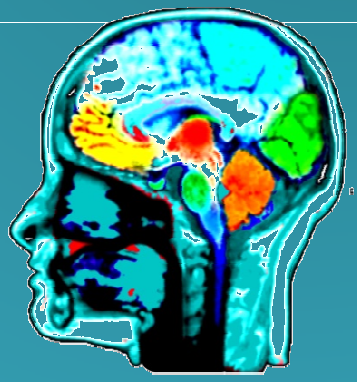
Introduction To Data Mining

Isfahan University of Technology (IUT)

Bahman 1401



Getting to Know Your Data

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

چرا سراغ ویژوالیزیشن میریم؟
یه دیدی از داده ها بمون میده، بمون کمک میکنه بفهمیم کل داده هامون توش چه خبره
پیداکردن یه دید کیفی از داده ها
پیدا کردن الگو در داده ها مثلا بگیم ابتدا یه کرلیشن مثبت داریم بعد یه کرلیشن منفی داریم
برای بهینه کردن پارامترها در انتخاب ابزارها تصمیم بگیریم چی را برداریم؟
ابزارهای مختلفی برای ویژوالیزیشن هست

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

دید کلی کیفی

با نگاشت داده ها بر روی نمونه های اولیه گرافیکی، بینش نسبت به فضای اطلاعاتی به دست می اوریم.
ارائه نمای کلی کیفی از مجموعه داده های بزرگ.
-جستجو برای الگوها، روندها، ساختار، بی نظمی ها، روابط بین داده ها.
به یافتن مناطق جالب و پارامترهای مناسب برای تجزیه و تحلیل کمی بیشتر کمک میکند.
ارائه یک مدرک بصری از نمایش های کامپیوتری به دست آمده

چهار یا پنج تا اتریبیوت داریم مثل درآمد، حد اعتبار، مقدار تراکنش، سن ابجکت هامون چیان؟کسانی که به بانک مراجعه کردن یا توی بانک حساب دارن ما میخایم ببینیم چه ارتباطی بین ویژگی های اون ابجکت ها وجود داره

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

مثلا اینجا که ۴تا اتریبیوت داریم، ۴تا نجره کشیده شده که مقدار اتریبیوت های هر رکورد یا ابجکت متناظرش رسم شده

از نظر سطح درامدی مرتب شدند اابجکت ها

اعتبار همون ابجکتی که مثلا سطح درامدش یه پیکسل خاصیه تو شکل اول رو میایم نگاه میکنیم یعنی این ویژگی محدودیت اعتبار رو هم اومدیم با بازه ی رنگ ها دسته بندی کردیم

به یه جایی از سطح درامد که رسیدیم تراکنش ها به شدت افزایش پیدا کرده که هیچ ربطی هم به سنشون نداره میتونه کم یا زیاد باشه سنشون

سطح درامد کم به زیاد

(a) Income (b) Credit Limit (c) transaction volume (d) age

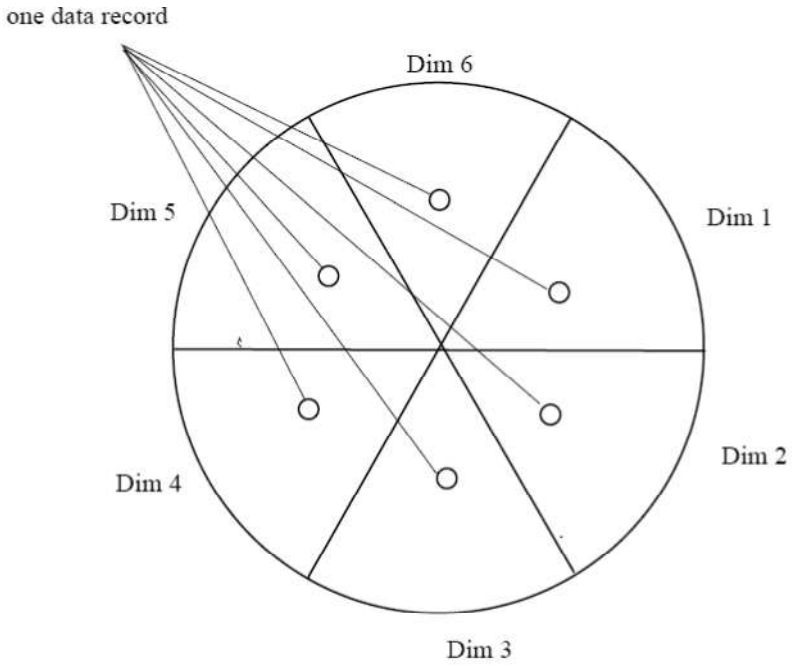
به ازای هر مقداری که سطح درامدشون داره، یه پیکسلی میسازیم که میزان سطح درامدشون با رنگ پیکسله مرتبطه وقتی سطح درامد زیاد شد، رنگ مشکی تر میشه اگه سطح درامد کم شد رنگ سفید

هرپیکسل نظیر یک ابجکت است

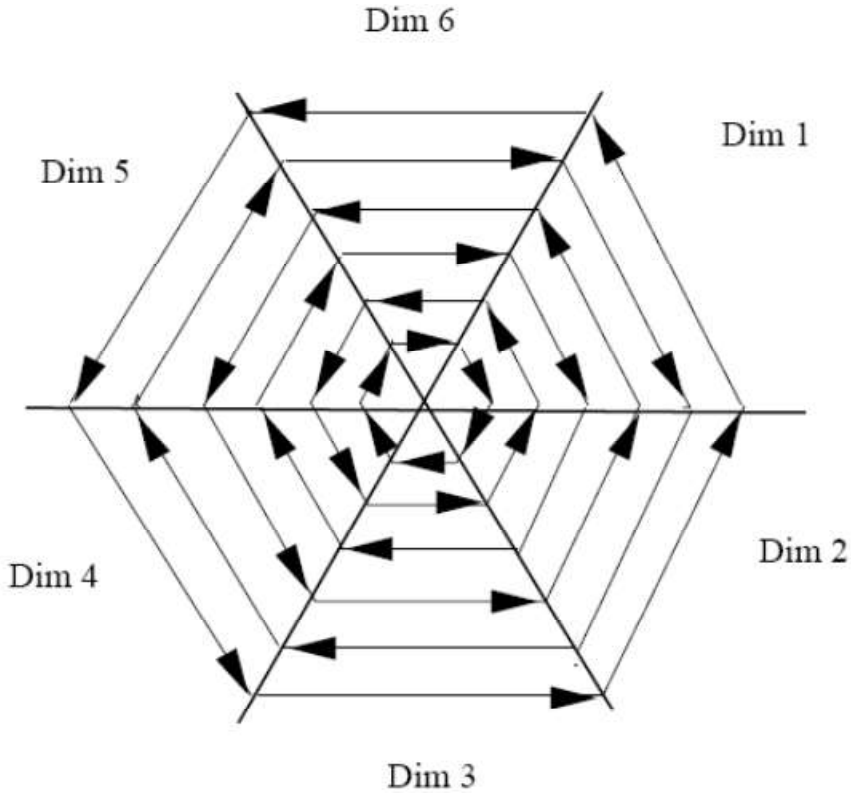
رابطه ی خاصی بین سن و سطح درامد وجود نداره

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 تجسم تبدیل های هندسی و پیش بینی داده ها طرح ها
- Methods
  - Scatterplot and scatterplot matrices
  - Landscapes
  - Parallel coordinates

