Home Concept

Dasar Clustering Analysis

Professional Academy **May, 30 2022**

Materi ini disiapkan untuk Professional Academy yang diadakan Kominfo bekerjasama dengan DQlab.

Explore More

Recap

Theme
Geometries
Aesthetics
Data

Intermezzo

Berikut adalah pernyataan yang benar mengenai default

JAWABAN

- Data dan aesthetic mapping pada plot adalah default untuk seluruh layer
- Data dan aesthetic mapping pada tiap layer akan override default pada plot
- Data dan aesthetic mapping pada tiap layer tidak bisa diberikan jika terdapat data dan aesthetic pada plot
- Data dan aesthetic bisa tidak dimasukkan pada plot object => need to test
- Semua benar

ukraina

Sign in yahoo!

Semua Gambar Video Berita Kapan saja

Tentang hasil pencarian 8.780.000

id.wikipedia.org › wiki › Ukraina

Ukraina - Wikipedia bahasa Indonesia, ensiklopedia bebas

Ukraina (bahasa **Ukraina**: Україна, bahasa Rusia: Украина) adalah sebuah negara di Eropa Timur yang berbatasan dengan Rusia di timur dan timur-laut; Belarus di barat-laut; Polandia dan Slowaki...

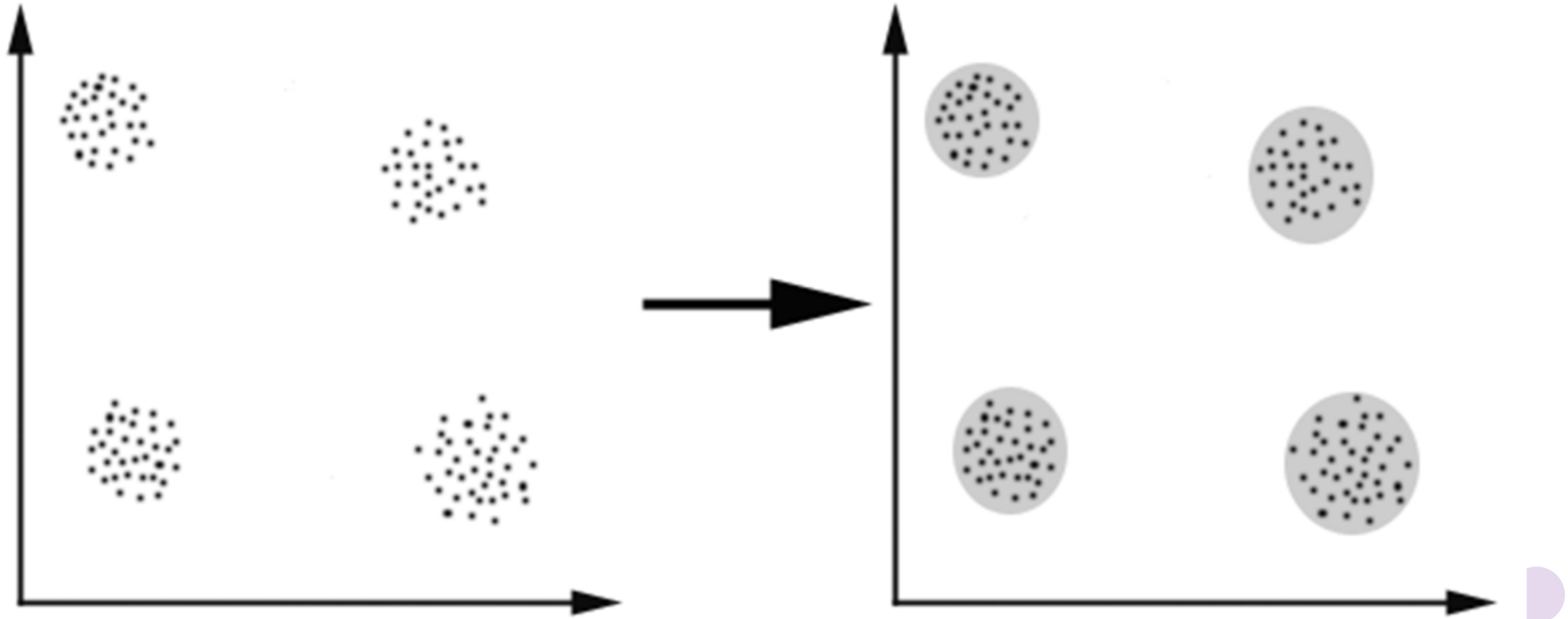
Gambar

Lihat semua

en.wikipedia.org › wiki › Ukraine

Ukraine - Wikipedia

Ukraine (Ukrainian: Україна, romanized : Ukraïna, pronounced [ʊkrɐˈjinɐ] (listen)) is a country in Eastern Europe. It is the second largest country in Europe after Russia, [12] which borders it to the...



Theory

Cluster: kumpulan objek data

Anggota cluster yang sama memiliki kemiripan satu sama lain, tetapi berbeda dengan anggota cluster lain.

Cluster analysis

Menemukan kemiripan data berdasarkan karakteristik dan mengelompokan data yang mirip ke dalam cluster.

Unsupervised learning:

class tidak ditentukan sebelumnya

Penggunaan

1. Tool untuk melihat distribusi data
2. Preprocessing untuk langkah berikutnya

APLIKASI

Clustering

- Pengenalan Pola
- Spatial Data Analysis :Cluster spatial
- Pemrosesan gambar
- Marketing: Membantu pihak pemasaran untuk menentukan grup khusus dan membuat program khusus untuk grup ini.
- Land use: Identifikasi area yang digunakan untuk hal yang sama.
- Asuransi: Identifikasi grup yang memiliki tingkat claim yang tinggi.
