

Tipe dan Jenis Data

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

November 16, 2019

Selayang Pandang

① Representasi Data

Tentang Data
Atribut

② Tipe Data

Record
Graf
Terurut

③ Masalah pada Data

Pengumuman

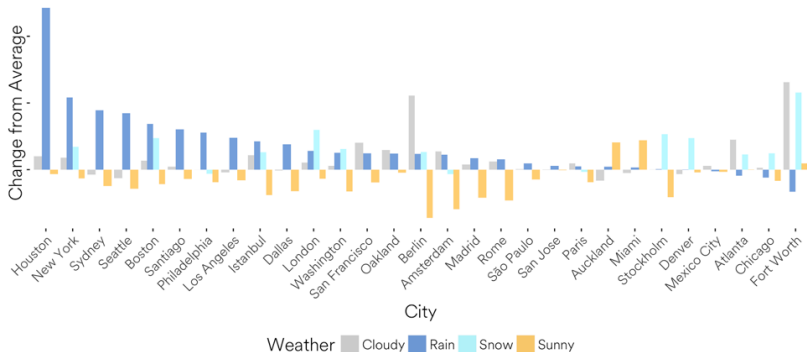
- 1 Segera *enroll* untuk kuliah Data Mining di e-learning!
- 2 Pengumpulan tugas akan dilakukan melalui e-learning
- 3 Ada baiknya untuk mengecek <http://www.inf.ed.ac.uk/teaching/courses/it/cribsheet.2up.pdf> karena akan mulai dipakai dalam waktu dekat
- 4 Siapkan Python, Pandas, dan Jupyter Notebook!

Representasi Data

Variabel seperti apa yang dapat dipakai oleh sistem rekomendasi berdasarkan konten dari aplikasi seperti Spotify?

Korelasi pada Spotify

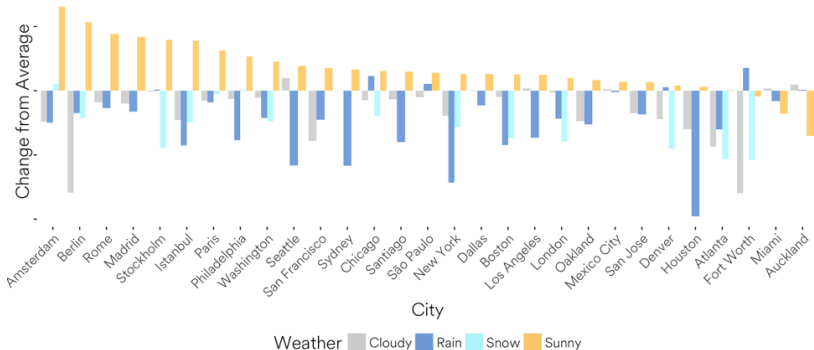
Impact of Weather on Acousticness



Gambar: Hubungan cuaca dengan “keakustikan” musik [van Buskirk, 2017] yang dilihat dari bunyi-bunyi alat akustik, e.g. gitar dan tamborin, dibandingkan dengan bunyi-bunyi elektronik, e.g. *synthesizer*

Korelasi pada Spotify

Impact of Weather on Energy



Gambar: Hubungan cuaca dengan “energi” musik [van Buskirk, 2017] yang dilihat dari kecepatan, volume, dan kebisingan, misalnya perbandingan kontras antara musik *death metal* dan komposisi Bach

Data, Atribut, dan Objek

Data

Data merupakan kumpulan objek (*instances*) yang memiliki atribut-atribut tertentu

Data, Atribut, dan Objek

Data

Data merupakan kumpulan objek (*instances*) yang memiliki atribut-atribut tertentu

Atribut

Karakteristik dari suatu objek, dikenal juga dengan nama **variabel** atau **fitur**

Data, Atribut, dan Objek

Data

Data merupakan kumpulan objek (*instances*) yang memiliki atribut-atribut tertentu

Atribut

Karakteristik dari suatu objek, dikenal juga dengan nama **variabel** atau **fitur**

Objek

Dikenal juga dengan nama **record**, **poin**, **sampel**, **entitas**, atau **instance**

Dari contoh kasus Spotify tadi, mana yang merupakan atributnya
dan mana yang merupakan objeknya?