طرح پراکنده
تکنیکی که یک تصویر و
دیدکلی از داده ها میده

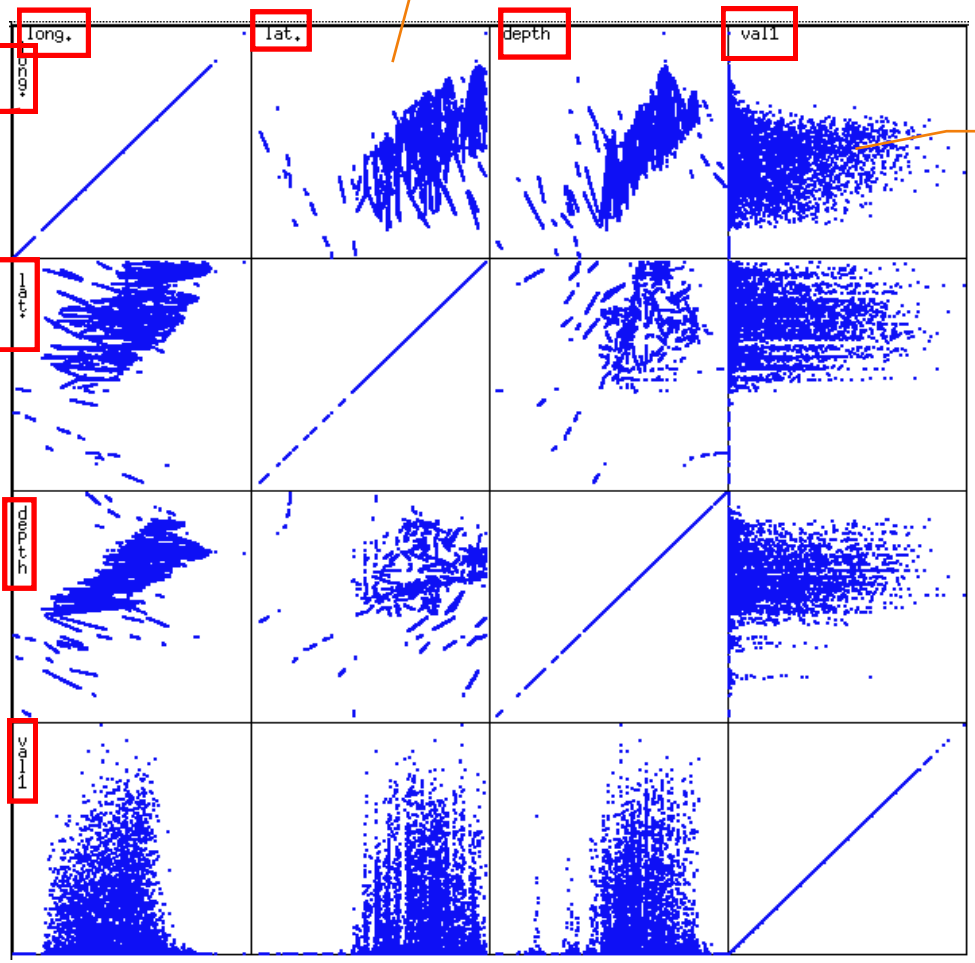
Scatterplot Matrices

رابطه ی اتریبیوت
یک با اتریبیوت دو

چهارتا اتریبیوت را درصفحه
ی ایکس و وای رسم میکنیم
میشه یه جدول مانند ۴ در ۴

اتریبیوت یک با خودش
چه رابطه ای داشته؟
به ازای هر ایکسی ازش
همون ایکس رو میگیریم
پس میشه y=x
داشتن رابطه ی یک
تریبیوت با خودش به چه
دردی میخوره؟

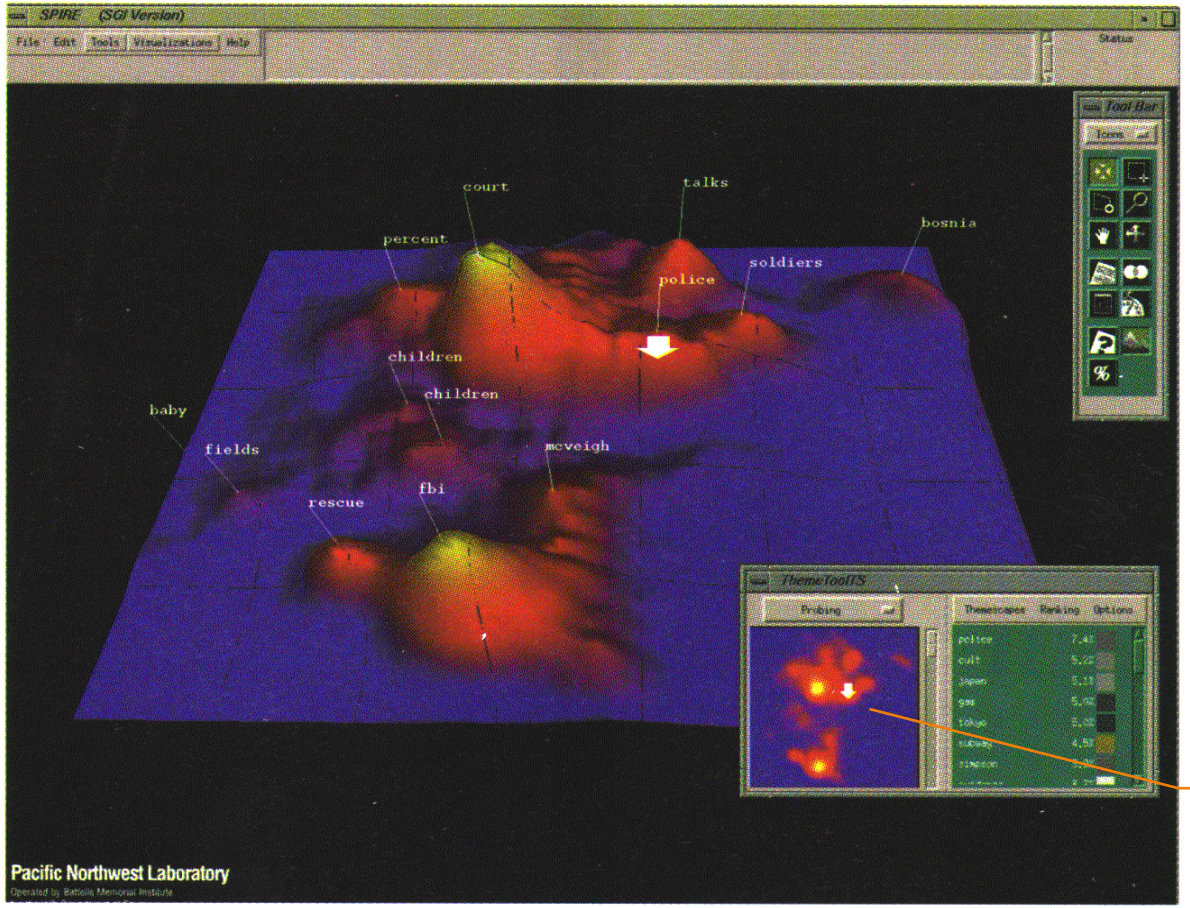
هرابجکتمون یک نقطه
ای میشه



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of (k2/2-k) scatterplots]

Landscapes

علاوه برایکس و وای بعد زد هم اضافه میشه برای نمایش اطلاعات جاهایی که با ۳تا اتریبیوت کار داریم ازاین نمودارها استفاده میشه مثلا میخایم ببینیم با مقدار ایکس از اتریبیوت یک و مقدار وای از تریبیوت ۲ ،اتریبیوت ۳ چه مقداری پیداکرده؟با سطح ارتفاع یا رنگ مقادیر اتریبیوت ها را نشان میده



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

news articles visualized as a landscape

از بالا به تصویر سه بعدی نگاه کنیم این شکل را میبینیم

- 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

همنوایی بین رفتارهای اتریبیوت ها را نشان میده

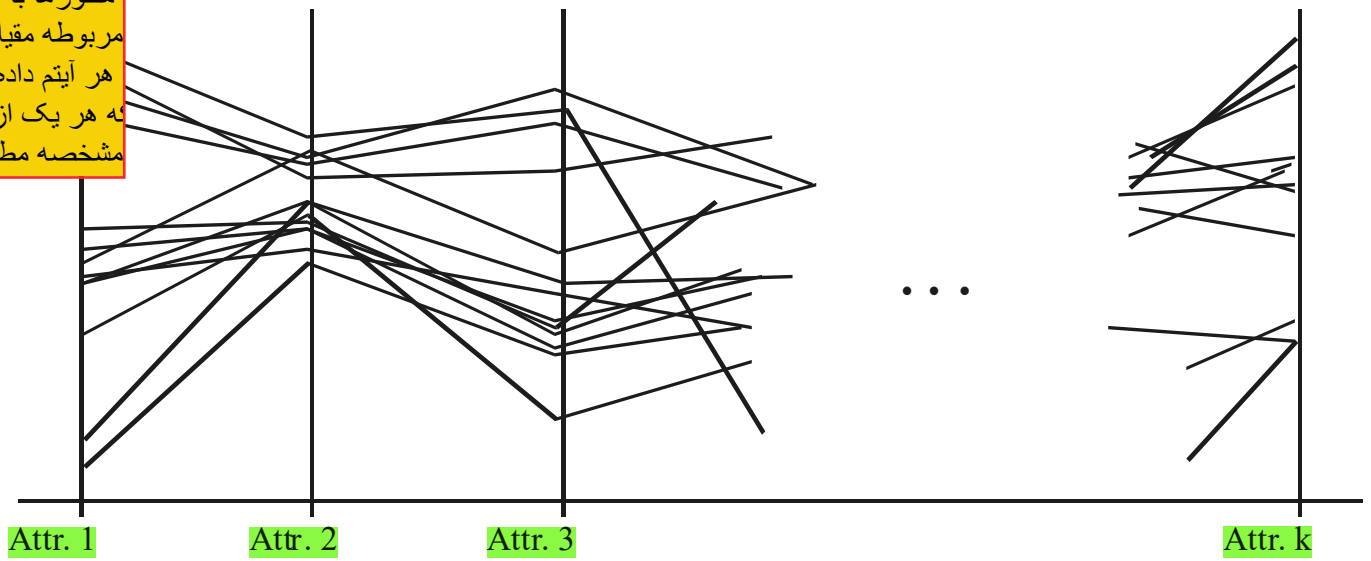
کنار هم گذاشتن اتریبیوت های مختلف ما چندین اتریبیوت داریم که هرکدوم یه مقدار مین و یه مقدار ماکس داره هر ابجکتی هم از هراتریبیوتی یک مقدار داره

Parallel Coordinates

مشاهده ی توده ی رفتارهای کلی مثلا میفهمیم اونایی که قدشون کمه و وزنشون کمه معدلشون بالاست!

