

Web Mining

DATA

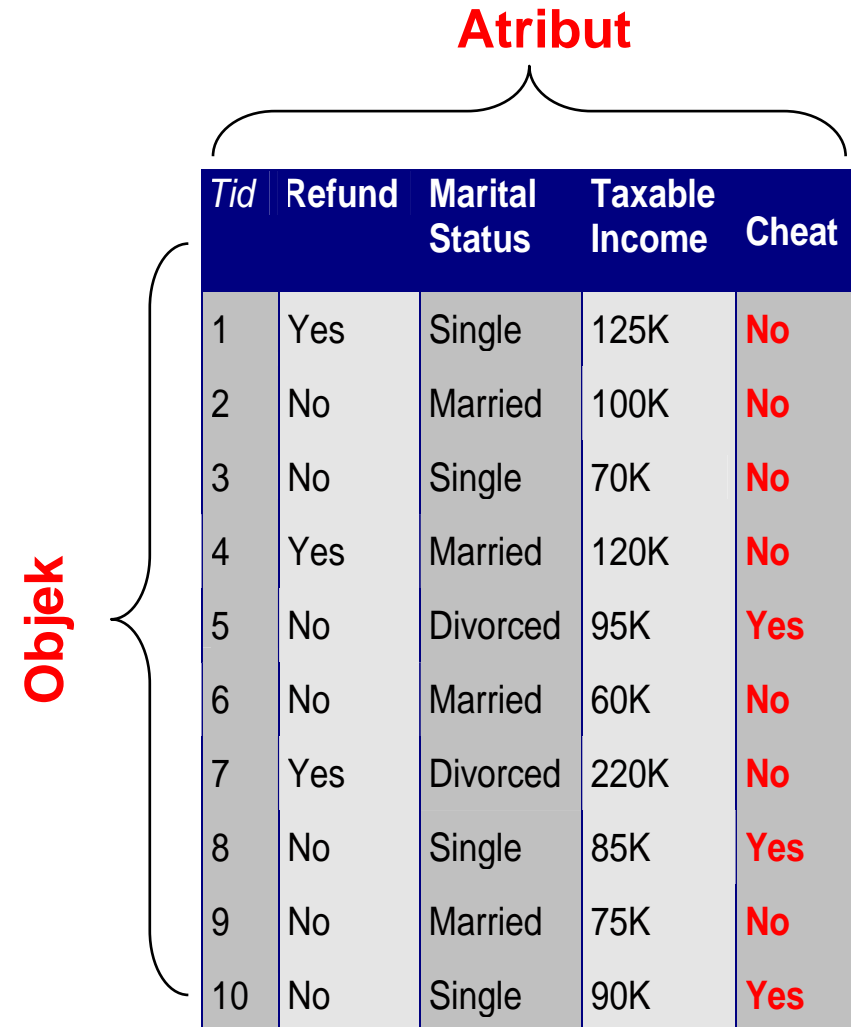
Prodi Teknik Informatika
Universitas Trunojoyo Madura
2024

Garis besar

- Atribut dan Objek
- Jenis Data
- Kualitas Data
- Kesamaan dan Jarak
- Pra-pemrosesan Data

Apa itu Data?

- Kumpulan **objek data** dan **atributnya**
- Atribut adalah properti atau karakteristik dari suatu **objek**
 - Contoh: warna mata seseorang, suhu, dll.
 - Atribut juga dikenal sebagai variabel, bidang (field), karakteristik, dimensi, atau fitur
- Kumpulan atribut menggambarkan **objek**
 - Objek juga dikenal sebagai record, titik, kasus, sampel, entitas, atau instan



The diagram illustrates the relationship between objects and attributes using a table. A vertical bracket on the left, labeled 'Objek' in red, groups the rows of the table. A horizontal bracket at the top, labeled 'Atribut' in red, groups the columns of 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

Tampilan Data Lebih Lengkap

- Data mungkin memiliki bagian
- Bagian data yang berbeda mungkin memiliki hubungan
- Lebih umum, data mungkin memiliki struktur
- Data bisa tidak lengkap
 - Akan dibahas ini lebih detail nanti

Nilai Atribut

- **Nilai atribut** adalah angka atau simbol yang diberikan pada atribut untuk objek tertentu
- Perbedaan antara atribut dan nilai atribut
 - Atribut yang sama dapat dipetakan ke nilai atribut yang berbeda
 - Contoh: tinggi dapat diukur dalam kaki atau meter
 - Atribut yang berbeda dapat dipetakan ke kumpulan nilai yang sama
 - Contoh: Nilai atribut untuk ID dan umur adalah bilangan bulat
 - Tetapi properti nilai atribut bisa berbeda

Pengukuran Panjang

- Cara mengukur atribut mungkin beda dengan properti atribut.

Skala ini hanya mempertahankan properti pengurutan panjang.



Skala ini mempertahankan sifat urutan dan penambahan panjang.

Jenis Atribut

- Ada berbagai jenis atribut
 - **Nominal**
 - Contoh: nomor ID, warna mata, kode pos
 - **Urut (Ordinal)**
 - Contoh: peringkat (misalnya, rasa keripik kentang dalam skala 1-10), tingkatan, tinggi {tinggi, sedang, pendek}
 - **Selang (Interval)**
 - Contoh: tanggal kalender, suhu dalam Celsius atau Fahrenheit.
 - **Perbandingan (Rasio)**
 - Contoh: suhu dalam Kelvin, panjang, waktu, hitungan

Properti Nilai Atribut

- Jenis atribut bergantung pada properti/operasi berikut yang dimilikinya:
 - Perbedaan: $= \neq$
 - Urutan: $< >$
 - Perbedaan bermakna : $+ -$
 - Rasio bermakna : $* /$
- Atribut nominal: perbedaan
- Atribut ordinal: perbedaan & keteraturan
- Atribut interval: perbedaan, keteraturan & perbedaan yang bermakna
- Atribut rasio: semua 4 properti/operasi

Perbedaan Antara Rasio dan Interval

- Apakah bermakna secara fisik untuk mengatakan bahwa suhu 10° dua kali lipat dari 5° pada
 - skala Celcius?
 - skala Fahrenheit?
 - skala Kelvin?
- Pertimbangkan untuk mengukur tinggi badan di atas rata-rata
 - Jika tinggi Bill tiga inci di atas rata-rata dan tinggi Bob enam inci di atas rata-rata, apakah kita akan mengatakan bahwa Bob dua kali lebih tinggi dari Bill?
 - Apakah situasi ini punya analogi sama dengan suhu?

		Attribute Type	Description	Examples	Operations
Categorical Qualitative		Nominal	Nominal attribute values only distinguish. (=, ≠)	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

Kategorisasi atribut ini berdasarkan SS 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.