Nilai dan Tipe dari Atribut

- 1 Nilai dari suatu atribut dapat berupa simbol maupun angka

Nilai dan Tipe dari Atribut

- 1 Nilai dari suatu atribut dapat berupa simbol maupun angka
- 2 Atribut yang sama dapat dipetakan ke beberapa nilai yang berbeda, misalnya karena beda satuan

Nilai dan Tipe dari Atribut

- 1 Nilai dari suatu atribut dapat berupa simbol maupun angka
- 2 Atribut yang sama dapat dipetakan ke beberapa nilai yang berbeda, misalnya karena beda satuan
- 3 Ada tiga tipe atribut secara umum: **categorical/nominal**, **ordinal**, **numeric**

Atribut Nominal

- 1 Atribut nominal bernilai saling lepas (*mutually exclusive*)
- 2 Perbandingan yang dapat dilakukan hanya menguji kesamaan ($=$, \neq)
- 3 Tidak dapat diurutkan maupun diukur jaraknya
- 4 Contoh: Warna mata, *genre* musik, pekerjaan

Atribut Ordinal

- 1 Terdapat urutan yang ada secara natural, e.g. {kecil, sedang, besar} atau {tidak suka, netral, suka}
- 2 Dikodekan sebagai angka untuk mempertahankan urutan sehingga dapat dibandingkan ($<$, $=$, $>$)
- 3 Terkadang sulit untuk dibedakan dengan nominal, e.g. apakah ada urutan untuk {belum menikah, menikah, bercerai}?

Atribut Numerik

- 1 Dapat bernilai bulat atau riil sehingga bisa dijumlahkan atau dirata-rata
- 2 Sensitif terhadap nilai ekstrem, e.g. tinggi:
 $\{165, 171, 182, 1850\}$
- 3 Terkadang dibedakan sebagai **ratio** dan **interval**

Ratio vs Interval

Ratio

Punya referensi nilai nol, e.g. berat, tinggi, jarak, suhu

Ratio vs Interval

Ratio

Punya referensi nilai nol, e.g. berat, tinggi, jarak, suhu

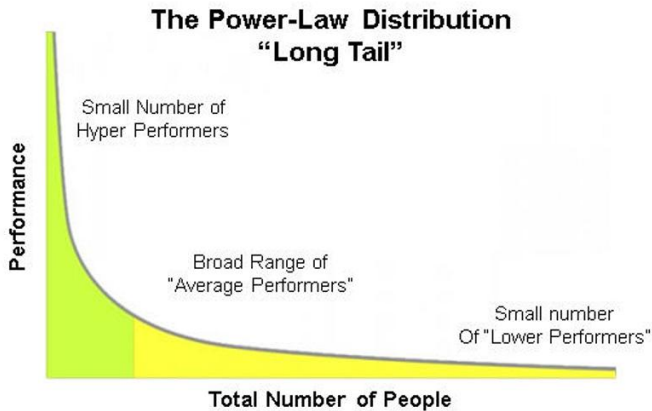
Interval

Tidak punya referensi nilai nol, e.g. tahun

Kasus dalam Atribut Numerik

- 1 Distribusi yang memiliki kecondongan, e.g. *power law distribution* $\rightarrow \log(x)$
- 2 Efek non-monotonik dari atribut, e.g. usia dalam menentukan pemenang marathon \rightarrow *binning*
- 3 Terkadang perlu dilakukan standardisasi, i.e. menjadi distribusi Gaussian standar, atau normalisasi, i.e. dalam rentang $[0, 1]$

Distribusi yang Condong



Gambar: *Power law distribution* [Bersin, 2014]