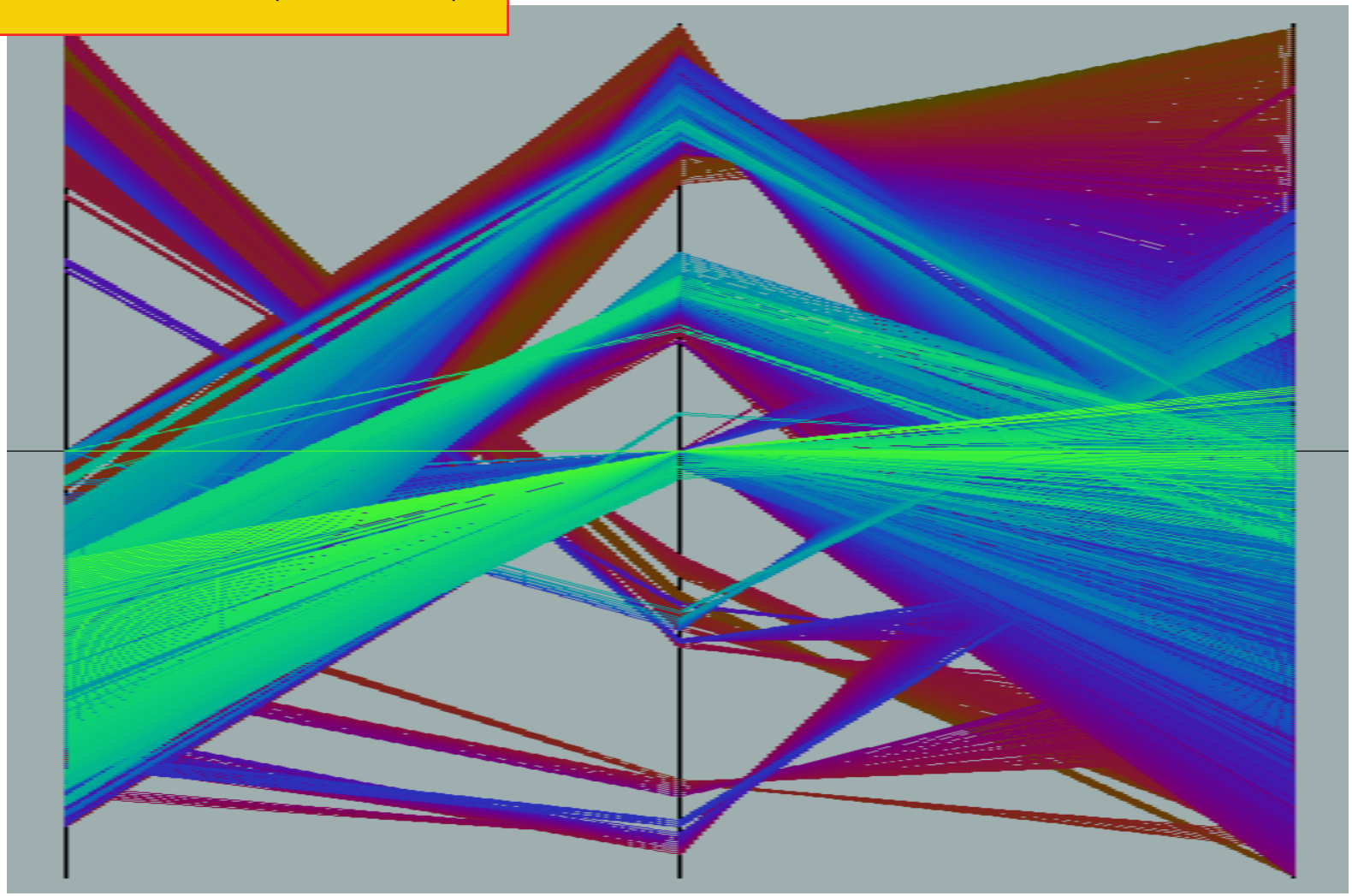
- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to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

n محور مساوی که با یکی از محورهای صفحه موازی هستند و با ویژگی ها مطابقت دارند.
محورها به [حداقل، حداکثر]: محدوده ویژگی مربوطه مقیاس می شوند.
هر آیتم داده مربوط به یک خط چند ضلعی است که هر یک از محورها را در نقطه ای که با مقدار مشخصه مطابقت دارد قطع می کند.



Parallel Coordinates of a Data Set

کاربرد: مثلا برای تقسیم کردن داده ها به دسته های مختلف مثل کلاسترینگ نمونه ها
برای پیداکردن یک سری پترن و الگو از نمونه ها

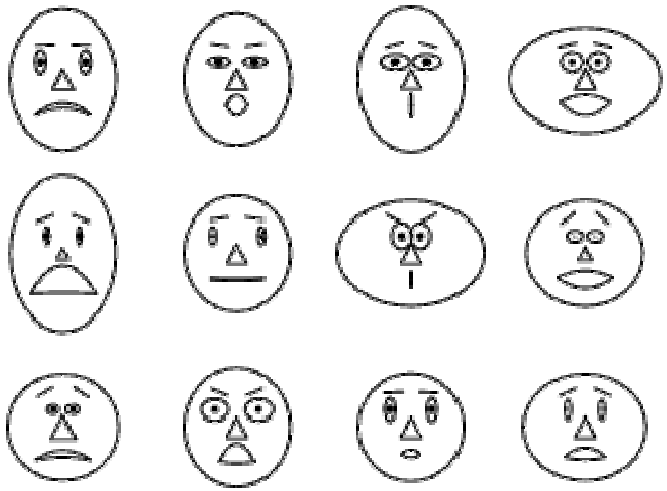


Icon-Based Visualization Techniques

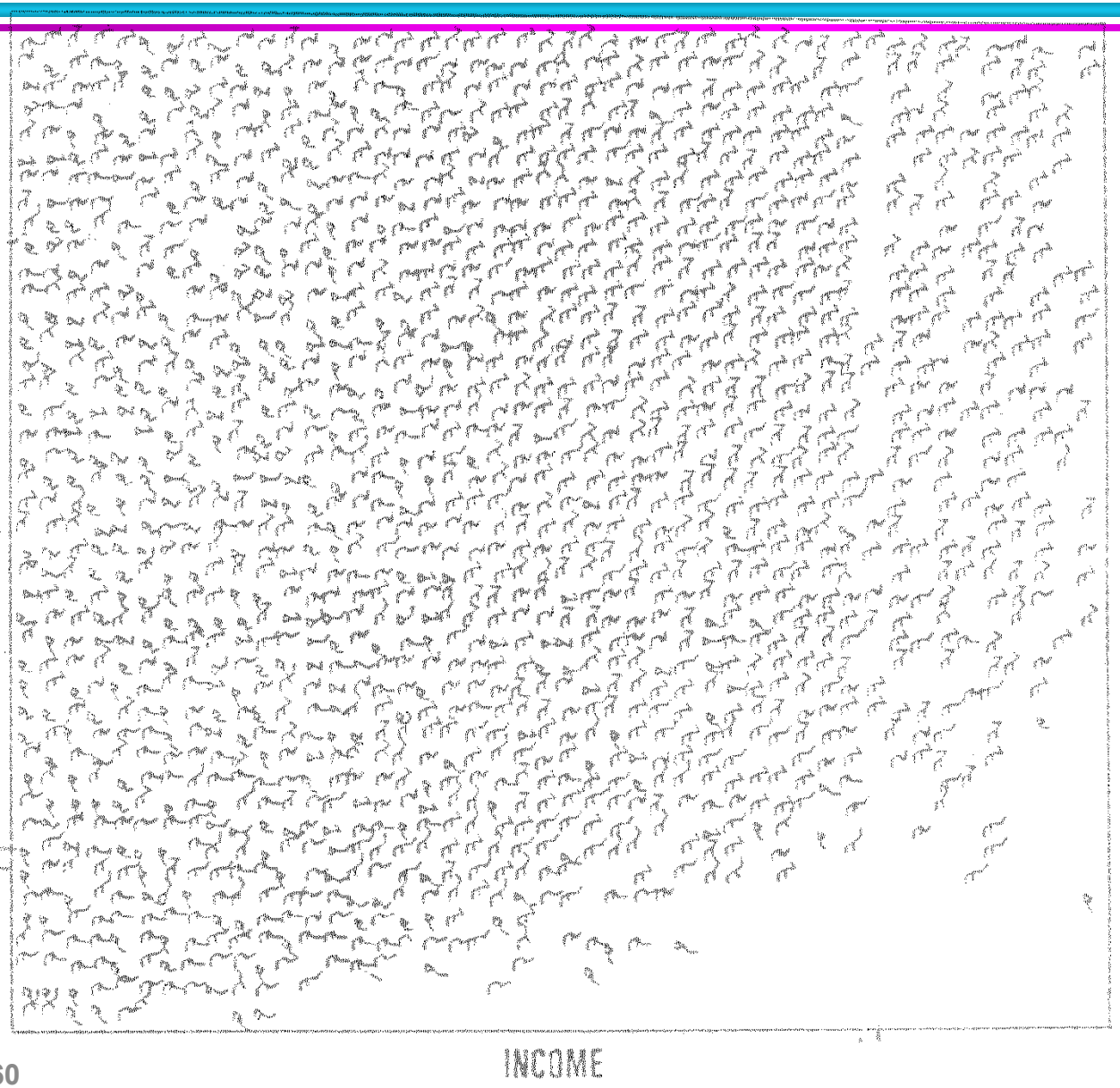
- Visualization of the data values as features of icons
- Typical visualization methods
  - Chernoff Faces
  - Stick Figures
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information
  - Tile bars: Use small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



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

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

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

SIMILARITY AND DISSIMILARITY MEASURES

ما انسان ها خودمون هم وقتی میخایم قضاوت کنیم دنبال شباهت ها میگردیم توی تاریخچه مغزمون میگیم این ابجکت شبیه به کدوم ابجکتی است که قبلا باش برخورد کردم و چه طوری برخورد کردم باش که الان با این جدیده هم همون طوری رفتار کنم