- Tata kota: Identifikasi rumah-rumah berdasrkan tipe, harga dan lokasi.

Ciri Good Cluster

Metode yang bagus akan menghasilkan:

- intra-class similarity yang tinggi (anggota di dalam kelas yang sama mirip)
- low inter-class similarity (anggota di kelas yang lain, jauh berbeda)

Kualitas cluster bergantung kepada ukuran kemiripan yang digunakan oleh metode clustering.

Kualitas juga ditentukan sejauh mana clustering dapat menemukan pola tersembunyi.

Ukuran Kesamaan

1. Kesamaan/kemiripan diukur berdasarkan fungsi jarak, d(i, j)
2. Definisi distance functions bisanya sangat berbeda untuk interval-scaled, boolean, categorical, ordinal ratio, and vector variables.
3. Bobot diasosiasikan dengan aplikasi dan arti data.
4. Sulit untuk mendefinsikan “cukup sama ” or “cukup bagus” karena subyektif.

Pendekatan Clustering

1. Partisi : Buat partisi dan evaluasi berdasarkan kriteria tertentu, misalnya meminimalkan sum of square errors. Metode: k-means, k-medoids, CLARANS
2. Hirarkis: Buat struktur hierarchical menggunakan kriteria tertentu. Metode: Diana, Agnes, BIRCH, ROCK, CAMELEON
3. Density-based : Berdasarkan connectivity dan density functions. Metode: DBSACN, OPTICS, DenClue
4. Yang lain: Grid-based approach, model-based, frequent pattern-based, user-guided or constraint-based:

K-Means

1. Partisi objek ke k nonempty subset
2. Hitung centroid (centroid adalah titik tengah cluster)
3. Masukkan setiap objek ke cluster dengan centroid terdekat
4. Kembali ke langkah 2, sampai tidak ada posisi yang berubah

Kelemahan: K-Means

1. Bila jumlah data tidak terlalu banyak, mudah untuk menentukan cluster awal
2. Jumlah cluster, sebanyak K, harus ditentukan sebelum dilakukan perhitungan
3. Tidak pernah mengetahui real cluster dengan menggunakan data yang sama, namun jika dengan cara yang berbeda mungkin dapat memproduksi cluster yang berbeda jika jumlah datanya sedikit
4. Tidak tahu kontribusi dari atribut dalam proses pengelompokan karena dianggap memiliki bobot yang sama

Contoh Case K-Means

Anda diminta mencluster 8 point berikut: A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9).
Gunakan K-Means dengan euclidean distance.

