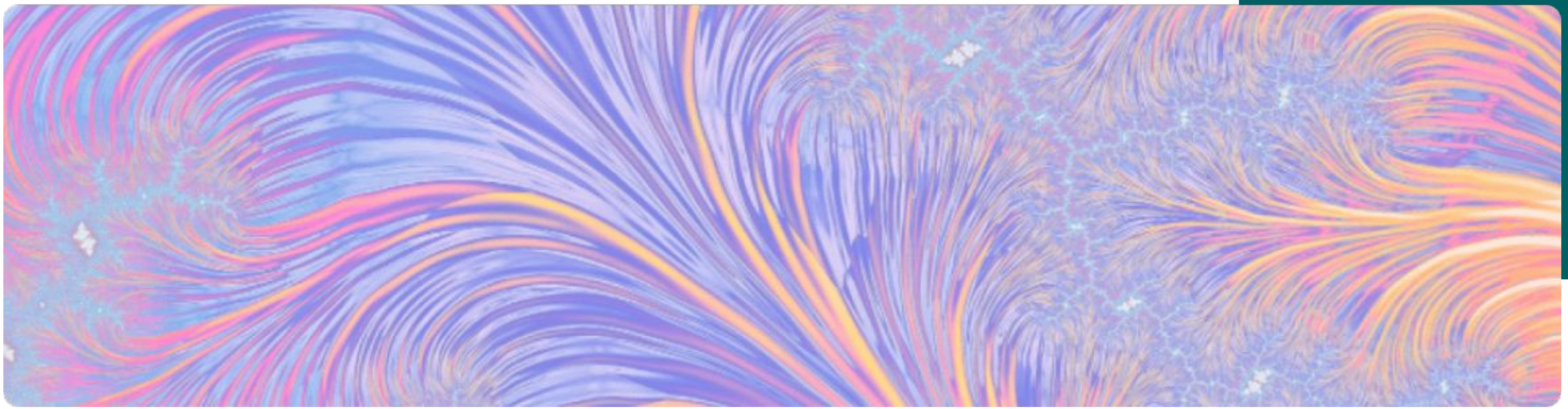
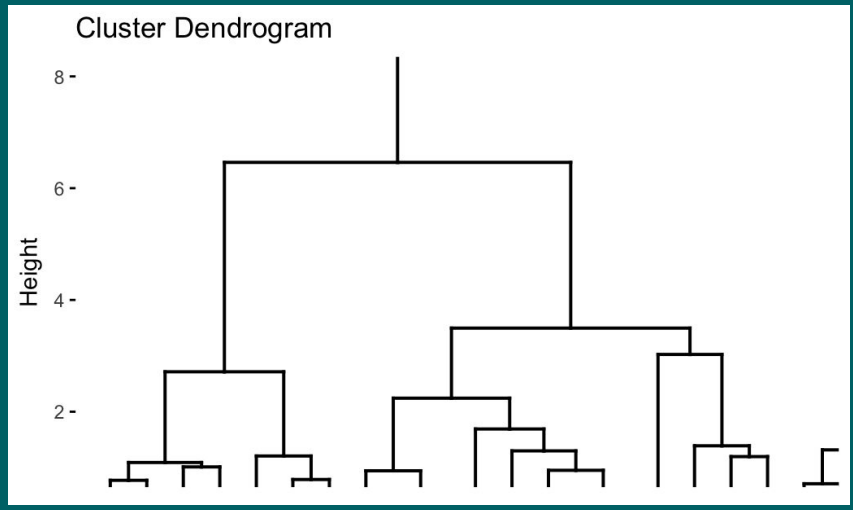