معیار های پیداکردن شباهت و تفاوت

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

پیداکردن یک توصیف عددی و کمی از شباهت دوتا ابجکت

دوتا ابجکت که شبیه هم هستند مقدار بیشتری بده اگه هم شبیه هم نیستند صفر بده

ر داده کاوی دنبال شباهت یابی هستیم یعنی دوست داریم که شباهت ها را در یک مساله کشف کنیم این طوری نسبت به داده ها بی یاز میشیم یعنی وقتی یه داده ی جدیدی اومد برحسب شباهتش به داده های قبلی دربارش تصمیم میگریم

نزدیکی یا مجاورت

مثلا اگه میخایم ببینیم یه فردی یه درسی رو پاس میکنه یا نه میبینیم نسبت به ادم قبلی ها چقدر فاصله داره اگه شبیه درس خوان کلاس باشه به احتمال زیاد هم پاس میکنه

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	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d,\ s = \frac{1}{1+d},\ s = e^{-d},$ $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

یشترین میزان فاصله بین دوتا ابجکت

چون میخایم رنج را نرمال کنیم و ببریم به بازه ی صفر و یک

رنگ چشم مثلا سبز و قرمز وابی
تیم های لیگ برتر مثلا
شباهت میشه مساوی هستند یا نه؟

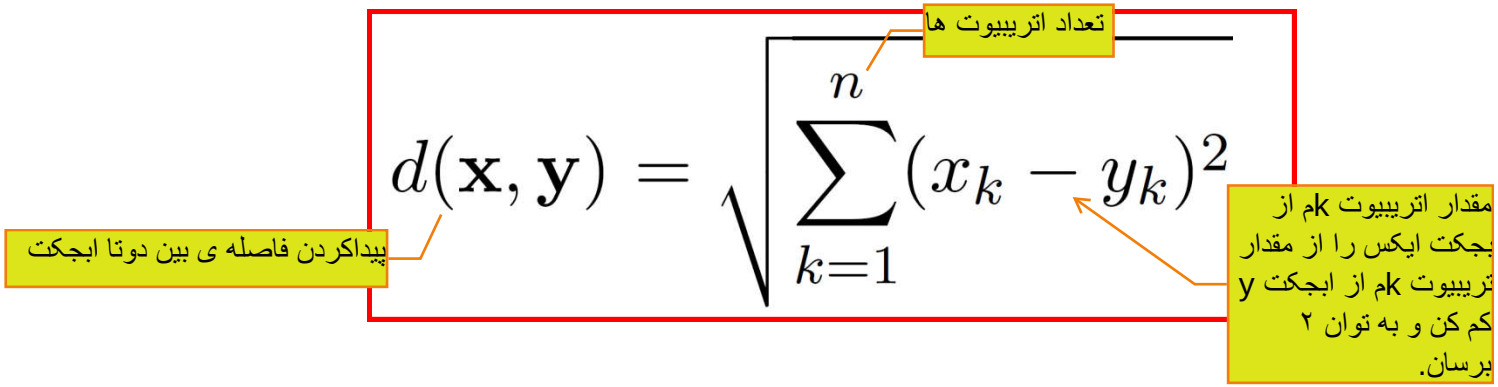
کوچک متوسط بزرگ
مثلا اندازه ی موبایل یا توپ فوتبال
کیفیت یک کالا : خوب عالی بد

برای مقایسه ی دوتا کارخانه که ماشین تولید میکنند میخاهیم کیفیتشون را مقایسه کنیم
باید یه عددی به مقدار کیفیت ها بدیم مثلاصفر یک دو
با نسبت دادن عدد، داده ها را برای مقایسه اماده میکنیم
شاید اندازه ی رنج و طیف خوب و بد بودن و عالی بودن در یک مساله متفاوت باشه
مثلا رنج عددی در خوب بودن خیلی بیشتر از رنج در حالت بد باشه

فاصله اقلیدسی

Euclidean Distance

- Euclidean Distance



where 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**.

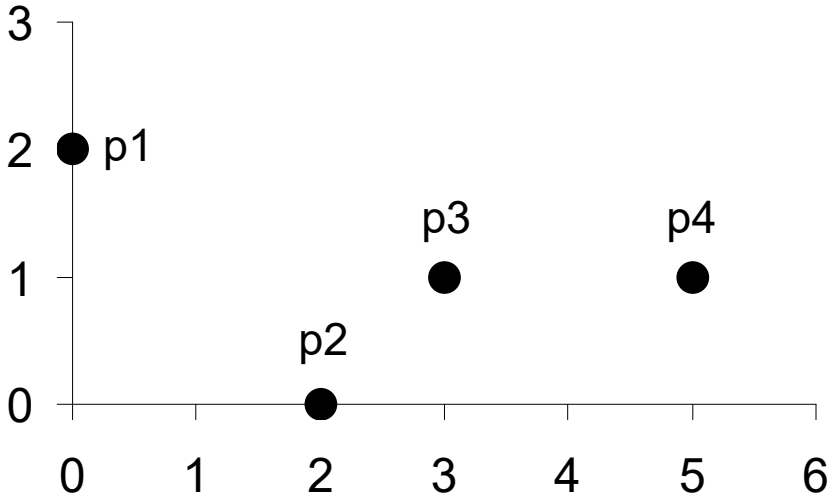
- Standardization is necessary, if scales differ.

The most popular distance measure is **Euclidean distance** (i.e., straight line or "as the crow flies"). Let $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ be two objects described by p numeric attributes. The Euclidean distance between objects i and j is defined as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}. \quad (2.16)$$

Euclidean Distance

وتر مثلث قائم الزاویه



مقادیر اتریبیوت های ایکس و وای

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

بین یه ابجکت با خودش هیچ فاصله ای نیست دیگه

Distance Matrix

تعمیم یافته ی فرمول اقلیدوسی

فاصله مینکوفسکی

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 $\boldsymbol{x}$ and $\boldsymbol{y}$.

Minkowski distance is a generalization of the Euclidean and Manhattan distances. It is defined as

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}, \qquad (2.18)$$

where h is a real number such that $h \geq 1$. (Such a distance is also called L_p **norm** in some literature, where the symbol p refers to our notation of h. We have kept p as the number of attributes to be consistent with the rest of this chapter.) It represents the Manhattan distance when $h = 1$ (i.e., L_1 norm) and Euclidean distance when $h = 2$ (i.e., L_2 norm).

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

- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors

Euclidean distance and Manhattan distance. Let $\boldsymbol{x}_1 = (1, 2)$ and $\boldsymbol{x}_2 = (3, 5)$ represent two objects as shown in Figure 2.23. The Euclidean distance between the two is $\sqrt{2^2 + 3^2} = 3.61$. The Manhattan distance between the two is $2 + 3 = 5$. ■

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

Another well-known measure is the **Manhattan (or city block) distance**, named so because it is the distance in blocks between any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks). It is defined as

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|. \quad (2.17)$$

1) Calculate the Minkowski distance of order 2 between points (1, 2, 3) and (4, 5, 6):
Minkowski distance = ((4-1)^2 + (5-2)^2 + (6-3)^2)^(1/2)
Minkowski distance = 5.196

2) Calculate the Minkowski distance of order 3 between points (0, 0, 0) and (1, 2, 3):
Minkowski distance = ((1-0)^3 + (2-0)^3 + (3-0)^3)^(1/3)
Minkowski distance = 3.301

3) Calculate the Minkowski distance of order 4 between points (-1, -2) and (3, 4):
Minkowski distance = ((3-(-1))^4 + (4-(-2))^4)^(1/4)
Minkowski distance = 6.211

4) Calculate the Minkowski distance of order 5 between points (2, -5, 1) and (-3, 7, -2):
Minkowski distance = ((-3-2)^5 + (7-(-5))^5 + (-2-1)^5)^(1/5)
Minkowski distance = 10.342

5) Calculate the Minkowski distance of order 6 between points (0, 0) and (5, -12):
Minkowski distance = ((5-0)^6 + (-12-0)^6)^(1/6)
Minkowski distance = 12.069

1) Calculate the Euclidean distance between points (1, 2) and (4, 6):
Euclidean distance = ((4-1)^2 + (6-2)^2)^(1/2)
Euclidean distance = 5

2) Calculate the Euclidean distance between points (-3, 0) and (0, 4):
Euclidean distance = ((0-(-3))^2 + (4-0)^2)^(1/2)
Euclidean distance = 5

3) Calculate the Euclidean distance between points (2, -5, 1) and (-3, 7, -2):
Euclidean distance = ((-3-2)^2 + (7-(-5))^2 + (-2-1)^2)^(1/2)
Euclidean distance = 13

4) Calculate the Euclidean distance between points (0, 0, 0) and (1, 2, 3):
Euclidean distance = ((1-0)^2 + (2-0)^2 + (3-0)^2)^(1/2)
Euclidean distance = 3.742

5) Calculate the Euclidean distance between points (-1, -2, -3) and (4, 5, 6):
Euclidean distance = ((4-(-1))^2 + (5-(-2))^2 + (6-(-3))^2)^(1/2)
Euclidean distance = 11.225

Minkowski Distance

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

جمع تفاوت ایکس ها و وای ها باهمدیگه
۲+۲ = ۴

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

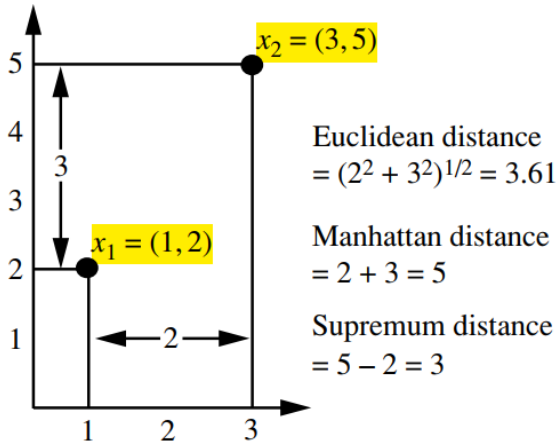
L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

The **supremum distance** (also referred to as L_{max}, L_∞ **norm** and as the **Chebyshev distance**) is a generalization of the Minkowski distance for $h \to \infty$. To compute it, we find the attribute f that gives the maximum difference in values between the two objects. This difference is the supremum distance, defined more formally as:

$$d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|. \tag{2.19}$$

The L^∞ norm is also known as the *uniform norm*.



