

K-MEANS CLUSTERING

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

May 19, 2020

① CLUSTERING

② ALGORITMA

③ EVALUASI

BAHAN BACAAN

- ① VanderPlas, J. (2016). Python Data Science Handbook. (In Depth: k-Means Clustering) <http://nbviewer.jupyter.org/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.11-K-Means.ipynb>
- ② “Clustering dengan k-Means.” *Cerita Tentang Data*. 14 November 2017. <https://tentangdata.wordpress.com/2017/11/14/clustering-dengan-k-means/>

CLUSTERING

CLUSTERING

- *Unsupervised learning*
- Subpopulasi apa yang ada dalam data?
- Apa kesamaan dari elemen di tiap subpopulasi?
- Bisa digunakan untuk menemukan pencilan

JENIS-JENIS CLUSTERING

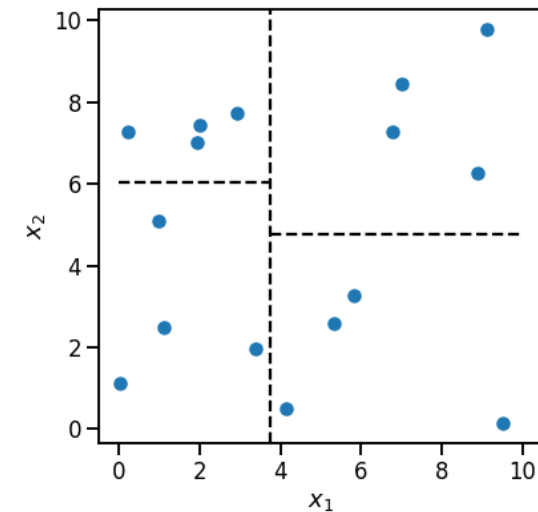
- Tujuan:
 - Monothetic: *common property*
 - Polythetic: kemiripan data dengan pengukuran jarak
- Irisan:
 - Hard clustering
 - Soft clustering
- Flat vs hierarchical

METODE CLUSTERING

Metode *clustering* yang akan dibahas dalam kuliah ini:

- K-D Trees: *monothetic, hard boundaries, hierarchical*
- k-Means: *polythetic, hard boundaries, flat*
- Gaussian mixtures (EM algorithm): *polythetic, soft boundaries, flat*
- Agglomerative clustering: *polythetic, hard boundaries, hierarchical*