Hierarchical Clustering

Memahami Algoritma, Cost Function, dan Hipotesis Function secara Mendalam

Pengantar Hierarchical Clustering



Definisi Konseptual

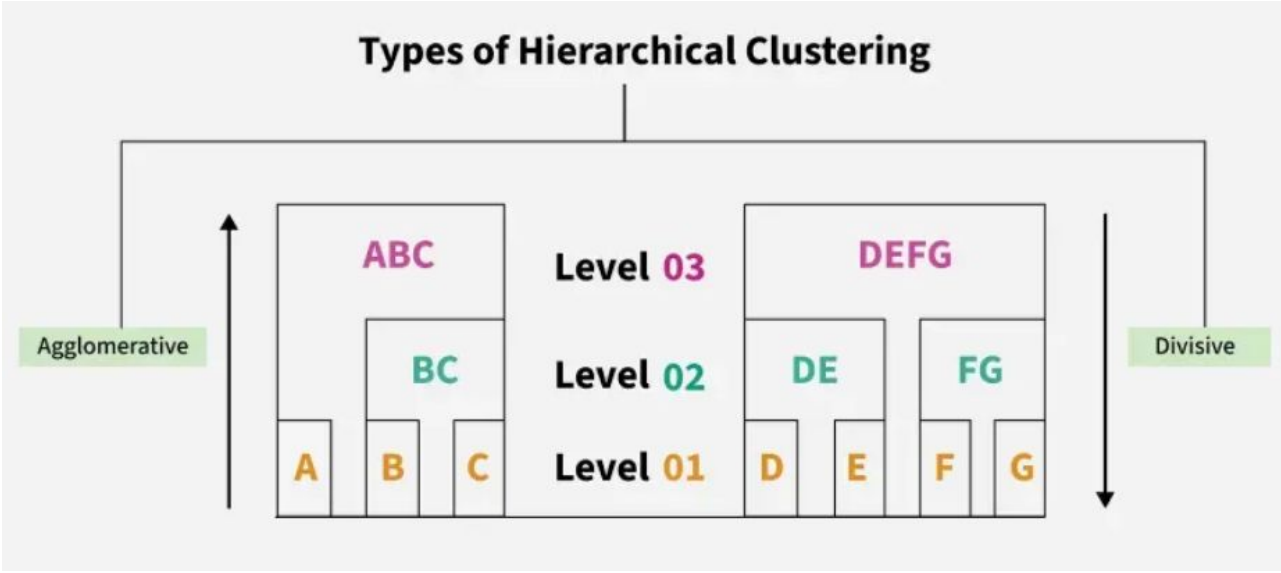
Hierarchical Clustering adalah algoritma unsupervised learning yang digunakan untuk mengelompokkan titik-titik data serupa ke dalam klaster. Algoritma ini membangun hierarki klaster bertingkat dengan menggabungkan klaster yang lebih kecil menjadi klaster yang lebih besar (aglomeratif) atau membagi klaster yang besar menjadi klaster yang lebih kecil (divisif). Hal ini menghasilkan struktur seperti pohon yang dikenal sebagai dendrogram.

Output Dendrogram

Hasil analisis ini divisualisasikan dalam bentuk struktur pohon yang disebut **Dendrogram**. Dendrogram bagaikan pohon keluarga untuk klaster. Dendrogram menunjukkan bagaimana titik data individual atau kelompok data bergabung. Bagian bawah menunjukkan setiap titik data sebagai kelompoknya sendiri, dan semakin ke atas, kelompok-kelompok yang serupa akan digabungkan. Semakin rendah titik penggabungan, semakin mirip kelompok-kelompok tersebut. Dendrogram membantu kita melihat bagaimana berbagai hal dikelompokkan selangkah demi selangkah.

Jenis-jenis Pengelompokan Hierarkis

- **Agglomerative (Bottom-Up):** Ini adalah pendekatan yang paling umum digunakan. Algoritma memulai proses dengan menganggap setiap data sebagai cluster terpisah, kemudian secara iteratif menggabungkan pasangan cluster terdekat hingga hanya tersisa satu cluster tunggal.
- **Divisive (Top-Down):** Pendekatan ini bekerja sebaliknya. Dimulai dengan satu cluster raksasa yang berisi seluruh dataset, algoritma kemudian memecah cluster tersebut secara rekursif menjadi bagian-bagian yang lebih kecil dan spesifik hingga tersisa cluster individu.



Tahapan Algoritma Agglomerative

1

Langkah Inisialisasi

Langkah pertama adalah menetapkan setiap titik data (N) sebagai cluster tersendiri. Kemudian, menghitung matriks jarak (distance matrix) yang mencatat jarak antara setiap pasangan cluster yang ada.