Supremum distance. Let's use the same two objects, $\boldsymbol{x_1} = (1, 2)$ and $\boldsymbol{x_2} = (3, 5)$, as in Figure 2.23. The second attribute gives the greatest difference between values for the objects, which is $5 - 2 = 3$. This is the supremum distance between both objects. ■

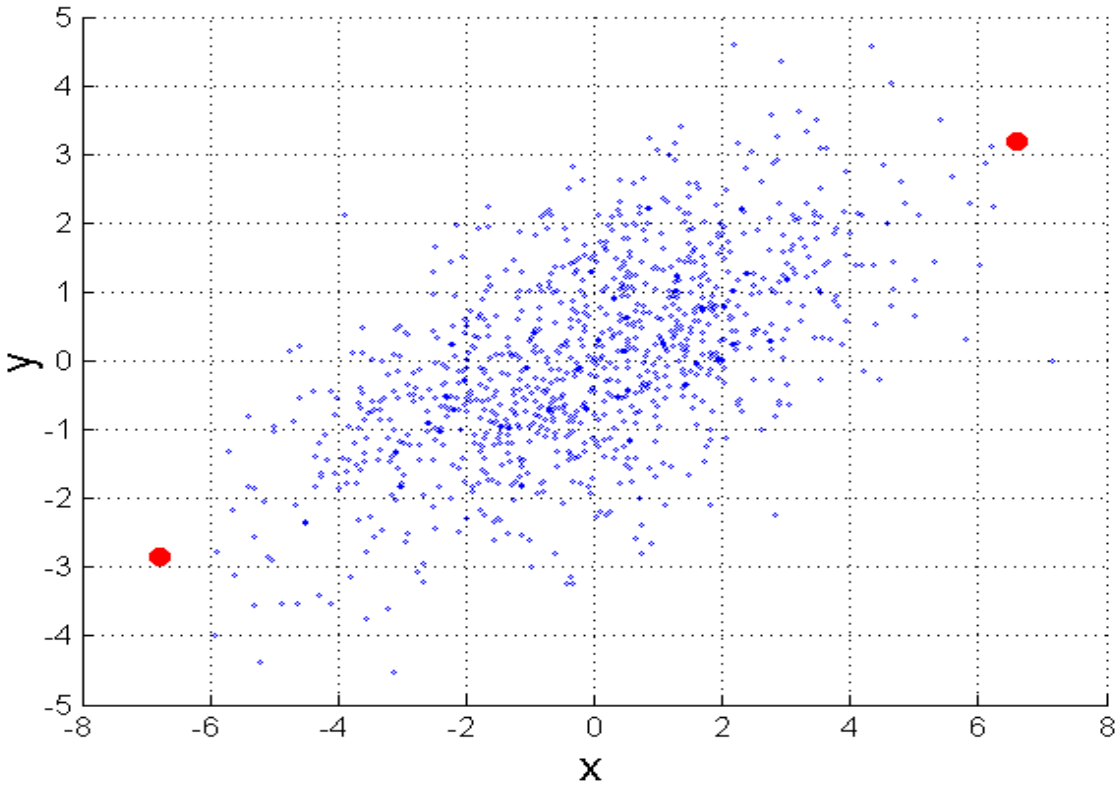
میخایم به بعدهای ایکس و وای به شکل وزن دار نگاه کنیم

Mahalanobis Distance

Mahalanobis distance is a measure of the distance between a point and a distribution. It takes into account the covariance between variables, which makes it useful for multivariate data analysis. The formula for Mahalanobis distance is:
$D^2 = (x - \mu)' \Sigma^{-1} (x - \mu)$
where D is the Mahalanobis distance, x is the vector of observations, μ is the vector of means, and Σ^{-1} is the inverse covariance matrix.

The Mahalanobis distance is calculated for each point, and those that are far away from the center of the distribution (i.e., have a high Mahalanobis distance) are considered outliers.

نسبت به کواریانس میایم فاصله سنجی میکنیم



Σ is the covariance matrix

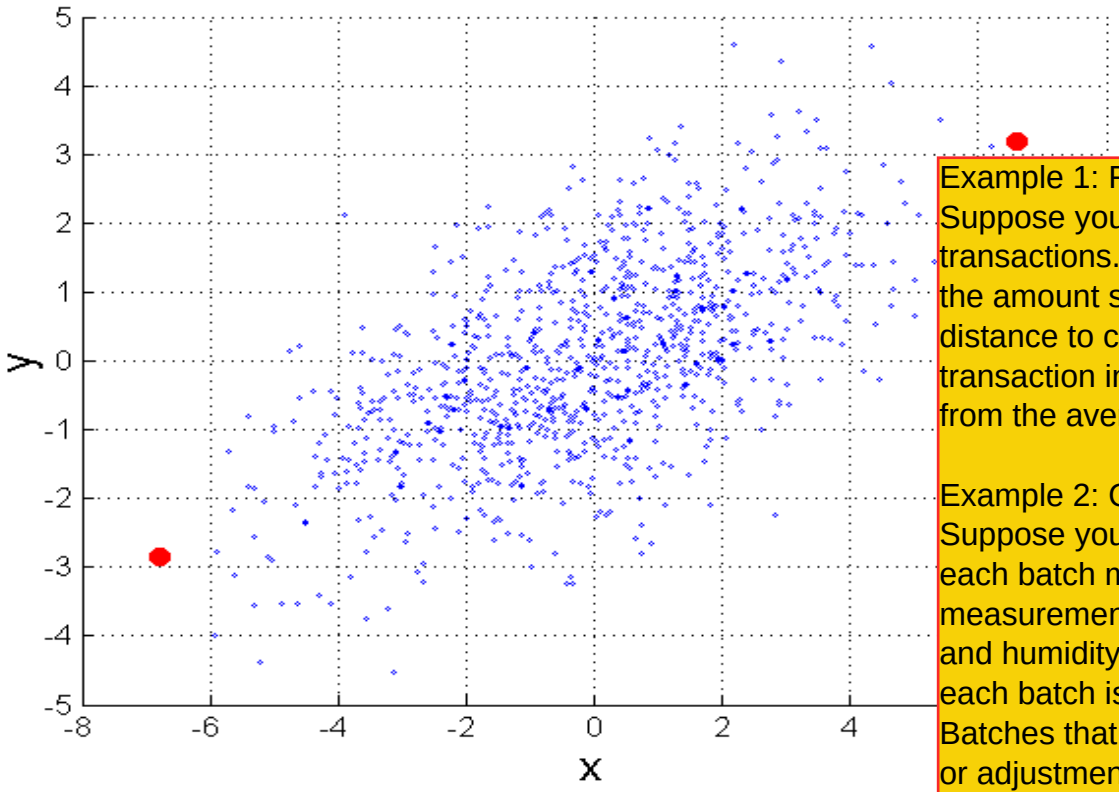
$$\begin{bmatrix} \mathrm{Var}(x_1) & \cdots & \mathrm{Cov}(x_n, x_1) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_n, x_1) & \cdots & \mathrm{Var}(x_n) \end{bmatrix}$$

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

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

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

Example 1: Fraud detection
Suppose you are working for a bank and you want to detect fraudulent transactions. You have data on customers' transaction history, such as the amount spent, location, and time of day. You can use Mahalanobis distance to calculate how far each transaction is from the average transaction in terms of these variables. Transactions that are far away from the average may be flagged as potentially fraudulent.

Example 2: Quality control
Suppose you are manufacturing a product and you want to ensure that each batch meets certain quality standards. You have data on various measurements taken during production, such as temperature, pressure, and humidity. You can use Mahalanobis distance to calculate how far each batch is from the average batch in terms of these variables. Batches that are far away from the average may need further inspection or adjustment.

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

اگه پراکندگی داده ها یکسان باشه ینی یه دایره تشکیل بده فاصله ی اقلیدسی و مهلنوبیس یکی میشه

Suppose we have a dataset with two variables, x and y, and three observations:

Observation 1: x = 2, y = 4
Observation 2: x = 3, y = 5
Observation 3: x = 4, y = 6

We want to calculate the Mahalanobis distance between observation 1 and observation 2. To do this, we first need to calculate the covariance matrix of the dataset:

Covariance matrix:

	x	y
x	1/2	1/2
y	1/2	1/2

Next, we need to calculate the inverse of the covariance matrix:

Inverse covariance matrix:

	x	y
x'	2	-2
y'	-2	2

Now we can calculate the Mahalanobis distance using the formula:

$D^2 = (x1 - x2)' * S^{-1} * (x1 - x2)$

where D is the Mahalanobis distance, x1 is the vector of variables for observation 1 (x=2, y=4), x2 is the vector of variables for observation 2 (x=3, y=5), and S^{-1} is the inverse covariance matrix.