Asumsikan A2, B2 dan C2 sebagai inisial cluster untuk cluster A, B dan C. Tampilkan perhitungan dan isi cluster (termasuk centroid cluster yang dihitung dengan rata-rata).

Contoh Case K-Means

Diketahui:
A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9).

Jarak antara setiap titik dengan setiap cluster.

Cluster A, centroid: (2,5)

Cluster B, centroid: (7,5)

Cluster C, centroid: (4,9)

A1 --> cluster A

$$d(A1,A)=\sqrt{(|2-2|^2+|10-5|^2}$$
$$d(A1,A)=5$$

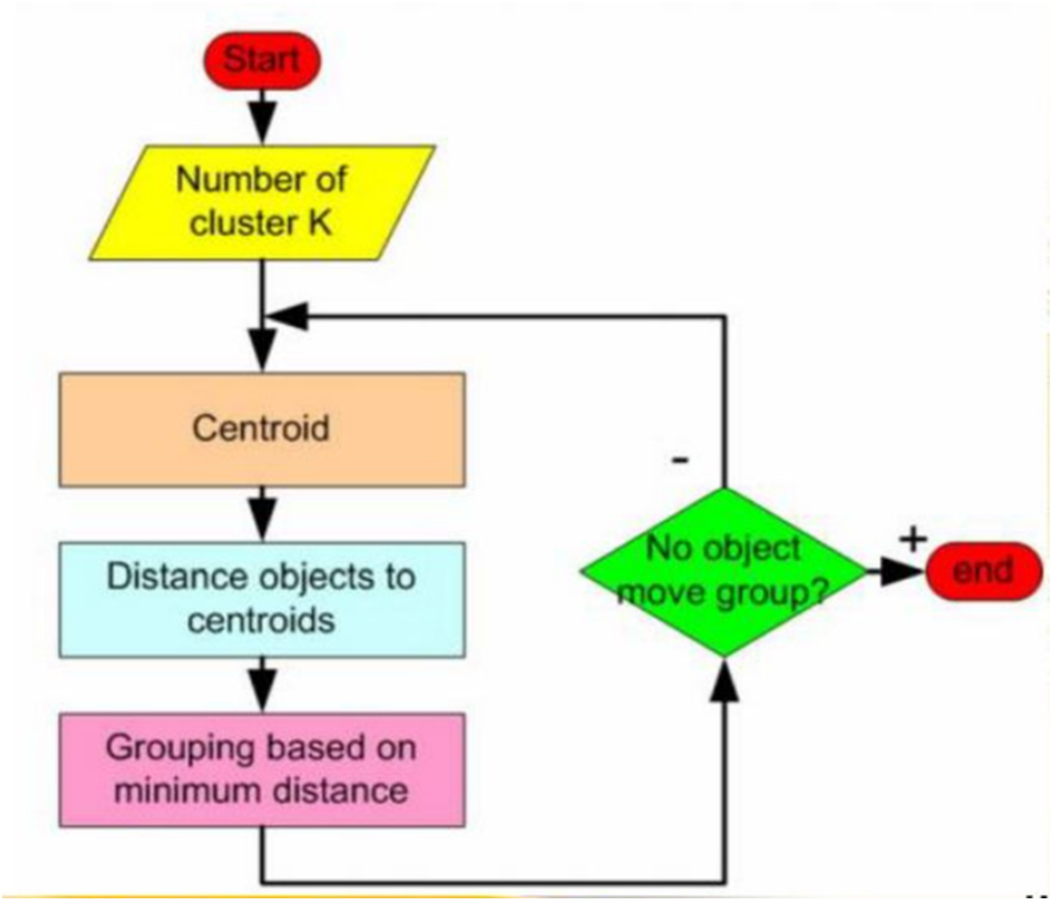
A3 --> cluster A, d(A3,A) =

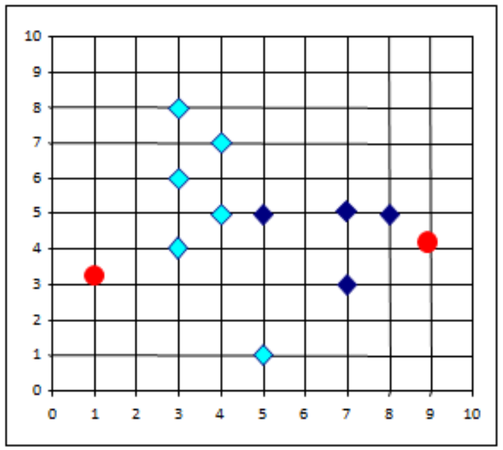