Tipe Data

Data Matriks

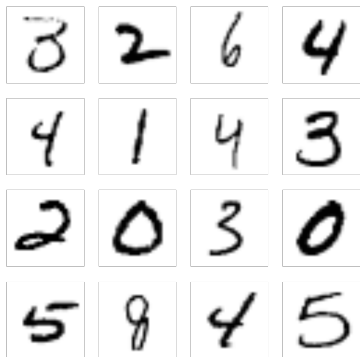
BUBUR.xlsx - LibreOffice Calc																			
File Edit View Insert Format Tools Data Window Help																			
A1																			
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Timestamp	Mana di	Alasan	Berapa kali rata	Merica	Seledri	Kerupuk/ke	Kacang	Topping	Kaldu	Cakue/cake	Kecap	Sambal	Bawang p	sum	Usus ayam	Kulit ayam	Telur puyu	Hati ayam
2	4/13/2015 21:57	Diaduk			5	0	1	1	0	0	1	1	1	0	5	1	0	0	0
3	4/13/2015 22:00	Diaduk	Rasanya lebih	Tidak tentu (da		0	1	1	0	0	1	1	1	0	5	1	0	0	0
4	4/13/2015 22:01	Diaduk	Biar rata semu		1	1	1	1	0	0	1	1	1	1	8	1	0	1	0
5	4/13/2015 22:02	Diaduk	Tidak tentu (da		1	1	0	0	1	1	1	0	1	1	6	0	0	1	0
6	4/13/2015 22:03	Diaduk	Biar bumbunya	Tidak tentu (da	1	0	1	1	0	0	1	0	1	0	4	0	0	0	0
7	4/13/2015 22:03	Diaduk	Tidak tentu (da		0	1	1	0	0	0	1	1	0	1	5	1	0	0	0
8	4/13/2015 22:03	Diaduk	Tidak tentu (da		0	1	1	0	1	1	1	1	1	1	7	0	0	1	1
9	4/13/2015 22:03	Diaduk	Tidak tentu (da		1	0	0	0	0	0	1	1	1	1	5	0	0	1	1
10	4/13/2015 22:03	Diaduk	Aneh mengapa	Tidak tentu (da	0	0	0	0	1	0	0	1	1	1	5	0	0	1	1
11	4/13/2015 22:04	Diaduk	Tidak tentu (da		0	0	0	1	0	0	0	1	0	1	2	0	0	1	0
12	4/13/2015 22:04	Diaduk	supaya semua	Tidak tentu (da	0	1	0	1	1	0	1	1	0	1	6	0	0	0	0
13	4/13/2015 22:04	Diaduk	Tidak tentu (da		1	0	1	0	1	1	1	1	0	0	5	1	0	0	0
14	4/13/2015 22:05	Diaduk	semua bagian	Tidak tentu (da	0	0	1	1	1	1	1	1	0	0	5	0	1	1	0
15	4/13/2015 22:05	Diaduk	biar nyampur	Tidak tentu (da	1	1	1	1	1	1	0	0	0	0	5	0	0	0	0
16	4/13/2015 22:06	Diaduk	kakau diaduk	je	Tidak tentu (da	0	0	1	0	1	1	1	0	0	4	1	0	0	0
17	4/13/2015 22:07	Diaduk	enak semua	n	Tidak tentu (da	1	1	1	0	1	1	1	0	0	1	0	0	0	0
18	4/13/2015 22:09	Diaduk	tercampur kalo	Tidak tentu (da	0	0	1	0	1	0	1	0	0	0	3	0	0	0	0
19	4/13/2015 22:09	Diaduk	Tidak tentu (da		1	1	0	0	1	1	0	0	1	0	5	1	0	0	0
20	4/13/2015 22:09	Diaduk	karena saya kr		0	1	0	1	0	0	0	1	4	1	0	0	1	0	0
21	4/13/2015 22:10	Diaduk	lah idat ga ny	Tidak tentu (da	0	1	1	0	1	1	1	1	0	0	6	0	0	0	0
22	4/13/2015 22:10	Diaduk	Suka2	Tidak tentu (da	0	1	1	0	1	1	1	1	0	0	5	1	0	0	0
23	4/13/2015 22:11	Diaduk	biar sama rata	Tidak tentu (da	0	0	1	1	1	1	1	1	0	0	6	1	1	1	0
24	4/13/2015 22:11	Diaduk	semuanya ny	Tidak tentu (da	1	1	1	1	1	1	1	1	1	1	9	1	1	1	0
25	4/13/2015 22:12	Diaduk	Kerupuknya	je	Tidak tentu (da	0	0	1	0	0	1	0	0	0	3	0	1	0	0
26	4/13/2015 22:12	Diaduk	biar cepet dino	Tidak tentu (da	1	0	1	0	0	1	1	1	1	0	5	1	0	1	0
27	4/13/2015 22:13	Diaduk	tidak jelas mo	Tidak tentu (da	1	1	1	0	1	1	1	1	1	1	8	1	0	1	0
28	4/13/2015 22:13	Diaduk	biar ga geli		0	1	1	0	1	1	1	1	0	1	7	0	0	1	0
29	4/13/2015 22:14	Diaduk	Tidak tentu (da		0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
30	4/13/2015 22:14	Diaduk	Sukasuka	Tidak tentu (da	0	0	1	0	1	1	1	1	1	1	6	0	0	1	0
31	4/13/2015 22:15	Diaduk	lucu	Tidak tentu (da	0	1	1	0	1	1	1	1	1	1	8	1	0	0	0
32	4/13/2015 22:15	Diaduk	Karena perpad	Tidak tentu (da	1	1	1	1	1	1	1	1	1	1	9	1	0	1	0
33	4/13/2015 22:17	Diaduk	Enak	Tidak tentu (da	1	1	1	1	1	1	1	1	1	1	9	1	1	1	0
34	4/13/2015 22:18	Diaduk	biar isinya kelo	Tidak tentu (da	1	1	1	0	1	1	1	1	1	1	8	1	0	1	0
35	4/13/2015 22:19	Diaduk	Tidak tentu (da		0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
36	4/13/2015 22:19	Diaduk	biar rasanya m	Tidak tentu (da	0	0	1	0	0	1	1	0	0	0	1	5	0	0	0
37	4/13/2015 22:20	Diaduk	liet orang sek	Tidak tentu (da	1	1	1	0	1	0	0	1	0	0	4	0	0	0	0
38	4/13/2015 22:20	Diaduk	Tidak tentu (da		1	1	1	0	1	1	0	1	1	1	7	0	1	0	0
39	4/13/2015 22:22	Diaduk	Tidak tentu (da		0	0	0	0	1	1	1	1	0	0	3	0	0	1	0