Plugging in our values:

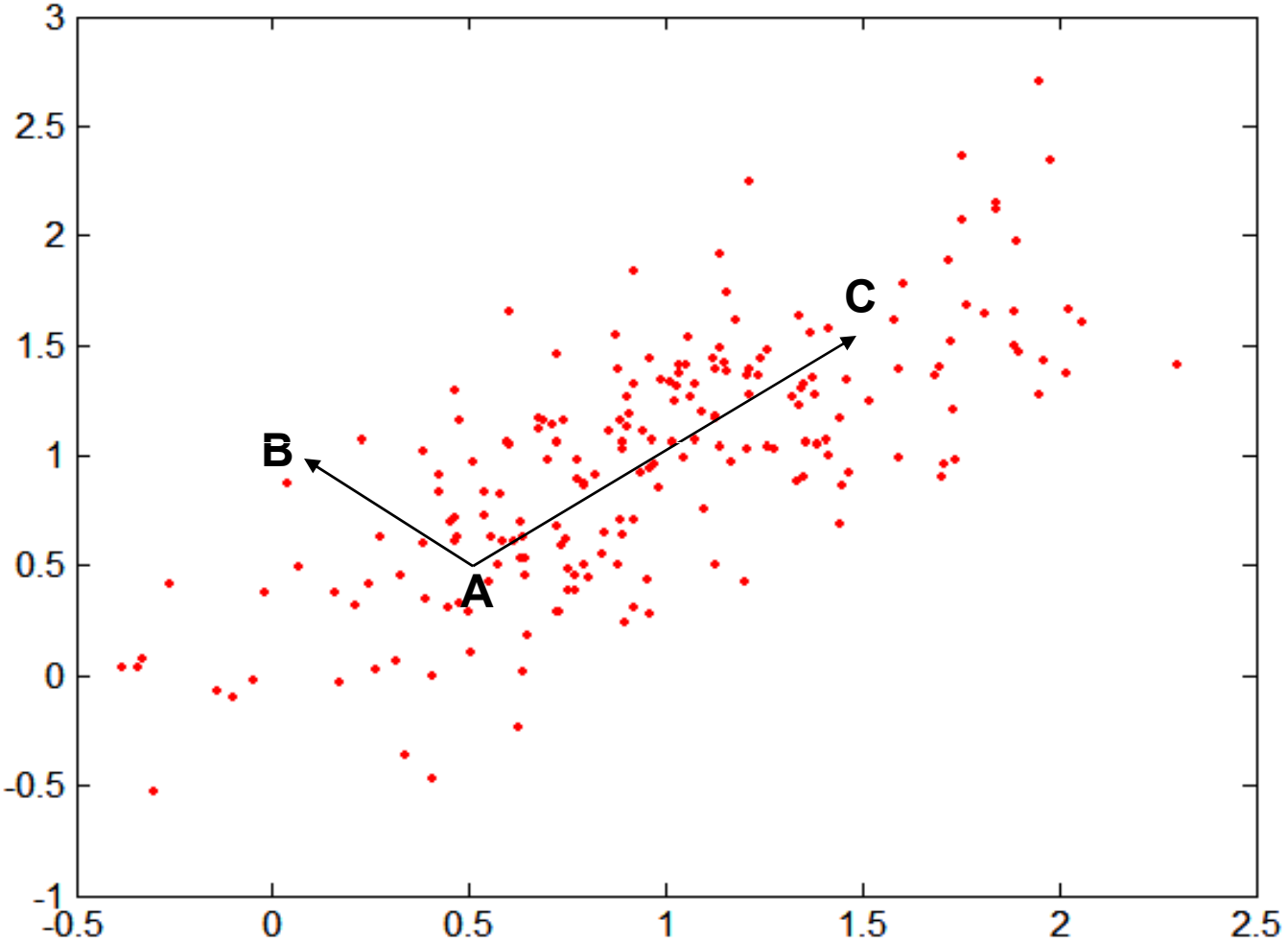
$D^2 = (([2,4] - [3,5])' * [[2,-2],[-2,2]] * ([2,4] - [3,5]))$
$= (([-1,-1])' * [[-4],[-4]] * ([-1,-1]))$
$= (-8)$

Taking the square root of this value gives us our final result:

$D = \sqrt{-8} = 2.83$

Therefore, the Mahalanobis distance between observation 1 and observation 2 is 2.83.

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)

Mahal(A,B) = 5

Mahal(A,C) = 4

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all *x* and *y* and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)
3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (Triangle Inequality)

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

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$.
   (does not always hold, e.g., cosine)
2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

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

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

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

SMC versus Jaccard: Example

$\mathbf{x}$ = 1 0 0 0 0 0 0 0 0 0

$\mathbf{y}$ = 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)

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

$= (0+7) / (2+1+0+7) = 0.7$

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

Cosine Similarity

دوتا وکتور یا بردار یا ابجکت داریم

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

ضرب داخلی

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

ضرب اندازه ی وکتورها برای نرمال کردن صورت که بدست اومد

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:

$$\mathbf{d}_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$

$$\mathbf{d}_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

$$\langle \mathbf{d}_1, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

محاسبه ی نرم ۲ این وکتورها

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

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

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

چه زمانی بیشترین شباهت بدست میاد؟ وقتی که وکتورها دقیقا مثل هم باشند که جواب cos(d1,d2)=1 میشه

دوتا ابجکت دی ۱ ودی ۲ چقدر دارن شبیه هم رفتار میکنن؟

1) Calculate the cosine similarity between vectors [2, 3, 4] and [5, 6, 7]:
cosine similarity = (2*5 + 3*6 + 4*7) / (sqrt(2^2 + 3^2 + 4^2) * sqrt(5^2 + 6^2 + 7^2))
cosine similarity = 0.994

2) Calculate the cosine similarity between vectors [1, 0, -1] and [-1, 0, 1]:
cosine similarity = (1*(-1) + 0*0 + (-1)*1) / (sqrt(1^2 + 0^2 + (-1)^2) * sqrt((-1)^2 + 0^2 + 1^2))
cosine similarity = -1

3) Calculate the cosine similarity between vectors [4, 5, 6] and [7, 8, 9]:
cosine similarity = (4*7 + 5*8 + 6*9) / (sqrt(4^2 + 5^2 + 6^2) * sqrt(7^2 + 8^2 + 9^2))
cosine similarity = 0.997

4) Calculate the cosine similarity between vectors [0, 1, 0] and [0, -1, 0]:
cosine similarity = (0*0 + 1*(-1) + 0*0) / (sqrt(0^2 + 1^2 + 0^2) * sqrt(0^2 + (-1)^2 + 0^2))
cosine similarity = -1

5) Calculate the cosine similarity between vectors [3, -4] and [-6, 8]:
cosine similarity = (3*(-6) + (-4)*8) / (sqrt(3^2 + (-4)^2) * sqrt((-6)^2 + 8^2))
cosine similarity = -1

Suppose we have the following two-dimensional data set:

	A_1	A_2
$\boldsymbol{x_1}$	1.5	1.7
$\boldsymbol{x_2}$	2	1.9
$\boldsymbol{x_3}$	1.6	1.8
$\boldsymbol{x_4}$	1.2	1.5
$\boldsymbol{x_5}$	1.5	1.0

(a) Consider the data as two-dimensional data points. Given a new data point, $\boldsymbol{x} = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using (1) Euclidean distance (Equation 7.5), and (2) cosine similarity (Equation 7.16).

The Euclidean distance of two n-dimensional vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, is defined as: $\sqrt[2]{\sum_{i=1}^{n}(x_i - y_i)^2}$. The cosine similarity of $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as: $\frac{\boldsymbol{x}^t \cdot \boldsymbol{y}}{||\boldsymbol{x}||||\boldsymbol{y}||}$, where $\boldsymbol{x}^t$ is a transposition of vector $\boldsymbol{x}$, $||\boldsymbol{x}||$ is the Euclidean norm of vector $\boldsymbol{x}$,[1] and $||\boldsymbol{y}||$ is the Euclidean norm of vector $\boldsymbol{y}$. Using these definitions we obtain the distance from each point to the query point.

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$
Euclidean distance	0.14	0.67	0.28	0.22	0.61
Cosine similarity	0.9999	0.9957	0.9999	0.9990	0.9653

Based on the Euclidean distance, the ranked order is $\boldsymbol{x_1}, \boldsymbol{x_4}, \boldsymbol{x_3}, \boldsymbol{x_5}, \boldsymbol{x_2}$. Based on the cosine similarity, the order is $\boldsymbol{x_1}, \boldsymbol{x_3}, \boldsymbol{x_4}, \boldsymbol{x_2}, \boldsymbol{x_5}$.

[1] The Euclidean normal of vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is defined as $\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$.

Cosine similarity between two term-frequency vectors. Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are the first two term-frequency vectors in Table 2.5. That is, $\boldsymbol{x} = (5,0,3,0,2,0,0,2,0,0)$ and $\boldsymbol{y} = (3,0,2,0,1,1,0,1,0,1)$. How similar are $\boldsymbol{x}$ and $\boldsymbol{y}$? Using Eq. (2.23) to compute the cosine similarity between the two vectors, we get:

$$\boldsymbol{x}^t \cdot \boldsymbol{y} = 5 \times 3 + 0 \times 0 + 3 \times 2 + 0 \times 0 + 2 \times 1 + 0 \times 1 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 0 \times 1 = 25$$

$$||\boldsymbol{x}|| = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$||\boldsymbol{y}|| = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = 4.12$$

$$sim(\boldsymbol{x}, \boldsymbol{y}) = 0.94$$

Therefore, if we were using the cosine similarity measure to compare these documents, they would be considered quite similar. ■

برای متغیر های پیوسته

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

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

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

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

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

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

The correlation coefficient is a measure of the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.