Kategorisasi atribut ini berdasarkan SS Stevens

Atribut Diskrit dan Kontinu

- Atribut Diskrit (Discrete Attribute)
 - Hanya memiliki kumpulan nilai yang terbatas atau tak terbatas
 - Contoh: kode pos, hitungan, atau rangkaian kata dalam kumpulan dokumen
 - Sering direpresentasikan sebagai variabel integer.
 - Catatan: **atribut biner** adalah kasus khusus dari atribut diskrit
- Atribut Berkelanjutan (Continuous Attribute)
 - Memiliki bilangan real sebagai nilai atribut
 - Contoh: suhu, tinggi, atau berat.
 - Praktisnya, nilai riil hanya dapat diukur dan direpresentasikan menggunakan jumlah digit yang terbatas.
 - Atribut kontinu biasanya direpresentasikan sebagai variabel floating-point.

Atribut Asimetris

- Hanya kehadiran (nilai atribut bukan nol) yang dianggap penting
 - Kata-kata yang ada dalam dokumen
 - Item yang ada dalam transaksi pelanggan
- Dibutuhkan dua atribut biner asimetris untuk mewakili satu atribut biner biasa
 - Analisis asosiasi menggunakan atribut asimetris
- Atribut asimetris biasanya muncul dari objek yang merupakan himpunan

Kritik

- Tidak lengkap (perlu kajian lebih lanjut)
 - Asymmetric binary
 - Cyclical
 - Multivariate
 - Partially ordered
 - Partial membership
 - Relationships between the data
- Data pada kenyataannya adalah perkiraan dan (noisy / tidak bersih)
 - Dapat mempersulit pengenalan jenis atribut yang tepat
 - Memperlakukan satu jenis atribut sebagai atribut lain mungkin kurang lebih benar

Kritik ...

- Tidak ada pendekatan baku untuk analisis statistik
 - Mungkin tidak perlu membatasi operasi dan hasil
 - Analisis statistik seringkali merupakan perkiraan
 - Jadi, misalnya, menggunakan analisis interval untuk nilai ordinal dapat dibenarkan
 - Transformasi umum digunakan tetapi tidak mempertahankan skala
 - Dapat mengubah data ke skala baru dengan properti statistik yang lebih baik
 - Banyak analisis statistik hanya bergantung pada distribusi

Contoh Tantangan

- nomor identitas
 - Nominal, ordinal, atau interval?
- Jumlah silinder pada mesin mobil
 - Nominal, ordinal, atau rasio?
- Skala Bias
 - Interval atau Rasio

Pesan Utama untuk Jenis Atribut

- Jenis operasi yang dipilih harus “bermakna” untuk jenis data yang dimiliki
 - Nominal, Ordinal, interval dan rasio hanyalah empat sifat data
 - Tipe data yang terlihat – seringkali berupa angka atau string – mungkin tidak menangkap semua properti atau mungkin menggambarkan properti yang tidak ada
 - Analisis mungkin bergantung pada properti data lainnya ini
 - Banyak analisis statistik hanya bergantung pada distribusi
 - Sering kali makna dari data diukur dengan signifikansi statistik
 - Namun pada akhirnya, makna data diukur dengan domainnya

Jenis dataset

- Record
 - Matriks Data
 - Data Dokumen
 - Data Transaksi
- Graph
 - World Wide Web
 - Struktur Molekul
- Ordered
 - Data spasial
 - Data Sementara
 - Data Berurutan
 - Data Urutan Genetik

Karakteristik Penting Data

- Dimensi (jumlah atribut)
 - Data dimensi tinggi membawa sejumlah tantangan
- Ketersebaran
 - Hanya kehadiran yang diperhitungkan
- Resolusi
 - Pola tergantung pada skala
- Ukuran
 - Jenis analisis mungkin bergantung pada ukuran data

Data Record

- Data yang terdiri dari kumpulan record, yang masing-masing terdiri dari kumpulan atribut tetap

<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

Data Matriks

- Jika objek data memiliki kumpulan atribut numerik tetap yang sama, maka objek data dapat dianggap sebagai titik dalam ruang multidimensi, di mana setiap dimensi mewakili atribut yang berbeda.
- Kumpulan data tersebut dapat diwakili oleh matriks $m \times n$, di mana ada m baris, satu untuk setiap objek, dan n kolom, satu untuk setiap atribut

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

Data Dokumen

- Setiap dokumen menjadi vektor '**term**'
 - Setiap **term** adalah **komponen** (atribut) dari **vektor**
 - Nilai setiap komponen adalah berapa kali **term** yang sesuai **muncul dalam dokumen**

	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

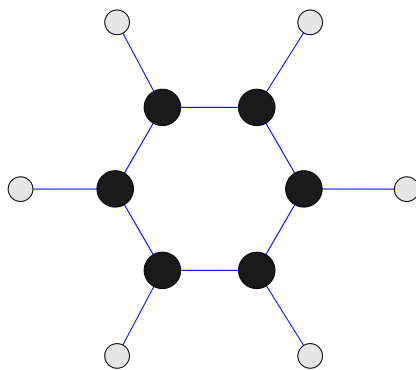
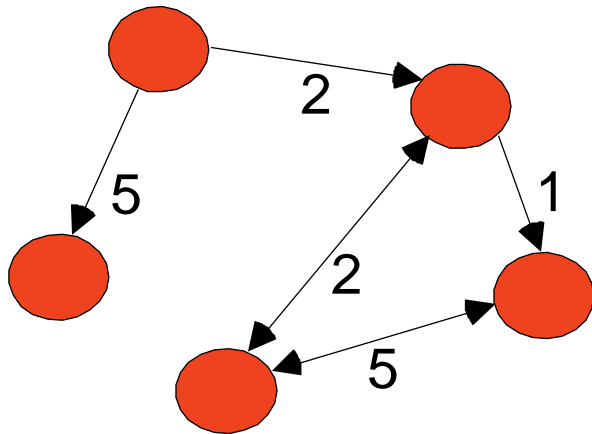
Data Transaksi

- Jenis data record khusus, di mana
 - Setiap record (transaksi) melibatkan satu set item.
 - Misalnya, pertimbangkan toko bahan makanan. Kumpulan produk yang dibeli oleh pelanggan selama satu perjalanan belanja merupakan transaksi, sedangkan produk individual yang dibeli adalah itemnya.