Gambar: Data preferensi bubur ayam

Data Matriks

- 1 Bentuk data paling sederhana
- 2 Sudah siap diolah
- 3 Dikenal juga sebagai **data terstruktur**
- 4 Contoh lain: data transaksi, hasil penapisan verbal

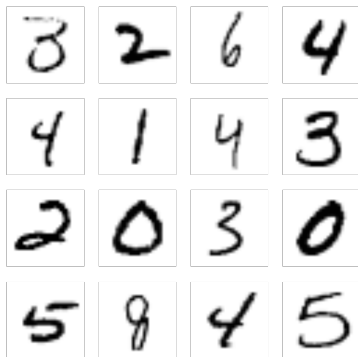
Gambar



- 1 Bagaimana cara merepresentasikan gambar?

Gambar: Contoh data MNIST
[O'Shea, 2016]

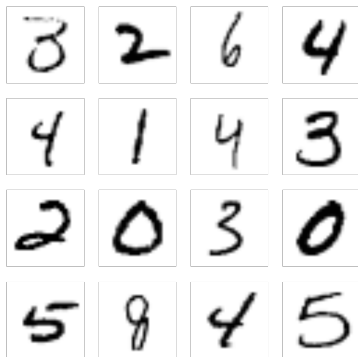
Gambar



- 1 Bagaimana cara merepresentasikan gambar?
- 2 Jika tiap pixel adalah atribut, berapa nilainya yang mungkin?