CLUSTERING DENGAN K-D TREES

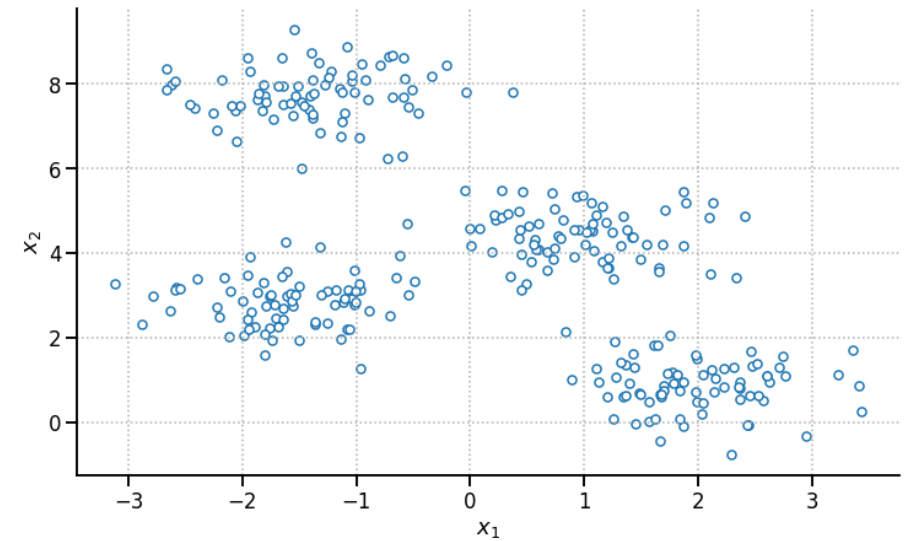


GAMBAR: Monothetic, hard boundaries, hierarchical

K-MEANS

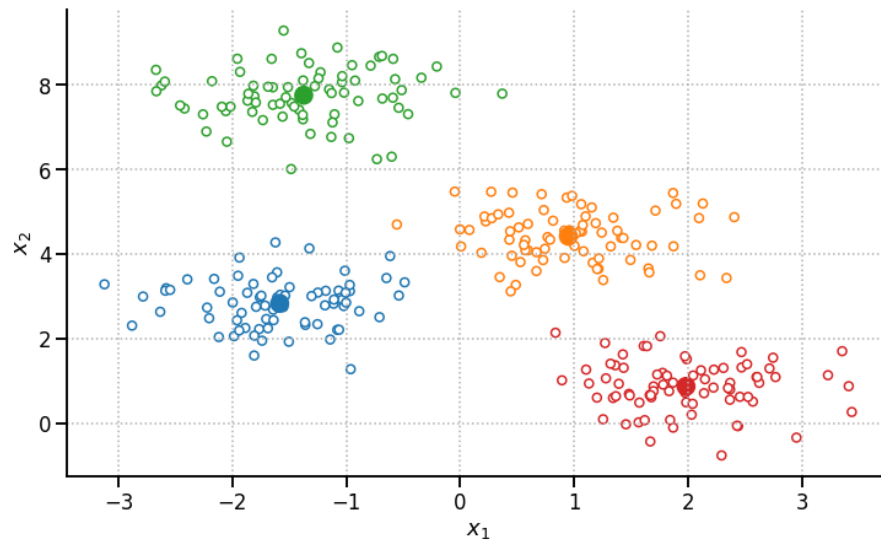
- Jumlah k ditentukan dari awal
- Tidak memerlukan label
- Menggunakan *centroid*, i.e. rata-rata nilai dari objek yang masuk dalam *cluster* tersebut
- Mencari *centroid* terdekat dari tiap objek

CONTOH DATA



GAMBAR: Contoh data dalam 2D

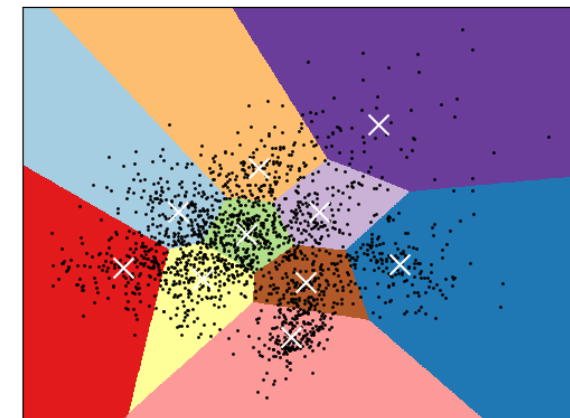
HASIL K-MEANS



GAMBAR: Setelah algoritma k-Means dijalankan

HASIL PCA + K-MEANS

K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross



GAMBAR: PCA + k-Means pada MNIST (Pedregosa et al., 2011)

ALGORITMA: EXPECTATION-MAXIMIZATION

ALGORITMA

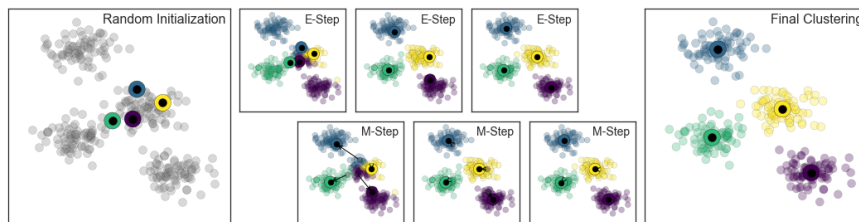
- ① Inisialisasi k *centroid* secara acak
- ② Ulangi hingga konvergen
 - A. E-step: Masukkan tiap titik/objek ke *centroid* terdekat

$$\arg \min_j D(x_i, c_j)$$

- B. M-step: Ubah nilai *centroid* menjadi rata-rata dari tiap titik/objek

$$c_j(a) = \frac{1}{n_j} \sum_{x_i \rightarrow c_j} x_i(a), \text{ for } a = 1..d$$

