

: میزان اطلاعاتی است که توی یک دنباله داده وجود داره Entropy

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails
- What is the entropy of a fair four-sided die?

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = -p \log_2 p - q \log_2 q$$

- For $p=0.5, q=0.5$ (fair coin) $H=1$
- For $p = 1$ or $q = 1$, $H = 0$

- What is the entropy of a fair four-sided die?

مثال:

یک سکه ای داریم که احتمال رو او مدنش p است و احتمال پشت او مدنش یک منهای p است

Entropy for Sample Data: Example

Hair Color	Count	p	$-p \log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

Maximum entropy is $\log_2 5 = 2.3219$

زمانی ما بزرگترین مقدار رو داریم که همسون با هم برابر باشند --> اگر همه احتمالاتمون با هم
برابر باشند این بزرگترین مقدار رو پیدا میکنه
چرا 5 ؟ چون ما 5 دسته داریم
وقتی که احتمال رخداد هر دسته $1/5$ باشه ؟؟

Entropy for Sample Data

- Suppose we have

- a number of observations (m) of some attribute, X ,
e.g., the hair color of students in the class,
- where there are n different possible values
- And the number of observation in the i^{th} category is m_i
- Then, for this sample

$$H(X) = - \sum_{i=1}^n \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

- For continuous data, the calculation is harder

Mutual Information

- Information one variable provides about another

Formally, $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where

$H(X, Y)$ is the joint entropy of X and Y ,

$$H(X, Y) = - \sum_i \sum_j p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together

- For discrete variables, this is easy to compute
- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y))$, where $n_X (n_Y)$ is the number of values of $X (Y)$

Mutual Information میاد از همین Entropy و بی نظمی کمک می گیره و ارتباط بین دو تا متغیر تصادفی رو برآمون کمی میکنه

Entropy روی یک متغیر تعریف میکنیم

Mutual Information میاد ارتباط بین دو تا متغیر رو تعریف میکنه

Mutual Information Example

توأم؛ Entropy

Student Status	Count	p	$-p \log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p \log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p \log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

Mutual information of Student Status and Grade = $0.9928 + 1.4406 - 2.2710 = 0.1624$

مثال:

میخوایم ببینیم ایا رابطه ای بین سطح تحصیلات و نمره دانشجو وجود داره؟ ایا ستون نمره و ستون وضعیت دانشجو رابطه ای بینشون است؟

کاری که باید انجام بشه این است که Entropy ستون اول رو حساب بکنیم و Entropy ستون دوم هم حساب بکنیم و توامشون هم حساب بکنیم

جواب وقتی یک میشه که:
???

جواب وقتی صفر میشه که:
??