Let's consider an example where we have two variables, X and Y, with the following values:

X = [1, 2, 3, 4, 5]
Y = [3, 5, 7, 9, 11]

To calculate the correlation coefficient between X and Y, we first need to calculate the mean and standard deviation of each variable.

The mean of X is:

mean(X) = (1 + 2 + 3 + 4 + 5) / 5 = 3

The mean of Y is:

mean(Y) = (3 + 5 + 7 + 9 + 11) / 5 = 7

The standard deviation of X is:

std(X) = sqrt(((1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2) / (5-1)) = 1.5811

The standard deviation of Y is:

std(Y) = sqrt(((3-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (11-7)^2) / (5-1)) = 2.8284

Next, we can calculate the covariance between X and Y:

cov(X,Y) = ((1-3)*(3-7) + (2-3)*(5-7) + (3-3)*(7-7) + (4-3)*(9-7) + (5-3)*(11-7)) / (5-1) = 5

Finally, we can calculate the correlation coefficient between X and Y:

corr(X,Y) = cov(X,Y) / (std(X) * std(Y)) = 5 / (1.5811 * 2.8284) = 0.8839

Therefore, the correlation coefficient between X and Y is 0.8839, indicating a strong positive linear relationship between the two variables.

Suppose we have two variables, X and Y, with the following values:

X = [10, 20, 30, 40, 50]
Y = [5, 15, 25, 35, 45]

The steps to calculate the correlation coefficient are similar to the previous example. First, we calculate the mean and standard deviation of each variable:

mean(X) = (10 + 20 + 30 + 40 + 50) / 5 = 30
std(X) = sqrt(((10-30)^2 + (20-30)^2 + (30-30)^2 + (40-30)^2 + (50-30)^2) / (5-1)) = 15.8114

mean(Y) = (5 + 15 + 25 + 35 + 45) / 5 = 25
std(Y) = sqrt(((5-25)^2 + (15-25)^2 + (25-25)^2 + (35-25)^2 + (45-25)^2) / (5-1)) = 15.8114

Next, we calculate the covariance between X and Y:

cov(X,Y) = ((10-30)*(5-25) + (20-30)*(15-25) + (30-30)*(25-25) + (40-30)*(35-25) + (50-30)*(45-25)) / (5-1) = 500

Finally, we calculate the correlation coefficient between X and Y:

corr(X,Y) = cov(X,Y) / (std(X) * std(Y)) = 500 / (15.8114 * 15.8114) = 0.9439

Therefore, the correlation coefficient between X and Y is 0.9439, indicating a strong positive linear relationship between the two variables.

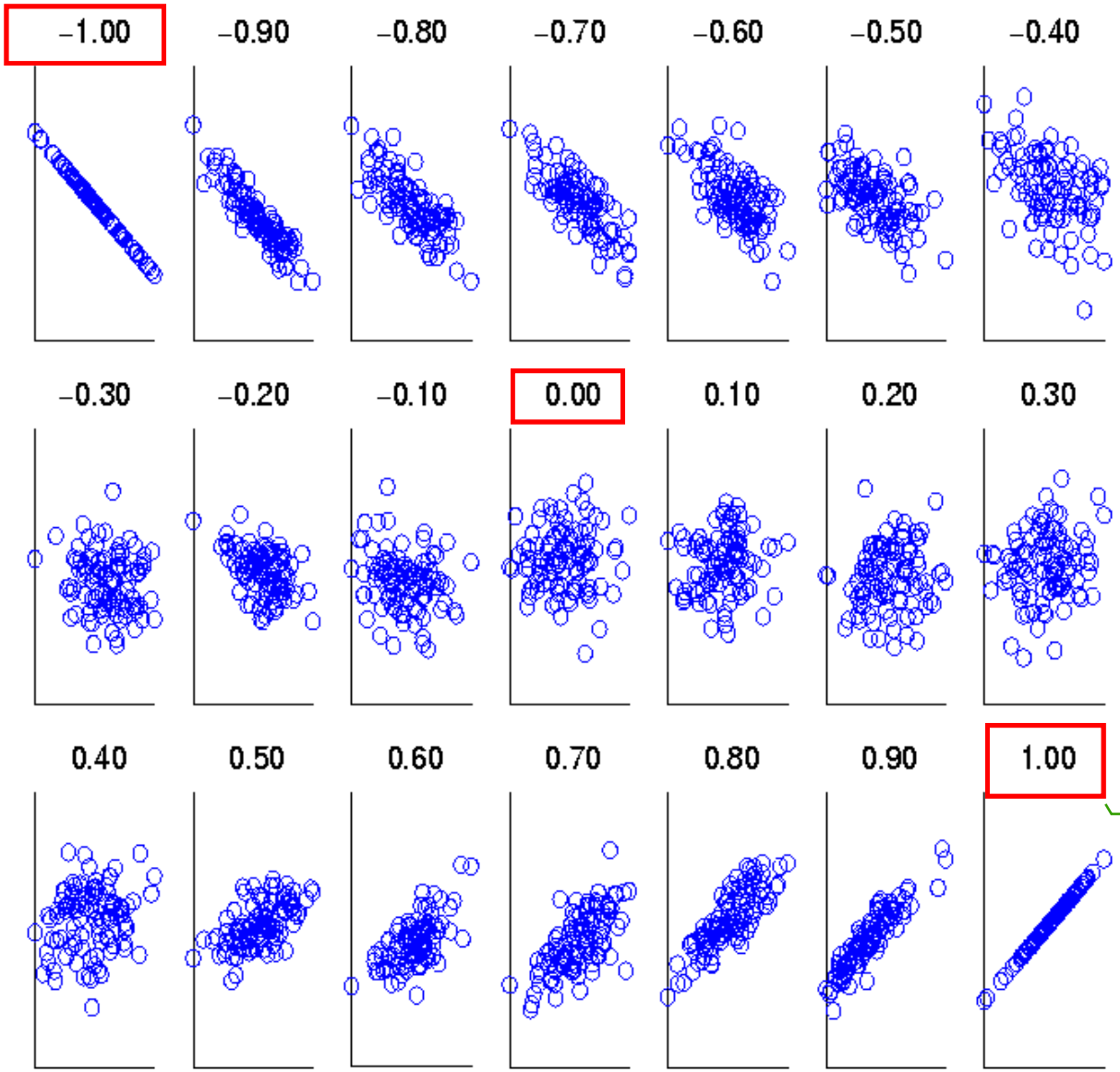
Drawbacks of correlation:
1. Correlation measures only linear relationships: Correlation measures only the linear relationship between two variables and does not capture nonlinear relationships. In machine learning, nonlinear relationships are often present in the data, and using correlation as a measure of distance can lead to inaccurate results.

2. Correlation is sensitive to outliers: Correlation is sensitive to outliers, which can have a significant impact on the correlation coefficient. In machine learning, outliers are common in real-world datasets, and using correlation as a measure of distance can lead to inaccurate results.

3. Correlation does not account for differences in scale: Correlation does not account for differences in the scale of the variables being compared. In machine learning, variables often have different scales, and using correlation as a measure of distance can lead to inaccurate results.

4. Correlation does not capture complex relationships: Correlation measures only the linear relationship between two variables and does not capture complex relationships, such as interactions between variables. In machine learning, complex relationships between variables are often present in the data, and using correlation as a measure of distance can lead to inaccurate results.

Visually Evaluating Correlation



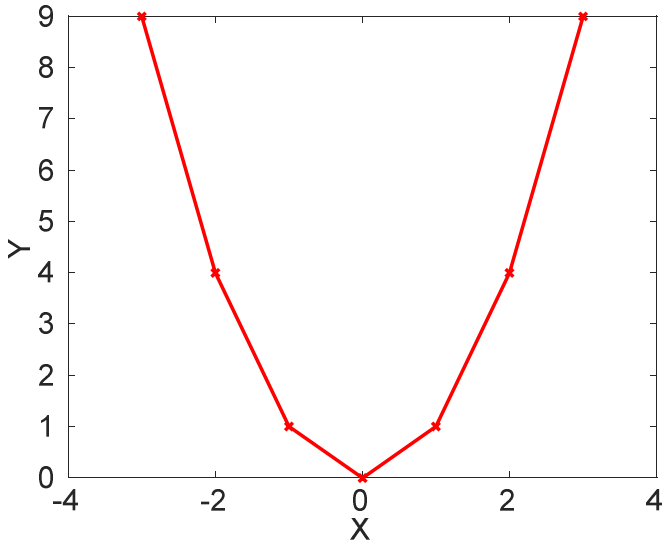
مقایسه ی خود اتریبیوت با خودش

Scatter plots showing the similarity from –1 to 1.