VISUALISASI EM



GAMBAR: Konvergensi kluster tercapai hanya dalam tiga iterasi [VanderPlas, 2016]

Perhatikan bahwa algoritma ini sangat bergantung pada inisialisasi *centroid*!

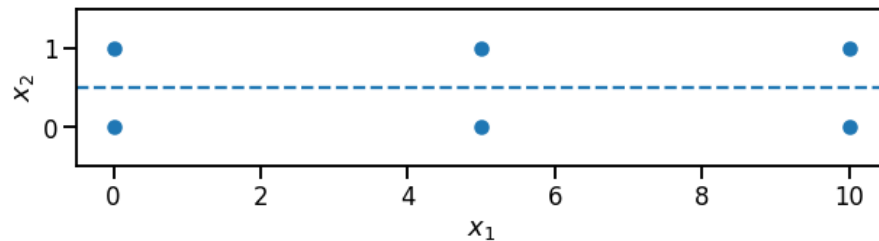
PROPERTI DARI K-MEANS

- Meminimalkan jarak agregat intra-kluster

$$V = \sum_j \sum_{x_i \rightarrow c_j} D(c_j, x_i)^2$$

- Konvergensi ke **minimum lokal**
- Poin yang berdekatan mungkin masuk ke kluster yang berbeda

Berapa nilai k yang optimal?



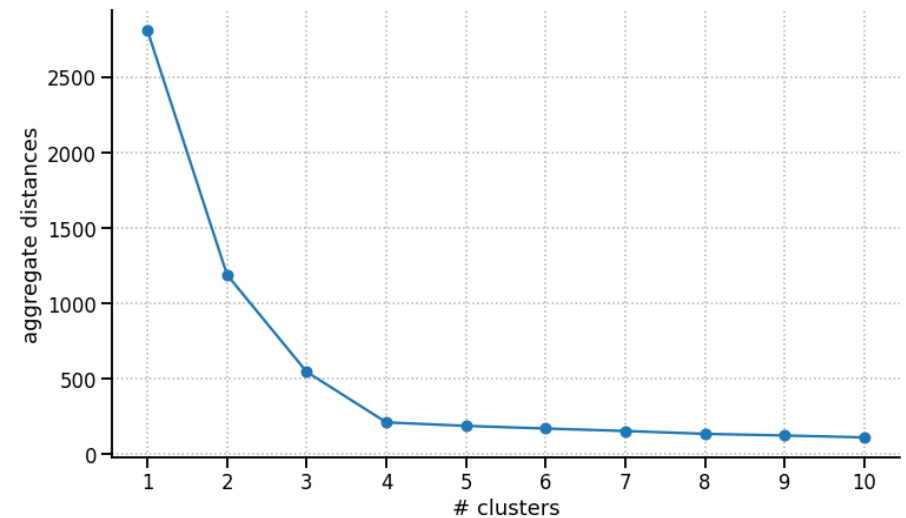
MENENTUKAN NILAI K

- Gunakan label kelas, e.g. 10 untuk MNIST
- Gunakan V untuk menggambarkan *scree plot*