Gambar: Contoh data MNIST
[O'Shea, 2016]

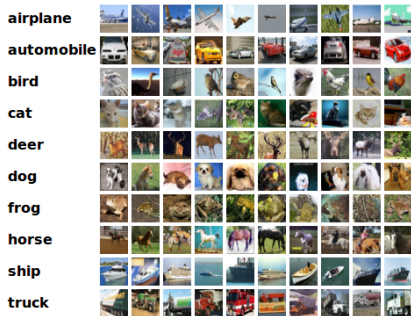
Gambar



Gambar: Contoh data MNIST
[O'Shea, 2016]

- 1 Bagaimana cara merepresentasikan gambar?
- 2 Jika tiap pixel adalah atribut, berapa nilainya yang mungkin?
- 3 Apa kelebihan dan kekurangannya?

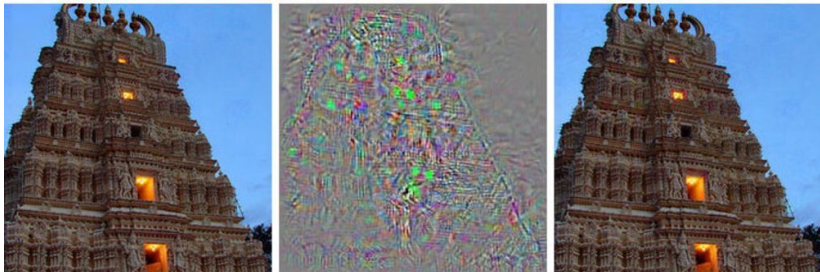
Gambar: Object Recognition



Gambar: Dataset CIFAR-10
[Krizhevsky, 2009]

- 1 Bagaimana dengan prediksi objek?
- 2 Tantangan: orientasi, skala, pencahayaan
- 3 Menggunakan pixels saja (mungkin) tidak cukup!
- 4 Bisa dibagi berdasarkan "region"

Misklasifikasi dalam Pengenalan Objek



Gambar: Gedung yang dianggap sebagai burung unta setelah diterapkan *noise*

Contoh tugas:

- ① berita → topik
- ② e-mail → spam
- ③ tweet → sentimen

Bagaimana merepresentasikannya?

Kata sebagai Atribut Numerik

- 1 Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**

Kata sebagai Atribut Numerik

- 1 Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**
- 2 Bernilai 1 jika terdapat di contoh teks, 0 jika tidak

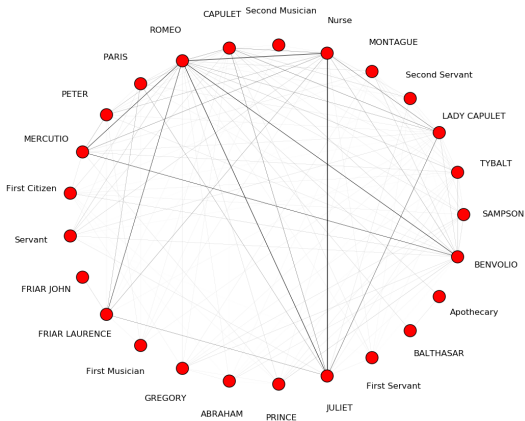
Kata sebagai Atribut Numerik

- 1 Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**
- 2 Bernilai 1 jika terdapat di contoh teks, 0 jika tidak
- 3 Dapat diubah menjadi frekuensi atau bobot (TF-IDF)

Kata sebagai Atribut Numerik

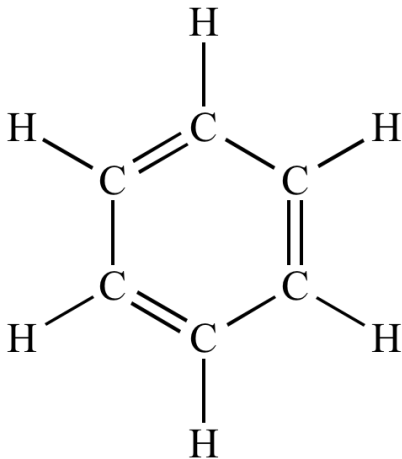
- 1 Representasi *bag-of-words* (BoW), i.e. **satu kata mewakili satu atribut**
- 2 Bernilai 1 jika terdapat di contoh teks, 0 jika tidak
- 3 Dapat diubah menjadi frekuensi atau bobot (TF-IDF)
- 4 **Catatan:** Dimensinya bisa jadi sangat besar dan matriksnya akan menjadi *sparse*