2

Proses Penggabungan

Algoritma mencari nilai minimum dalam matriks jarak untuk menemukan dua cluster yang paling mirip. Kedua cluster ini kemudian digabungkan (merge) menjadi satu cluster baru yang lebih besar.

3

Pembaruan Iteratif

Setelah penggabungan, matriks jarak diperbarui untuk mencerminkan jarak cluster baru terhadap cluster lainnya. Proses ini diulang terus-menerus hingga semua data tergabung dalam satu cluster.

Metrik Dasar & Contoh Perhitungan

1. Euclidean Distance

Metode ini adalah cara paling intuitif untuk mengukur jarak garis lurus ("as the crow flies") antara dua titik. Berasal dari prinsip Teorema Pythagoras, Euclidean sangat standar untuk data geometri kontinu, namun sensitif terhadap perbedaan skala

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Contoh Titik A(1, 2) dan B(4, 6):

$d = \sqrt{((4-1)^2 + (6-2)^2)}$
$d = \sqrt{(3^2 + 4^2)} = \sqrt{(9+16)}$
$d = \sqrt{25} =$
5

2. Manhattan Distance

Dikenal juga sebagai "Taxicab Geometry", metrik ini mengukur jarak dengan menyusuri sumbu grid (siku-siku), seperti taksi di blok kota yang tidak bisa menembus gedung. Ini dihitung dari jumlah selisih

$$d = |x_2 - x_1| + |y_2 - y_1|$$

Contoh Titik A(1, 2) dan B(4, 6):

$d = |4 - 1| + |6 - 2|$
$d = |3| + |4|$
$d = 3 + 4 =$
7

Metrik Lanjutan (Advanced Metrics)

3. Minkowski Distance

Minkowski adalah bentuk generalisasi matematis yang memayungi Euclidean dan Manhattan. Dengan mengubah parameter 'p', kita bisa mengatur sensitivitas jarak. Jika p=1, ia menjadi Manhattan; jika p=2, ia menjadi Euclidean. Ini memberikan fleksibilitas tinggi.

$$d = \sum | x_i - y_i |^{p\,1/p}$$

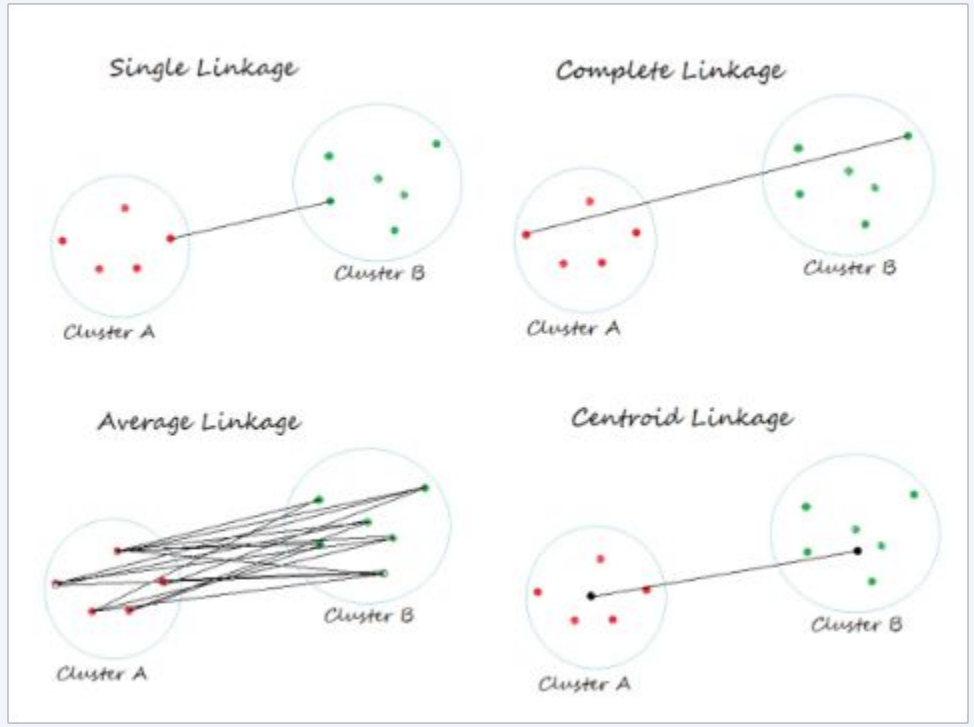
4. Cosine Similarity

Cosine Similarity mengukur kosinus sudut antara dua vektor bukan nol. Metrik ini ideal untuk data teks atau ketika besaran tidak menjadi pertimbangan (misalnya, vektor frekuensi kata).