$$V = \sum_j \sum_{x_i \rightarrow c_j} D(c_j, x_i)^2$$

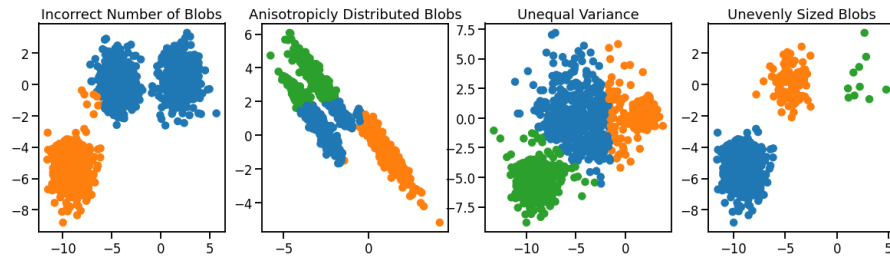
lalu gunakan *elbow method*, i.e. nilainya dapat dicari dengan menggunakan nilai optimal turunan kedua

SCREE PLOT



GAMBAR: Secara visual, scree plot menunjukkan nilai optimal $k = 4$

ASUMSI K-MEANS



GAMBAR: Tiga gambar pertama menunjukkan klaster yang tidak intuitif karena asumsi implisit k-Means (Roth, 2017)

EVALUASI

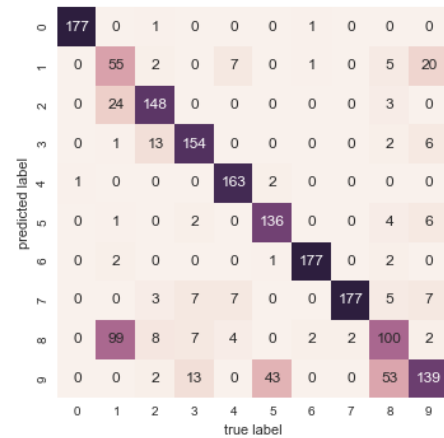
EVALUASI KLASTER

- Ekstrinsik: untuk mengerjakan tugas lain
 - Representasi gambar dengan fitur berupa klaster
 - Menemukan pencilan
- Intrinsik: untuk diri sendiri
 - Memahami data – deskriptif
 - Klaster \sim kelas, e.g. MNIST \rightarrow 10 klaster
 - Perbandingan pasangan data dari klaster oleh manusia

EVALUASI INTRINSIK: KLASTER \sim KELAS

- Klaster c_1, c_2, \dots, c_K
- Kelas R_1, R_2, \dots, R_N
- Cocokkan R_i dengan c_j , hitung akurasi atau F1
 - Bagaimana jika $N \neq K$?
 - Ada banyak cara, paling mudah dengan pendekatan *greedy*

CONTOH EVALUASI INTRINSIK



GAMBAR: Confusion matrix dari MNIST clustering [VanderPlas, 2016]

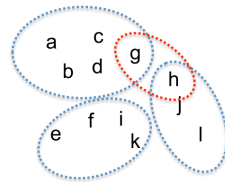
CONTOH EVALUASI INTRINSIK

	G1	G2	G3	G4	G5	G6
C1	1	7	0	1	4	0
C2	0	0	0	0	2	7
C3	0	0	2	0	0	0
C4	3	1	0	0	1	0

GAMBAR: Kluster karakter dalam Julius Caesar

EVALUASI INTRINSIK: PERBANDINGAN ANTARPASANGAN

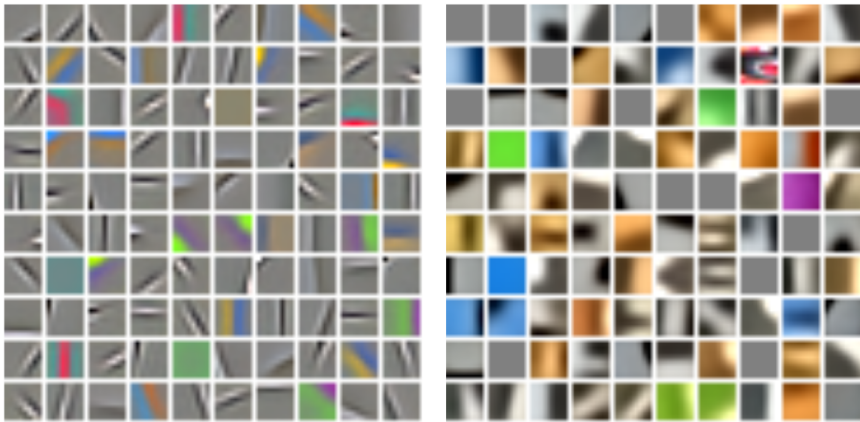
- Pasangan x_i, x_j apakah seharusnya berada dalam kluster yang sama?
- Hitung error, akurasi, F1
 - FN: pasangan x_i, x_j yang harusnya cocok, tapi berada dalam kluster yang lain (e,h)
 - FP: pasangan x_i, x_j yang harus tidak cocok, tapi berada dalam kluster yang sama (c,d)