<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

Data Grafik

- Contoh: Grafik generik, molekul , dan halaman web



Molekul Benzena: 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 uthurasamy, "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

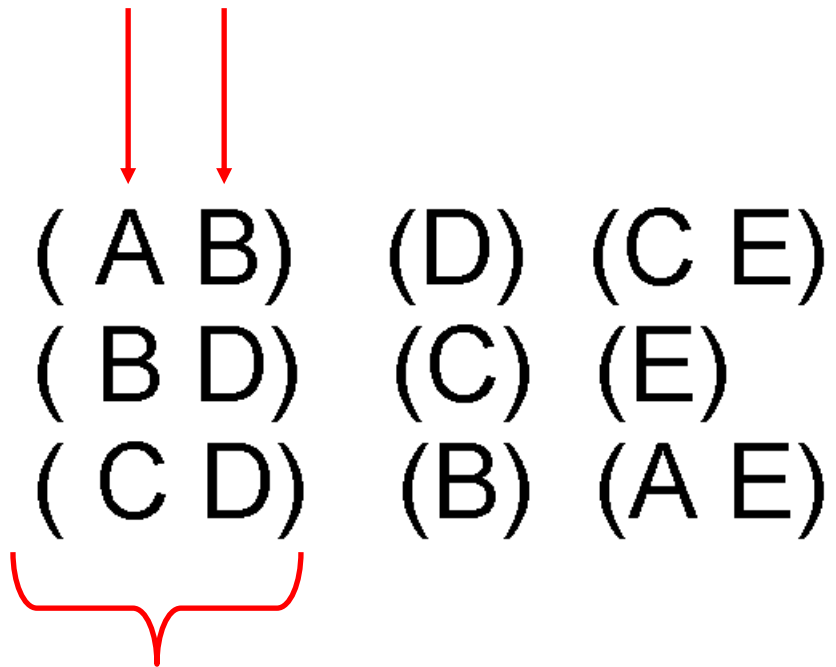
Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

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.

Data terurut (ordered)

- Urutan transaksi

Items/Events



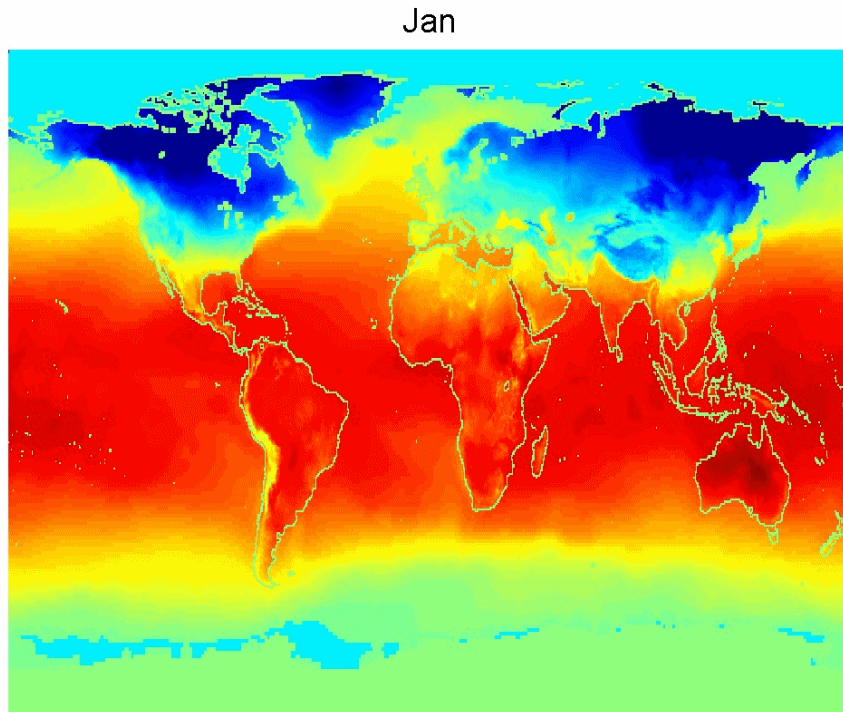
**Sebuah elemen
dari urutan**

Data terurut (ordered)

- Data urutan genom

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

Data terurut (ordered)



- Data spasial dan temporal (Ruang-Waktu)

**Suhu rata-rata
bulanan daratan
dan lautan**

Kualitas data

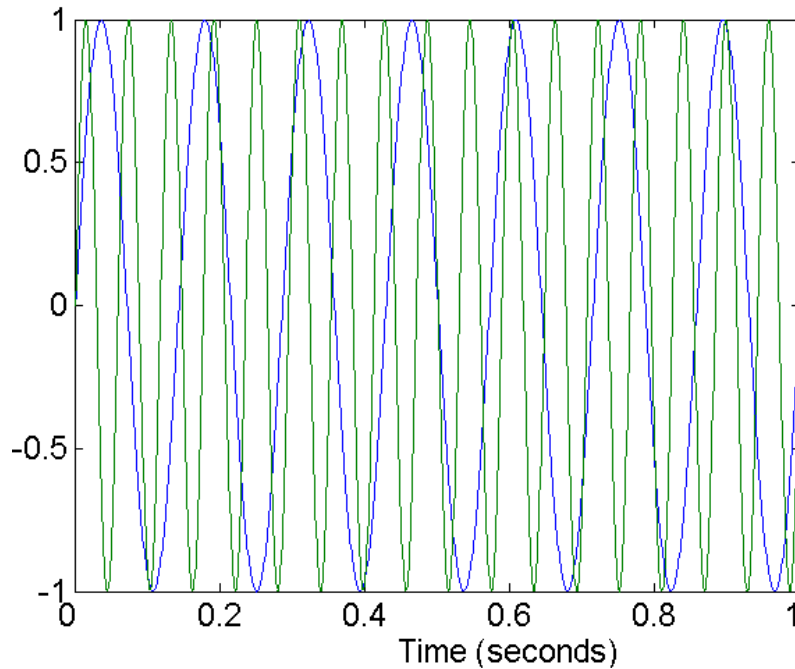
- Kualitas data yang buruk secara negatif mempengaruhi banyak upaya pemrosesan data
- Contoh penambangan data: model klasifikasi untuk mendeteksi orang yang pinjamannya berisiko dibangun menggunakan data yang buruk
 - Beberapa kandidat yang layak kredit ditolak pinjamannya
 - Lebih banyak pinjaman diberikan kepada individu yang gagal bayar

Kualitas data ...

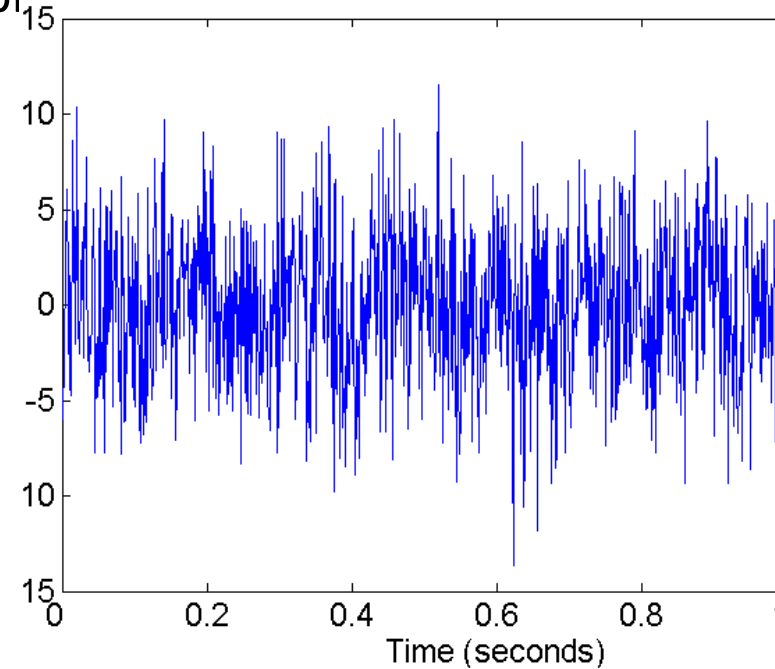
- Masalah kualitas data seperti apa?
- Bagaimana kita dapat mendeteksi masalah dengan data?
- Apa yang dapat kita lakukan tentang masalah ini?
- Contoh masalah kualitas data:
 - Kebisingan (noise) dan outlier
 - Nilai yang hilang
 - Duplikat data
 - Data yang salah