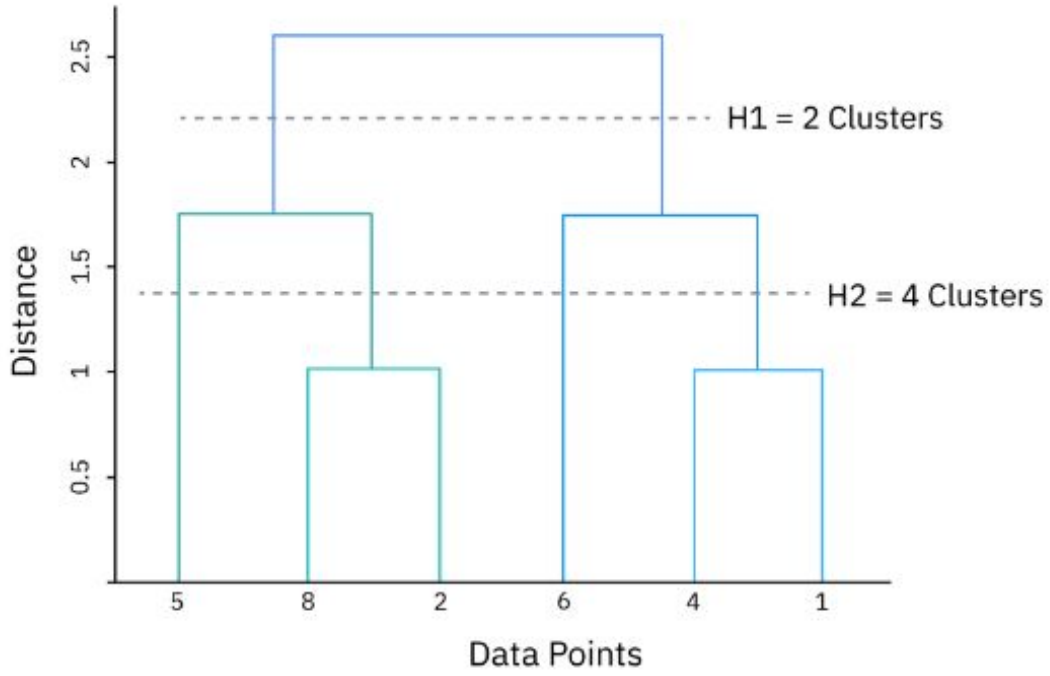
$$\text{Similarity} = \frac{A \dot{c} B}{\| A \| \quad \| B \|}$$

Linkage Criteria (Kriteria Jarak)

- **Single Linkage:** Mengukur jarak terdekat antar titik di dua cluster. Metode ini cenderung menghasilkan cluster yang memanjang (chaining effect) dan sangat sensitif terhadap noise.
- **Complete Linkage:** Mengukur jarak terjauh antar anggotanya. Ini menghasilkan cluster yang lebih kompak dan bulat, namun bisa memecah cluster besar.
- **Average Linkage:** Menghitung rata-rata jarak antara semua pasangan titik di dua cluster, memberikan kompromi yang seimbang.
- **Ward's Method:** Berfokus meminimalkan varians internal cluster.



Hipotesis Function: The Cut



Berbeda dengan supervised learning di mana hipotesis adalah prediksi output, di sini hipotesis adalah **aturan keputusan (decision rule)** tentang bagaimana data dikelompokkan.

Logika Pemotongan

Menarik sebuah garis horizontal imajiner pada Dendrogram yang berfungsi sebagai "pisau pemotong".

- Jika kita memotong di bagian atas, kita mendapatkan sedikit kelompok yang bersifat umum (General).
- Jika kita memotong di bagian bawah, kita mendapatkan banyak kelompok yang sangat spesifik (Detail).

Jadi, fungsi hipotesisnya adalah: "Di mana kita memutuskan untuk berhenti mengelompokkan data?"