Jejaring Sosial



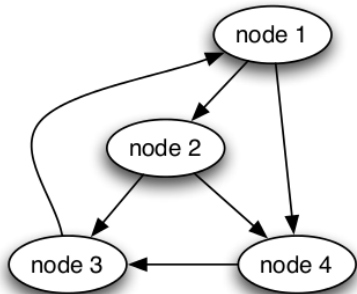
Gambar: Graf dari drama Romeo dan Juliet berdasarkan kemunculan karakter di satu babak yang sama

Struktur Kimia



Gambar: Struktur kimia benzena [Hardinger, 2017]

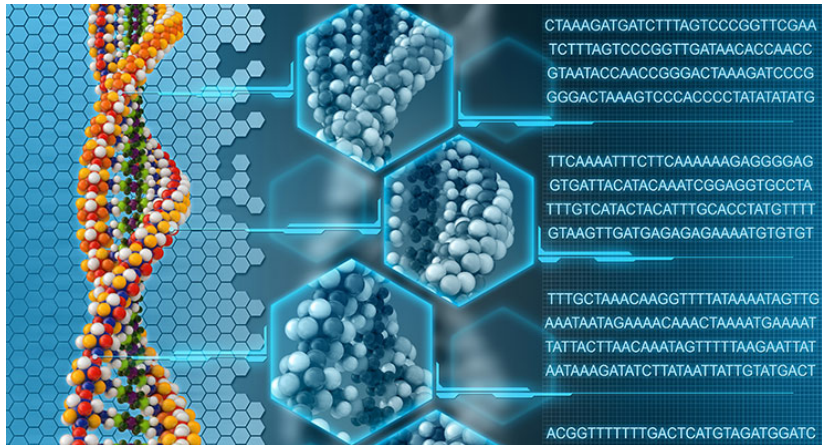
Adjacency Matrix



0	1	0	1
0	0	1	1
1	0	0	0
0	0	1	0

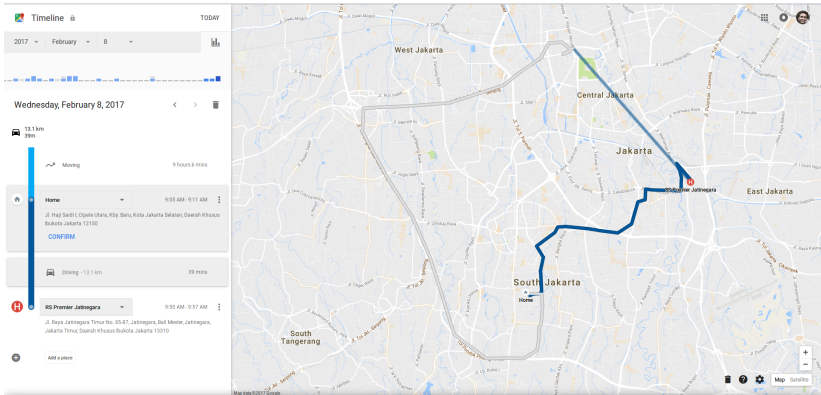
Gambar: Adjacency matrix dari graf [Easley dan Kleinberg, 2010]

Genomic Sequence



Gambar: Urutan genom [Global Biodefense, 2014]

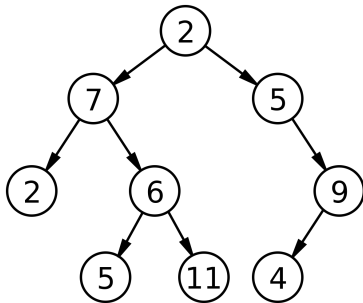
Spatio-Temporal



Gambar: Peta perjalanan seseorang yang direkam oleh Google Maps

Masalah pada Data

Dealing with Structures



Gambar: Data yang strukturnya berbentuk pohon

- 1 Atribut dapat berupa jalur dari akar ke daun
- 2 Contoh: {2-7-2-NA, 2-7-6-5, 2-7-6-11, ...}