Kebisingan / Noise

- Untuk objek, kebisingan/noise adalah objek asing
- Untuk atribut, noise mengacu pada modifikasi nilai asli
- Contoh: distorsi suara seseorang saat berbicara di telepon yang buruk



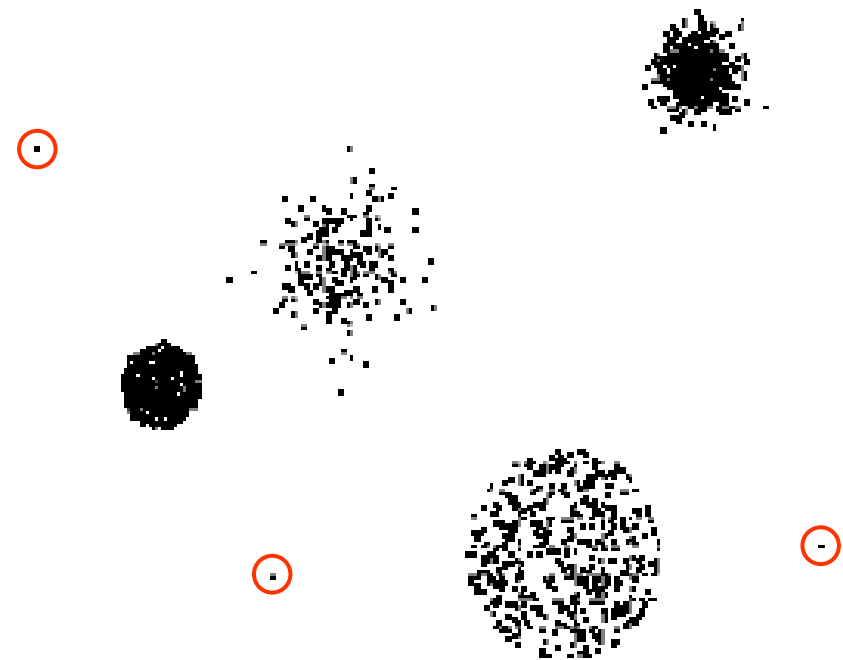
Dua Gelombang Sinus



Dua Gelombang Sinus + Kebisingan

Penyimpangan / outlier

- **Outlier** adalah objek data dengan karakteristik yang sangat berbeda dari sebagian besar objek data lainnya dalam kumpulan data
 - **Kasus 1:** Outlier adalah noise yang mengganggu analisis data
 - **Kasus 2:** Outlier adalah tujuan dari analisis kami
 - Penipuan kartu kredit
 - Deteksi gangguan
- Penyebab?



Missing Value / Nilai yang hilang

- Alasan terjadinya missing value
 - Informasi tidak dikumpulkan (misalnya, orang menolak menyebutkan usia dan berat badan mereka)
 - Atribut mungkin tidak berlaku untuk semua kasus (misalnya, pendapatan tahunan tidak berlaku untuk anak-anak)
- Menangani missing value
 - Hilangkan objek atau variabel data
 - Perkirakan nilai yang hilang
 - Contoh: deret waktu suhu
 - Contoh: hasil sensus
 - Abaikan nilai yang hilang selama analisis

Nilai yang hilang ...

- Missing completely at random (MCAR)
 - Hilangnya suatu nilai tidak tergantung pada atribut
 - Isikan nilai berdasarkan atribut
 - Analisis mungkin tidak bias secara keseluruhan
- Missing at Random (MAR)
 - Missingness terkait dengan variabel lain
 - Isikan nilai berdasarkan nilai lain
 - Hampir selalu menghasilkan bias dalam analisis
- Missing Not at Random (MNAR)
 - Missingness terkait dengan pengukuran yang tidak teramati
 - informasi atau tidak dapat diabaikan
- Tidak mungkin untuk mengetahui situasi dari data

Duplikat Data

- Kumpulan data dapat mencakup objek data yang merupakan duplikat, atau hampir duplikat satu sama lain
 - Masalah utama saat menggabungkan data dari sumber yang heterogen
- Contoh:
 - Orang yang sama dengan beberapa alamat email
- Pembersihan data
 - Proses menangani masalah data duplikat
- Kapan sebaiknya data duplikat tidak dihapus?

Ukuran Kemiripan dan Perbedaan

- Ukuran kesamaan
 - Ukuran numerik tentang seberapa mirip dua objek data.
 - Lebih tinggi ketika objek lebih mirip.
 - Sering jatuh dalam kisaran $[0,1]$
- Ukuran ketidaksamaan
 - Ukuran numerik tentang perbedaan dua objek data
 - Lebih rendah saat objek lebih mirip
 - Perbedaan minimum seringkali 0
 - Batas atas bervariasi
- **Kedekatan** mengacu pada kesamaan atau ketidaksamaan

Kesamaan/Ketidaksamaan untuk Atribut Sederhana

- Tabel berikut menunjukkan kesamaan dan ketidaksamaan antara dua objek, x dan y , *terhadap* satu atribut sederhana.

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 $	$s = -d, s = \frac{1}{1+d}, s = e^{-d},$ $s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Jarak Euclidean

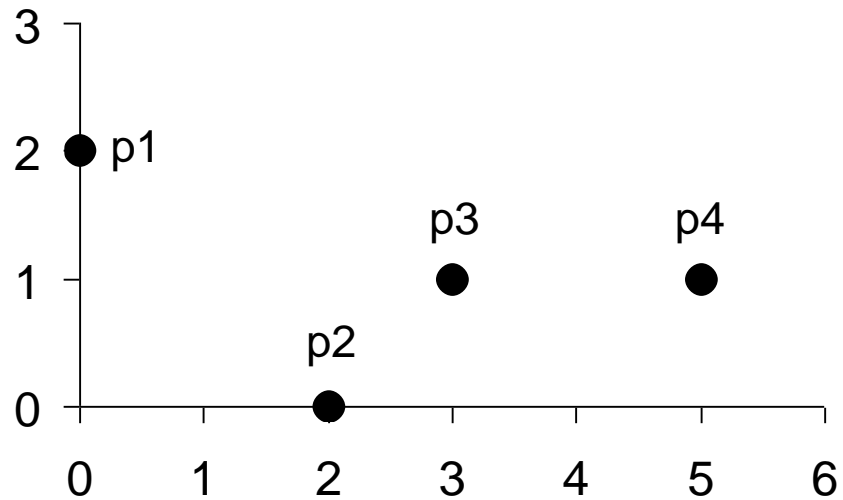
- Jarak Euclidean

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

di mana n adalah jumlah dimensi (atribut) dan x_k dan y_k adalah k^{th} atribut (komponen) atau objek data \mathbf{x} dan \mathbf{y} .