Drawback of Correlation

نمیتواند روابط غیر خطی را تشخیص بده
رابطه ای که بین اتریبیوت ها هست غیر خطی است
خطی یعنی اگه ایکس را دو برابر کردیم وای متناظرش هم دوبرابر بشه

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



$$y_i = x_i^2$$

- mean(**x**) = 0, mean(**y**) = 4
- std(**x**) = 2.16, std(**y**) = 3.74
- corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)

= 0

ایکس و وای باهم ارتباط دارند ولی داره میگه ارتباطشون صفر است! یعنی نتوانست ارتباط اینهارا تشخیص بده

xi - mean(x)
-3 - 0 = -3

yi - mean(y)
9 - 5 = 4

n=7
n-1 = 6

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

آیا رفتارشون با تغییر اتریبیوت ها تغییر میکنه یا نه؟

ثابت بودن جواب در اثر اسکیل کردن داده ها ینی در یه عددی ضرب کنیم مثلا

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

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

ثابت بودن در اثر جمع داده ها با یه مقدار ثابت

Measure	(*x* , *y*)	(*x* , y_s)	(*x* , y_t)
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

درهرشرایطی ثابت است و تغییر نمیکنه

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

اندازه وبزرگی

  - 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.

وقتی میخایم یه مدل بسازیم برامون مهم میشه از کدوم اتریبیوت ها باید استفاده کنیم کدوم هارا انتخاب کنیم و کدوم ها را کنار بگذاریم؟ مثلا ممکنه هزارتا اتریبیوت از یه ابجکت داشته باشیم ولی نمیدونیم کدومش را انتخاب کنیم؟ همش را نمیشه استفاده کرد
اگه همش را بخایم استفاده کنیم اصلا مدل سازی نخواهیم داشت

- Correlation: Correlation measures the linear relationship between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. Correlation is commonly used in statistical analysis to measure the strength and direction of the relationship between two variables.

- Cosine similarity: Cosine similarity measures the cosine of the angle between two vectors. It ranges from -1 to 1, where -1 indicates that the two vectors are diametrically opposed, 0 indicates that the two vectors are orthogonal, and 1 indicates that the two vectors are identical. Cosine similarity is commonly used in information retrieval and text processing to measure the similarity between two documents or two vectors of word frequencies.

- Euclidean distance: Euclidean distance measures the distance between two points in n-dimensional space. It is calculated as the square root of the sum of the squared differences between the corresponding coordinates of the two points. Euclidean distance is commonly used in machine learning and data mining to measure the similarity between two data points or to cluster data points based on their similarity.

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

ابزارهای اندازه گیری براساس تئوری اطلاعات
باید ببینیم چقدر عدم قطعیت توی اطلاعاتش هست؟ چقدر اطلاعاتش پراکنده است؟ احتمال رخدادن داده ها از نظر قد دانشجویان چقدره؟

Information Based Measures

- Information theory is a well-developed and fundamental disciple with broad applications

پس یه جایی باید اتریبیوت هامون را باهم مقایسه کنیم
اگه بخایم دوتا اتریبیوتی که هیچ ربطی به هم ندارن را مقایسه کنیم چطوری باید اینکارو کنیم؟

- 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

Information and Probability

اگه یه سکه بندازیم که همش رو بیاد هیچ اطلاعاتی به ما نمیده چون هیچ عدم قطعیتی توش نیست

- 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 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

پدیده هایی که احتمال رخدادشون کمتر باشه اطلاعاتی که میدن بیشتر است

اتریبیوت ها را در یک فضای مشترک مثل فضای اطلاعات میبریم وبعد مقایسشون میکنیم چون مستقیم نمیتونیم مقایسشون کنیم

information scale: فضای اطلاعاتی

اطلاعات یک رابطه ی نزدیک با احتمال و عدم قطعیت داره پدیده ای که همیشه قطعی است و عدم قطعیت نداره هیچ اطلاعاتی به ما نمیده اگه سکه طوری باشه که ما نتوانیم پیشبینی کنیم که رو میاد یا پشت در دفعه ی بعدی میگیم یه اطلاعاتی توش هست ولی اگه قرار باشه سکه همش رو بیاد اطلاعاتی نمیده نتیجه حاصل از انداختن سکه

Entropy

ارتباط و شباهت بین اتریبیوت ها را میسنجیم مثل ارتباط بین وزن و قد

مثلا تاس شش تا مقدار داره و هرکدوم یه احتمالی داره که بیاد

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

باید احتمال متغیرتصادفی ها را دربیاریم
ینی به ازای هرمقداری که متغیرتصادفیمون پیدا میکنه باید احتمالش را حساب کنیم
اگه متغیر تصادفیمون انداختن یک تاس باشه پس ۶ تا حالت داره متغیرمون پس شش تا احتمال باید بدست بیاریم

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

چقدر اطلاعات توی متغیر تصادفی مون هست؟

اگه احتمال همه مقادیر یکسان باشه ینی هرکدام 1/n باشد
انتروپی حاصل میشه یک
انتروپی:چندتا بیت لازم داریم تا اطلاعات را ذخیره کنیم؟

- 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

اگه انتروپی صفر بشه ینی احتمال یک مقدار از متغیر تصادفی یک است و بقیه صفر است ینی اطلاعاتی بمون نمیده

ماکس مقدار یه متغیر را اگه ازش لگاریتم در مبنای دو بگیریم تعداد بیت لازم برای ذخیره کردنش درمیاد

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 $q = 1$, $H = 0$

احتمال رخدادن مقادیر مختلف سکه یکسان بود و هرکدام یک دوم بود پس انتروپی میشه یک

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

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

سوال :
انتروپی رنگ
مو چیست؟

Maximum entropy is $\log_2 5 = 2.3219$

انتروپی حد پایین داره ولی حد
بالاش معلوم نیست
ولی راجع به پراکندگی اون
متغیرتصادفی یه سری اطلاعات
بمون میده

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

Suppose we have a dataset with 10 samples, where each sample belongs to one of two classes, A or B. The number of samples in each class is as follows:

Class A: 6 samples
Class B: 4 samples

To calculate the entropy of this dataset, we first need to calculate the proportion of samples in each class:

Proportion of class A = 6 / 10 = 0.6
Proportion of class B = 4 / 10 = 0.4

Next, we can use the entropy formula to calculate the entropy of the dataset:

Entropy = - (Proportion of class A * log2(Proportion of class A) + Proportion of class B * log2(Proportion of class B))

Entropy = - (0.6 * log2(0.6) + 0.4 * log2(0.4))

Entropy = 0.971

Therefore, the entropy of the dataset is 0.971.

Suppose we have a dataset with 20 samples, where each sample belongs to one of three classes, A, B, or C. The number of samples in each class is as follows:

Class A: 8 samples
Class B: 6 samples
Class C: 6 samples

To calculate the entropy of this dataset, we first need to calculate the proportion of samples in each class:

Proportion of class A = 8 / 20 = 0.4
Proportion of class B = 6 / 20 = 0.3
Proportion of class C = 6 / 20 = 0.3

Next, we can use the entropy formula to calculate the entropy of the dataset:

Entropy = - (Proportion of class A * log2(Proportion of class A) + Proportion of class B * log2(Proportion of class B) + Proportion of class C * log2(Proportion of class C))

Entropy = - (0.4 * log2(0.4) + 0.3 * log2(0.3) + 0.3 * log2(0.3))

Entropy = 1.576

Mutual Information

راجع به ارتباط دو تا متغیر اطلاعات کسب کنیم با رویکردهای تئوری اطلاعاتی

- 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,

عددی که بدست میاد میگه چقدر این دوتا متغیر به هم ربط دارند

مثلا قد و وزن را داریم

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

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 Example

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

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = E_{P_{XY}} \log \frac{P_{XY}}{P_X P_Y} .$$

Mutual information is a measure of the amount of information that two variables share. It is often used in machine learning and data analysis to determine the relationship between two variables.

Here are a few numeric examples:

1. Suppose we have two binary variables, X and Y, where X can take on the values 0 or 1 and Y can also take on the values 0 or 1. If we observe that X and Y are perfectly correlated (i.e., whenever X=1, Y=1 and whenever X=0, Y=0), then the mutual information between X and Y would be 1 bit.

2. Let's say we have two continuous variables, A and B, where A represents a person's age (in years) and B represents their income (in thousands of dollars). If we find that there is a strong positive correlation between A and B (i.e., as age increases, so does income), then the mutual information between A and B would be relatively high.

3. Consider two categorical variables C and D, where C represents a person's gender (either male or female) and D represents their favorite color (red, blue, or green). If we observe that males tend to prefer blue while females tend to prefer red, then the mutual information between C and D would be non-zero but relatively low.

Suppose we have two binary variables X and Y with the following data:

	Y=0	Y=1
X=0	10	30
X=1	20	40

To calculate the mutual information between X and Y, we can use the same formula as in Example 1:

$I(X;Y) = \sum\sum p(x,y) \log_2(p(x,y) / (p(x) * p(y)))$

First, we need to calculate the marginal probabilities:

p(X=0) = 10 + 30 = 40 / 100 = 0.4
p(X=1) = 20 + 40 = 60 / 100 = 0.6
p(Y=0) = 10 + 20 = 30 / 100 = 0.3
p(Y=1) = 30 + 40 = 70 / 100 = 0.7

Next, we can calculate the joint probabilities:

p(X=0, Y=0) = 10 / 100 = 0.1
p(X=0, Y=1) = 30 / 100 = 0.3
p(X=1, Y=0) = 20 / 100 = 0.2
p(X=1, Y=1) = 40 / 100 = 0.4

Now we can calculate the mutual information:

$I(X;Y) = 0.1 \log_2(0.1 / (0.4 * 0.3)) +$
$0.3 \log_2(0.3 / (0.4 * 0.7)) +$
$0.2 \log_2(0.2 / (0.6 * 0.3)) +$
$0.4 \log_2(0.4 / (0.6 * 0.7)) = 0.029$

Therefore, the mutual information between X and Y is 0.029.