Cost Function: Metode Ward

Dalam Hierarchical Clustering, konsep "Cost Function" paling baik dijelaskan melalui **Metode Ward**. Tujuannya bukan sekadar jarak, melainkan meminimalkan total varians intra-cluster.

Pada setiap langkah penggabungan, algoritma menghitung "biaya" penggabungan, yaitu seberapa besar kenaikan *Error Sum of Squares* (ESS) yang akan terjadi jika dua cluster digabungkan. Ward method memilih pasangan yang memberikan kenaikan ESS terkecil.

Ini menjamin bahwa cluster yang terbentuk tetap sekompak mungkin dan varians data di dalamnya tetap rendah.

$$\Delta(A,B) = \frac{|A||B|}{|A|+|B|} \|\mu_A - \mu_B\|^2$$

$$\Delta = \frac{10 \cdot 20}{10+20} \cdot 5 = \frac{200}{30} \cdot 5 = 33.33$$

μA = centroid kluster A

μB = centroid kluster B

μA−μB2 = jarak kuadrat antar centroid

Cara Membaca Dendrogram

Sumbu Y (Tinggi/Jarak)

Sumbu vertikal ini sangat krusial karena merepresentasikan jarak atau tingkat ketidakmiripan (dissimilarity) antar cluster saat mereka digabungkan. Semakin tinggi posisi garis horizontal penghubung, semakin besar perbedaan antara cluster yang digabungkan tersebut.

Garis & Cluster

Garis-garis vertikal mewakili anggota cluster, dan palang horizontal mewakili peristiwa penggabungan. Panjang garis vertikal sebelum bertemu palang horizontal menunjukkan seberapa "terpisah" atau stabil cluster tersebut dari cluster lainnya dalam hierarki.

Contoh Aplikasi di Dunia Nyata

Biologi & Taksonomi

Digunakan secara luas dalam biologi evolusioner untuk membangun pohon filogenetik, yang memetakan hubungan kekerabatan antar spesies berdasarkan kemiripan urutan DNA atau karakteristik fisik.

Analisis Pasar

Pemasar menggunakannya untuk segmentasi pelanggan yang mendalam, mengidentifikasi kelompok perilaku konsumen dari yang sangat spesifik (niche) hingga kategori belanja yang lebih umum.

Organisasi Dokumen

Dalam pemrosesan bahasa alami (NLP), teknik ini membantu mengelompokkan ribuan artikel berita atau dokumen hukum ke dalam hierarki topik utama dan sub-topik secara otomatis.

Studi Kasus

Pengelompokan Pelanggan Mall

Studi kasus ini bertujuan untuk mengelompokkan pengunjung mall berdasarkan Pendapatan tahunan dan pengeluaran. Feature nya adalah Annual Income dan Spending score.

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
47	Female	21	99.8	73.6
42	Male	35	93.6	60.5
17	Male	41	17.5	92.3
9	Male	30	15	51.3
39	Male	26	77.3	5.3
43	Male	36	79.2	88.1
38	Male	25	96.7	39.3
37	Male	72	95.4	6.2
5	Female	62	30.2	8.8
20	Female	52	40.2	87.8
46	Male	25	98.3	84
14	Female	55	27.6	82.4
27	Female	19	65.2	69.2
1	Female	73	46.7	26.3
32	Male	66	95.2	25.8
36	Male	64	97.9	11.6
3	Male	57	32.3	22.3
13	Male	25	30.1	68.1
2	Male	25	30	12.9
41	Male	22	96.5	87
25	Female	21	49.4	22.3
24	Female	70	86.3	35.3
16	Female	70	19	80
23	Male	27	60.8	53.7
33	Female	38	95.9	8.5
22	Female	56	64.2	35.2
40	Female	34	85.5	36.9
7	Male	70	18.2	22.3

Kesimpulan Akhir

Hierarchical Clustering menawarkan keseimbangan unik antara wawasan visual yang mendalam dan analisis struktur data yang kompleks. Meskipun memiliki tantangan dalam hal efisiensi komputasi pada data besar, kemampuannya untuk mengungkap hubungan hierarkis tanpa asumsi awal menjadikannya alat yang tak ternilai dalam tahap eksplorasi data (Exploratory Data Analysis).

Sesi Tanya Jawab