B1 --> cluster A, d(B1,A) =

B3 --> cluster A, d(B3,A) =

C1 --> cluster A, d(C1,A) =

Recap

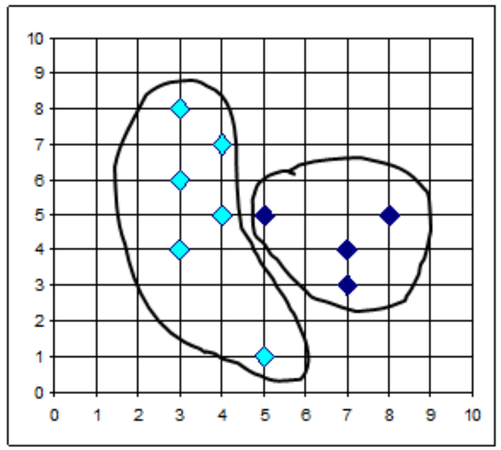

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end



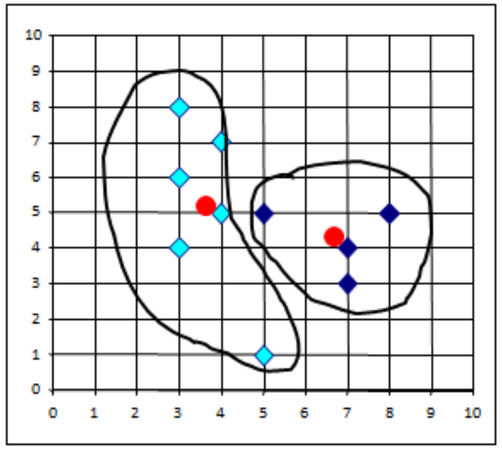
K=2

Arbitrarily choose K object as initial cluster center

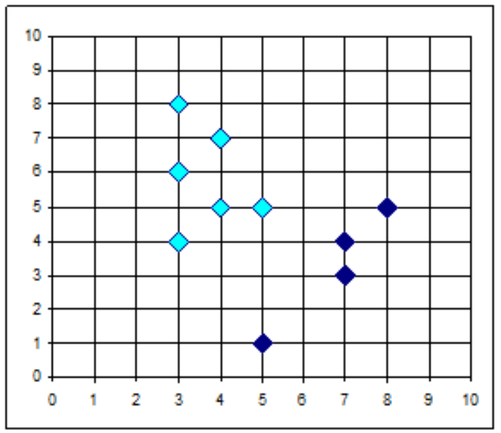
Assign each objects to most similar center



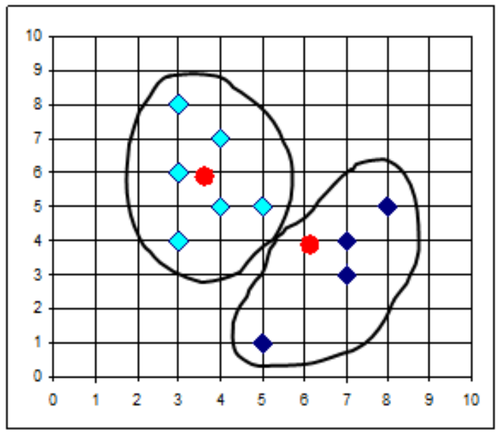
Update the cluster means



reassign



Update the cluster means



reassign

Praktik

Best Practice

Selalulah berlatih sampai terbiasa dan tetaplah semangat untuk belajar hal baru