Missing Values

- ① Tipe: tidak diketahui, tidak tersimpan, tidak relevan
- ② Penyebab: perubahan desain eksperimen, penggabungan dataset, dsb.
- ③ *Sangat mungkin terjadi!*

Missing Values - Solusi

- 1 Nominal: Gunakan label spesial, e.g. "NA"
- 2 Numerik: Diganti nilainya, e.g. rata-rata atau median atribut tersebut
- 3 Algoritma: Beberapa algoritma, e.g. Naïve Bayes dan *decision trees* dapat menyelesaikan kasus ini
- 4 Buang *instance*-nya

Inaccurate Values

- ① Kasus-kasus pencilan, kesalahan pengukuran, duplikat
- ② *Pahami datanya!*
- ③ Dapat dibuang dengan konsekuensi terhadap akurasi model

Imbalanced Data

- ① Kasus umum pada klasifikasi, e.g. diagnosis pasien
- ② Frekuensi salah satu kelas lebih banyak dibanding kelas lain
- ③ Mungkin perlu *metrics* selain akurasi
- ④ Ongkos kesalahan klasifikasi yang mungkin perlu dibuat tidak seimbang
- ⑤ Lihat [Kotsiantis, et al., 2006]!

Perbedaan Rentang

- Nilai dari atribut yang berbeda bisa ada pada rentang yang berbeda
- Dapat mempengaruhi kinerja algoritma pembelajaran mesin, e.g. SVM atau k-NN
- Perlu “diseragamkan”

Perbedaan Rentang - Solusi

Standardisasi

Mengubah nilai menjadi terdistribusi normal standar

$$x' = \frac{x - \mu}{\sigma}$$

Normalisasi

Mengubah nilai menjadi dalam rentang [0,1]

$$x' = \frac{x - \min}{\max - \min}$$

Atribut Nominal

- Beberapa algoritma pembelajaran mesin tidak bisa mengolah atribut nominal secara langsung, e.g. regresi linear, SVM, neural network
- Data kategorikal tidak selalu bersifat ordinal, e.g. merk ponsel, negara
- Solusi: **one-hot encoder**

One-hot Encoder

origin		origin_europe	origin_japan	origin_usa
usa		0	0	1
usa		0	0	1
usa	→	0	0	1
europe		1	0	0
usa		0	0	1
japan		0	1	0
japan		0	1	0

Referensi



Eliot Van Buskirk (7 Februari 2017)

Spotify, Accuweather Reveal How Weather Affects Music Listening

<https://insights.spotify.com/us/2017/02/07/spotify-accuweather-music-and-weather/>



Josh Bersin (19 Februari 2014)

The Myth Of The Bell Curve: Look For The Hyper-Performers

<https://www.forbes.com/sites/joshbersin/2014/02/19/the-myth-of-the-bell-curve-look-for-the-hyper-performers/>



Tim O'Shea (Juli 2016)

MNIST Generative Adversarial Model in Keras

<http://www.kdnuggets.com/2016/07/mnist-generative-adversarial-model-keras.html>



Alex Krizhevsky (2009)

Learning Multiple Layers of Features from Tiny Images

<https://www.cs.toronto.edu/~kriz/cifar.html>

Referensi



Steve Hardinger (diakses 27 Februari 2017)

Illustrated Glossary of Organic Chemistry

http://web.chem.ucla.edu/~harding/IGOC/B/benzene_ring.html



David Easley & Jon Kleinberg (2010)

Networks, crowds, and markets: Reasoning about a highly connected world

Cambridge University Press



Global Biodefense (25 Juni 2014)

USAMRIID Leads Effort on Viral Genome Sequencing Standards

<https://globalbiodefense.com/2014/06/25/>

[usamriid-leads-effort-viral-genome-sequencing-standards/](#)



Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas (2006)

Handling imbalanced datasets: A review

GESTS International Transactions on Computer Science and Engineering, 30(1), 25-36.

Terima kasih