- Standardisasi diperlukan, jika skalanya berbeda.

Jarak Euclidean



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

Matriks Jarak

Jarak Minkowski

- Minkowski ada :

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

Di mana r adalah parameter, n adalah jumlah dimensi (atribut) dan x_k dan y_k masing-masing adalah k^{th} atribut (komponen) atau objek data \mathbf{x} dan \mathbf{y} .

Jarak Minkowski: Contoh

- $r = 1$. Jarak blok kota (Manhattan, taxicab, L_1 norm).
 - Contoh umum dari hal ini adalah jarak Hamming, yang hanya merupakan jumlah bit yang berbeda antara dua vektor biner
- $r = 2$. Jarak Euclidean
- $R \rightarrow \infty$. “ tertinggi ” (L_{maks} norma, L_{∞} norma) jarak.
 - Ini adalah perbedaan maksimum antara setiap komponen vektor
- Jangan bingung r dengan n , yaitu, semua jarak ini ditentukan untuk semua jumlah dimensi.

Jarak Minkowski

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

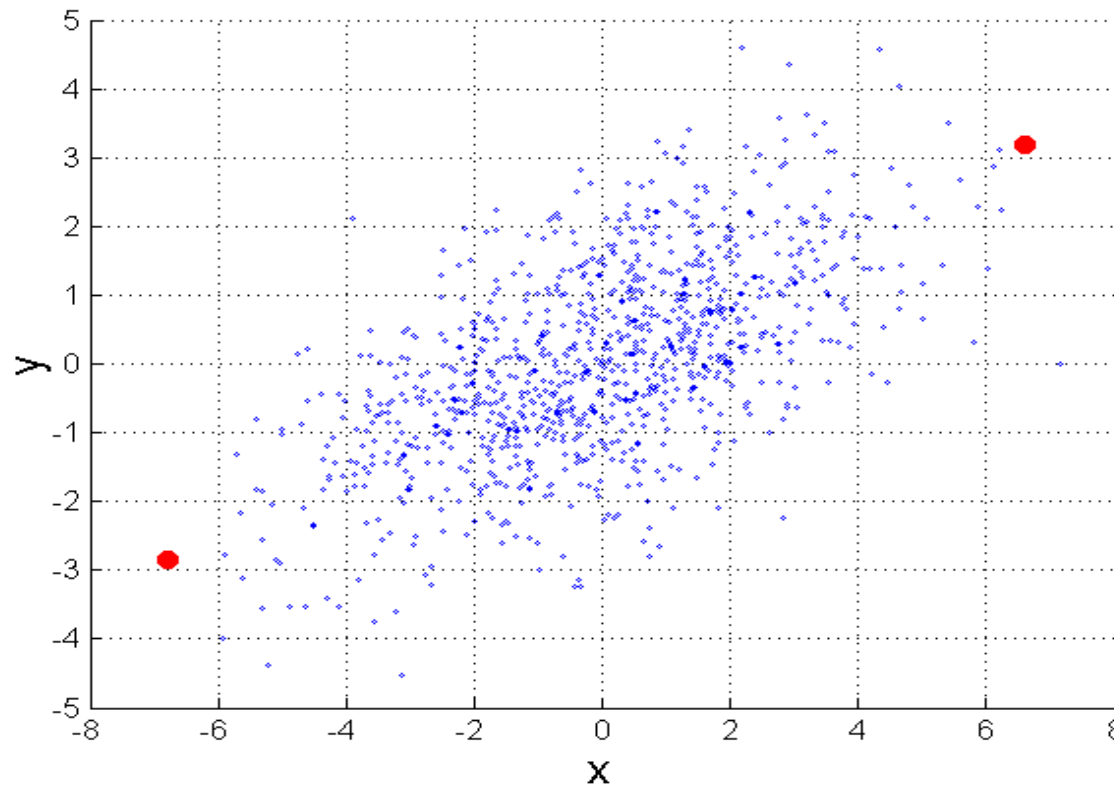
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

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

Matriks Jarak

Jarak Mahalanobis

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



Σ adalah matriks covariance
(hubungan arah antar dua
variable acak)

Untuk titik merah, jarak Euclidean adalah 14,7, jarak Mahalanobis adalah 6.

Jarak Mahalanobis

Matriks Kovariansi:

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

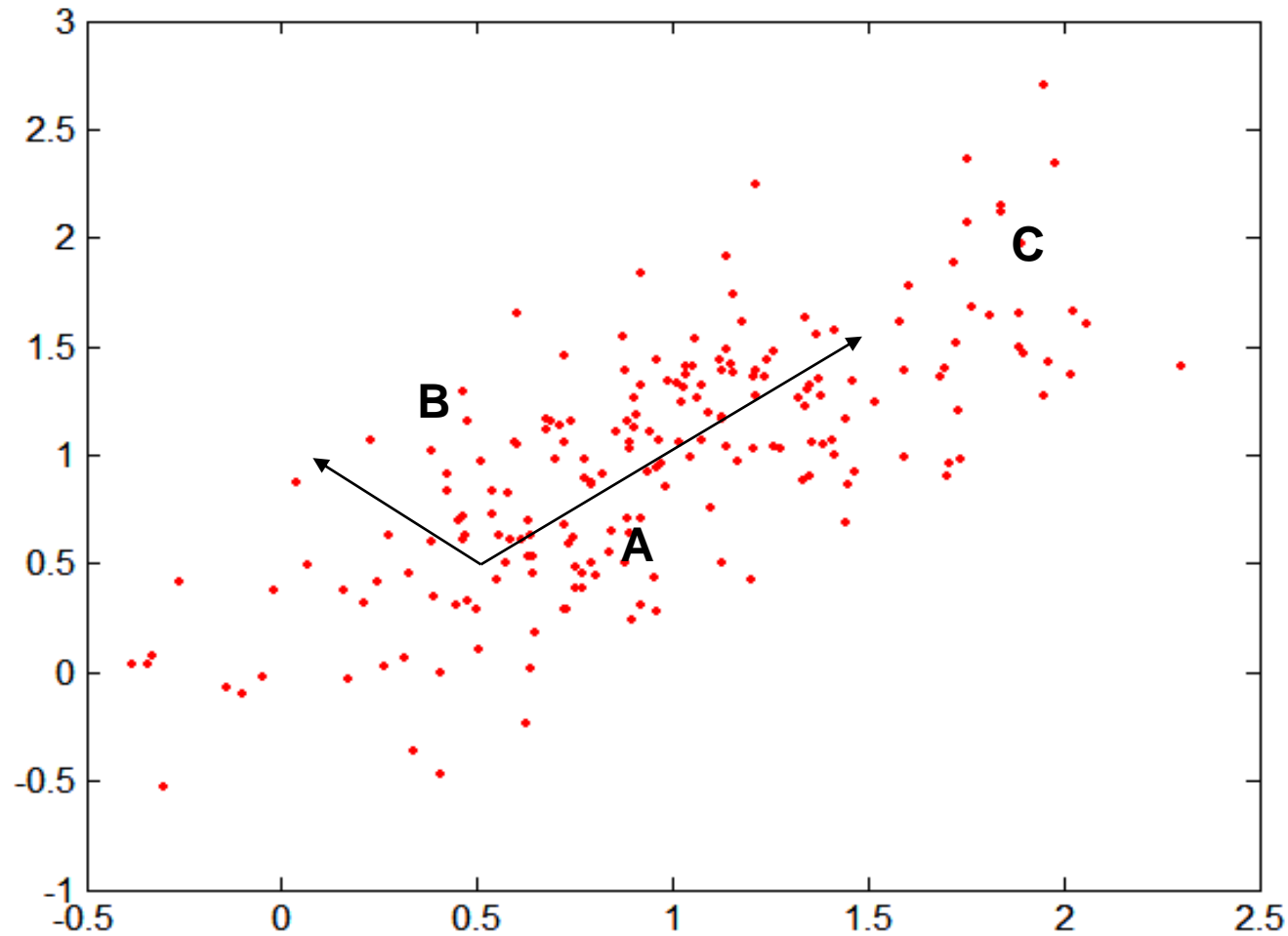
J: (0,5, 0,5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4