APLIKASI CLUSTERING

- Pemelajaran fitur [Coates, 2012]
- Kompresi gambar [VanderPlas, 2016]
- Sistem rekomendasi

APLIKASI: PEMELAJARAN FITUR



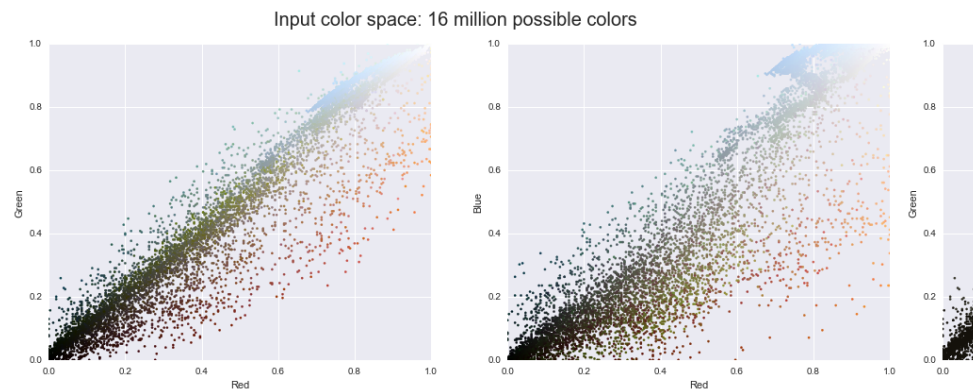
GAMBAR: Centroids dari CIFAR-10 dengan dan tanpa pemutihan [Coates, 2012]

APLIKASI: KOMPRESI GAMBAR



GAMBAR: Gambar yang akan dikompresi dengan *clustering* [VanderPlas, 2016]

KLASTER WARNA



GAMBAR: *Clustering* warna dengan kompresi [VanderPlas, 2016]

HASIL KOMPRESI GAMBAR



GAMBAR: Kompresi dengan faktor hingga 1 juta dengan *clustering* [VanderPlas, 2016]

IKHTISAR

- 1 Jenis metode clustering dibagi berdasarkan tiga atribut: tujuan, irisan, dan struktur
- 2 k-Means bekerja dengan algoritma EM untuk menghasilkan klaster yang diwakili oleh *centroid*
- 3 k-Means sangat tergantung dengan inisialisasi *centroid*
- 4 Ada asumsi implisit yang dibuat oleh k-Means
- 5 Klaster yang dibentuk dapat dievaluasi secara ekstrinsik dan intrinsik

Salindia ini dibuat dengan
sangat dipengaruhi oleh Lavrenko (2014)

REFERENSI



Jake VanderPlas (2016)

In Depth: k-Means Clustering

[http://nbviewer.jupyter.org/github/jakevdp/
PythonDataScienceHandbook/blob/master/notebooks/05.
11-K-Means.ipynb](http://nbviewer.jupyter.org/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.11-K-Means.ipynb)



Adam Coates & Andrew Y. Ng (2012)

Learning feature representations with k-means.

Neural networks: Tricks of the trade (pp. 561-580). Springer Berlin Heidelberg.

Terima kasih