Sifat Umum Jarak

- Jarak, seperti jarak Euclidean, memiliki beberapa sifat yang terkenal.
 1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ untuk semua \mathbf{x} dan \mathbf{y} dan $d(\mathbf{x}, \mathbf{y}) = 0$ hanya jika $\mathbf{x} = \mathbf{y}$. (Kepastian positif)
 2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ untuk semua \mathbf{x} dan \mathbf{y} . (Simetri)
 3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ untuk semua titik \mathbf{x} , \mathbf{y} , dan \mathbf{z} . (Pertidaksamaan Segitiga)

dimana $d(\mathbf{x}, \mathbf{y})$ adalah jarak (perbedaan) antara titik (objek data), \mathbf{x} dan \mathbf{y} .

- Jarak yang memenuhi sifat-sifat ini adalah **metrik**

Sifat Umum dari Kesamaan

- Kesamaan, juga memiliki beberapa sifat.
 1. $s(\mathbf{x}, \mathbf{y}) = 1$ (atau kesamaan maksimum) hanya jika $\mathbf{x} = \mathbf{y}$.
 2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ untuk semua \mathbf{x} dan \mathbf{y} . (Simetri)

dimana $s(\mathbf{x}, \mathbf{y})$ adalah kesamaan antara titik (objek data), \mathbf{x} dan \mathbf{y} .

Kesamaan Antara Vektor Biner

- Situasi umum adalah objek, p dan q , hanya memiliki atribut biner
- Hitung kesamaan menggunakan jumlah berikut

f_{01} = jumlah atribut dimana p adalah 0 dan q adalah 1

f_{10} = jumlah atribut dimana p adalah 1 dan q adalah 0

f_{00} = jumlah atribut dimana p adalah 0 dan q adalah 0

f_{11} = jumlah atribut dimana p adalah 1 dan q adalah 1

- Simple Matching dan Koefisien Jaccard

SMC = jumlah kecocokan / jumlah atribut

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

J = jumlah 11 kecocokan / jumlah atribut bukan nol

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

SMC versus Jaccard: Contoh

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

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

$f_{01} = 2$ (jumlah atribut dimana x adalah 0 dan y adalah 1)

$f_{10} = 1$ (jumlah atribut di mana x adalah 1 dan y adalah 0)

$f_{00} = 7$ (jumlah atribut dimana x adalah 0 dan y adalah 0)

$f_{11} = 0$ (jumlah atribut dimana x adalah 1 dan y adalah 1)

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

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

Kesamaan Cosinus / Cosine Similarity

- Jika \mathbf{d}_1 dan \mathbf{d}_2 adalah dua vektor dokumen, maka

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

di mana $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ menunjukkan perkalian dalam atau perkalian titik vektor dari vektor, \mathbf{d}_1 Dan \mathbf{d}_2 , dan $\|\mathbf{d}\|$ adalah panjang vektor \mathbf{d} .

- Contoh:

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

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

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

$$\|\mathbf{d}_1\| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0,5} = (42)^{0,5} = 6,481$$

$$\|\mathbf{d}_2\| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0,5} = (6)^{0,5} = 2,449$$

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0,3150$$

Koefisien Jaccard Diperpanjang (Tanimoto)

- Variasi Jaccard untuk atribut kontinyu atau hitungan
 - Dikurangi menjadi Jaccard untuk atribut biner

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

Korelasi mengukur hubungan linier antar objek

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

where we are using the following standard statistical notation and definitions

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

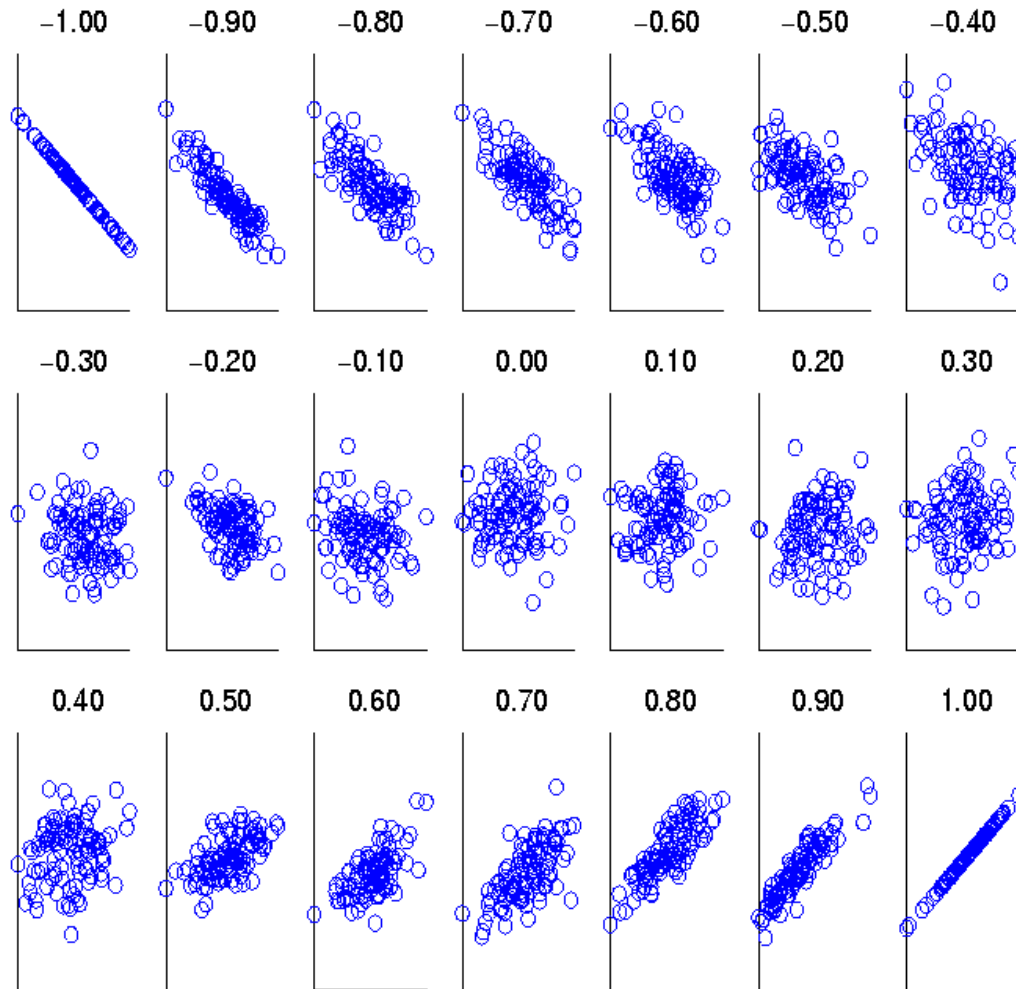
$$\text{standard_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$$

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

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

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

Mengevaluasi Korelasi Secara Visual



**Plot pencar
menunjukkan
kesamaan dari
-1 ke 1.**

Kekurangan Korelasi

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

$$\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$$

$$y_i = x_i^2$$

$$\text{mean}(\mathbf{x}) = 0, \text{mean}(\mathbf{y}) = 4$$

$$\text{std}(\mathbf{x}) = 2.16, \text{std}(\mathbf{y}) = 3.74$$

$$\begin{aligned} \text{corr} &= (-3)(5) + (-2)(0) + (-1)(-3) + (0)(-4) + (1)(-3) + (2)(0) + 3(5) / (6 * 2,16 * 3,74) \\ &= 0 \end{aligned}$$

Perbandingan Tindakan Kedekatan

- Domain aplikasi
 - Ukuran kesamaan cenderung spesifik untuk jenis atribut dan data
 - Rekam data, gambar, grafik, urutan, struktur protein 3D, dll cenderung memiliki ukuran yang berbeda
- Namun, seseorang dapat berbicara tentang berbagai properti yang Anda ingin memiliki ukuran kedekatan
 - Simetri adalah hal yang umum
 - Toleransi terhadap kebisingan dan outlier adalah hal lain
 - Kemampuan untuk menemukan lebih banyak jenis pola?
 - Banyak yang lain mungkin
- Pengukuran harus dapat diterapkan pada data dan menghasilkan hasil yang sesuai dengan pengetahuan domain

Tindakan Berbasis Informasi

- Teori informasi adalah murid yang berkembang dengan baik dan mendasar dengan aplikasi yang luas
- Beberapa ukuran kesamaan didasarkan pada teori informasi
 - Saling informasi dalam berbagai versi
 - Koefisien Informasi Maksimal (MIC) dan tindakan terkait
 - Umum dan dapat menangani hubungan non-linear
 - Bisa rumit dan memakan waktu untuk menghitung

Informasi dan Probabilitas

- Informasi berkaitan dengan kemungkinan hasil dari suatu peristiwa
 - transmisi pesan, lemparan koin, atau pengukuran sepotong data
- Semakin pasti suatu hasil, semakin sedikit informasi yang dikandungnya dan sebaliknya
 - Misalnya, jika sebuah koin memiliki dua kepala, maka hasil kepala tidak memberikan informasi
 - Secara lebih kuantitatif, informasi tersebut terkait dengan probabilitas suatu hasil
 - Semakin kecil probabilitas suatu hasil, semakin banyak informasi yang diberikan dan sebaliknya
 - Entropi adalah ukuran yang umum digunakan



Entropi

- Untuk
 - variabel (peristiwa), X ,
 - dengan n kemungkinan nilai (hasil), $x_1, x_2 \dots, x_n$
 - setiap hasil memiliki probabilitas, $p_1, p_2 \dots$, *hal*
 - entropi dari X , $H(X)$, diberikan oleh

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

- Entropi antara 0 dan $\log_2 n$ dan diukur dalam bit
 - Jadi, entropi adalah ukuran berapa banyak bit yang diperlukan untuk mewakili rata-rata pengamatan

Contoh Entropi

- Untuk koin dengan probabilitas p kepala dan probabilitas $q = 1 - p$ ekor

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

- Untuk $p = 0,5$, $q = 0,5$ (koin wajar) $H = 1$
 - Untuk $p = 1$ atau $q = 1$, $H = 0$
-
- Berapa entropi dadu empat sisi yang adil?

Entropi untuk Data Sampel: Contoh

Warna rambut	Menghitung	P	$-p \log_2 hal$
Hitam	75	0,75	0,3113
Cokelat	15	0,15	0,4105
Berambut pirang	5	0,05	0,2161
Merah	0	0,00	0
Lainnya	5	0,05	0,2161
Total	100	1.0	1.1540

Entropi maksimum adalah $\log_2 5 = 2,3219$

Entropi untuk Data Sampel

- Misalkan kita punya
 - sejumlah pengamatan (m) terhadap beberapa atribut, X , misalnya warna rambut siswa di kelas tersebut,
 - dimana terdapat n nilai yang mungkin berbeda
 - Dan jumlah observasi pada kategori ^{ke} i adalah m_i
 - Kemudian, untuk sampel ini

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

- Untuk data kontinu, perhitungannya lebih sulit

Mutual Informations

- Informasi satu variabel menyediakan tentang yang lain

Secara formal, $I(X, Y) = H(X) + H(Y) - H(X, Y)$, di mana $H(X, Y)$ adalah entropi gabungan dari X dan Y ,

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

Di mana p_{ij} adalah probabilitas bahwa nilai ke- i dari X dan nilai ke- j dari Y terjadi bersamaan

- Untuk variabel diskrit, ini mudah dihitung
- Informasi timbal balik maksimum untuk variabel diskrit adalah $\log_2 (\min (n_X, n_Y))$, di mana n_X (n_Y) adalah jumlah nilai dari X (Y)

Contoh Informasi Bersama

Status Mahasiswa	Menghitung	P	$-p \log_2 hal$
Sarjana	45	0,45	0,5184
Lulusan	55	0,55	0,4744
Total	100	1.00	0,9928

Nilai	Menghitung	P	$-p \log_2 hal$
A	35	0,35	0,5301
B	50	0,50	0,5000
C	15	0,15	0,4105
Total	100	1.00	1.4406

Status Mahasiswa	Nilai	Menghitung	P	$-p \log_2 hal$
Sarjana	A	5	0,05	0,2161
Sarjana	B	30	0,30	0,5211
Sarjana	C	10	0,10	0,3322
Lulusan	A	30	0,30	0,5211
Lulusan	B	20	0,20	0,4644
Lulusan	C	5	0,05	0,2161
Total		100	1.00	2.2710

Informasi mutual Status dan Nilai Siswa = $0.9928 + 1.4406 - 2.2710 = 0.1624$

Koefisien Informasi Maksimal

- Reshef , David N., Yakir A. Reshef , Hilary K. Finucane , Sharon R. Grossman, Gilean McVean , Peter J. Turnbaugh , Eric S. Lander , Michael Mitzenmacher , dan Pardis C. Sabeti . "Mendeteksi asosiasi baru dalam kumpulan data besar." *sains* 334, no. 6062 (2011): 1518-1524 .
- Menerapkan informasi timbal balik ke dua variabel kontinu
- Pertimbangkan kemungkinan binning variabel ke dalam kategori diskrit
 - $n_X \times n_Y \leq N^{0,6}$ **Di mana**
 - n_X adalah banyaknya nilai dari X
 - n_Y adalah jumlah nilai dari Y
 - N adalah jumlah sampel (pengamatan, objek data)
- Hitung informasi timbal balik
 - Dinormalkan dengan $\log_2 (\min (n_X , n_Y)$
- Ambil nilai tertinggi

Pendekatan Umum untuk Menggabungkan Kesamaan

- Terkadang atribut terdiri dari berbagai jenis, tetapi kesamaan secara keseluruhan diperlukan.

1: Untuk atribut ke- k , hitung kesamaan, $s_k(\mathbf{x}, \mathbf{y})$, dalam rentang $[0, 1]$.

2: Tentukan variabel indikator, δ_k , untuk atribut k^{th} sebagai berikut:

$\delta_k = 0$ jika k^{th} atribut adalah atribut asimetris dan

kedua objek memiliki nilai 0, atau jika salah satu objek memiliki nilai yang hilang untuk atribut ke- k

$\delta_k = 1$ sebaliknya

3. Hitung

$$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \delta_k}$$

Menggunakan Bobot untuk Menggabungkan Kesamaan

- Mungkin tidak ingin memperlakukan semua atribut sama.
 - Gunakan bobot non-negatif ω_k

- $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \omega_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \omega_k \delta_k}$

- Bisa juga mendefinisikan bentuk jarak tertimbang

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

Kepadatan

- Mengukur sejauh mana objek data dekat satu sama lain di area tertentu
- Pengertian kerapatan erat kaitannya dengan kedekatan
- Konsep kepadatan biasanya digunakan untuk pengelompokan dan deteksi anomali
- Contoh:
 - Kepadatan Euclidean
 - Kepadatan Euclidean = jumlah titik per satuan volume
 - Kepadatan probabilitas
 - Perkirakan seperti apa distribusi data itu
 - Kepadatan berbasis grafik
 - Konektivitas

Kepadatan Euclidean: Pendekatan Berbasis Grid

- Pendekatan yang paling sederhana adalah dengan membagi wilayah menjadi sejumlah sel persegi panjang dengan volume yang sama dan menentukan densitas sebagai # titik yang dikandung sel

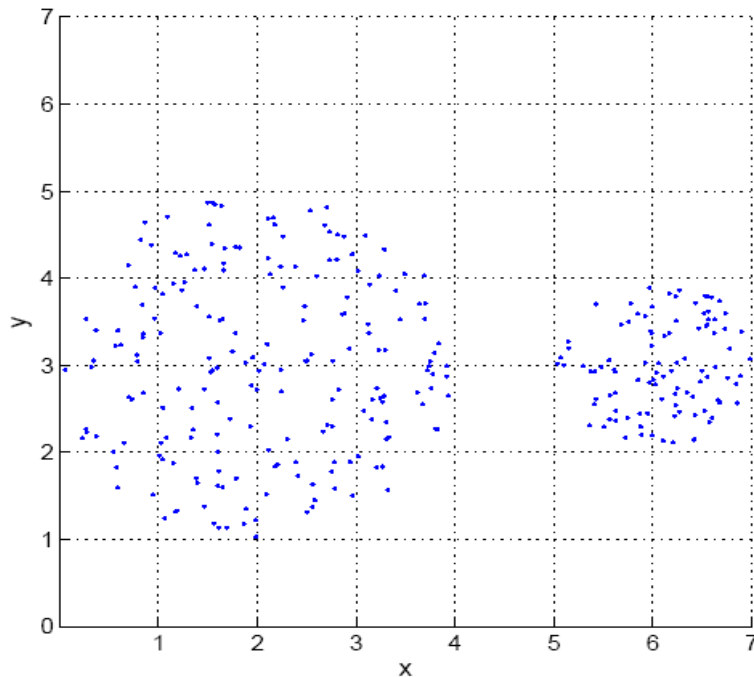


Figure 7.13. Cell-based density.

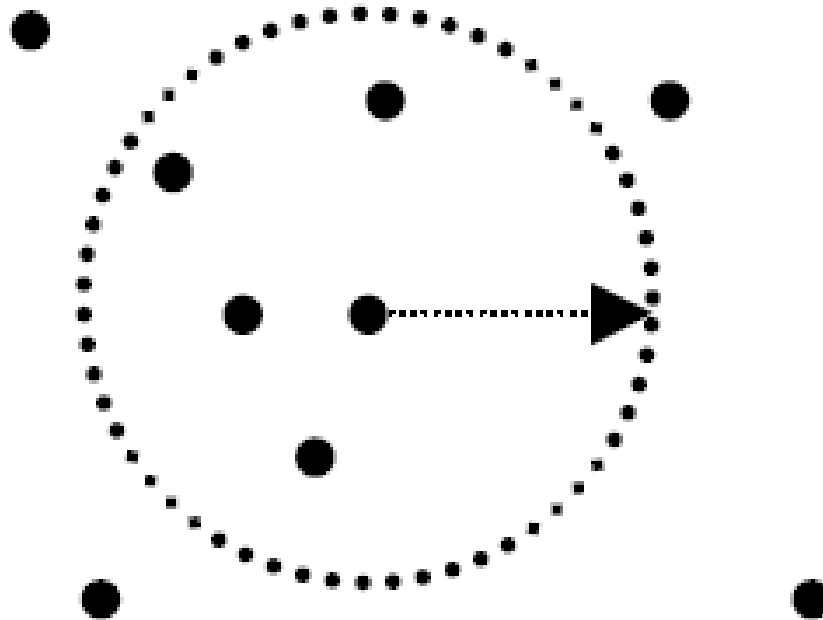
0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Table 7.6. Point counts for each grid cell.

Kepadatan berbasis grid. Jumlah untuk setiap sel.

Kepadatan Euclidean: Berbasis Pusat

- Kepadatan Euclidean adalah jumlah titik dalam radius tertentu dari titik tersebut



Ilustrasi kepadatan berbasis pusat.