IMPLEMENTASI VEKTOR & MATRIKS

TK13023
COMPUTATION II

KELAS B DAN C

DOSEN: LELY HIRYANTO

UNTAR
Universitas Tarumanagara
UNTAR untuk INDONESIA

CLASSIFICATION AND CLUSTERING

Implementasi Vektor

UNTAR
Universitas Tarumanagara
UNTAR untuk INDONESIA

Konsep Classification vs Clustering

Classification (Supervised Learning)

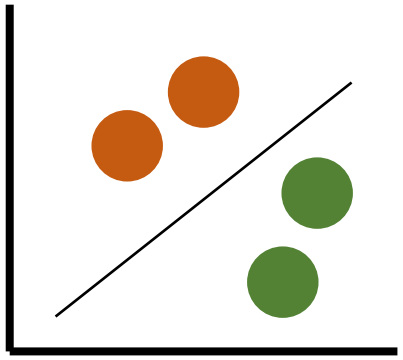
- Penentuan jumlah dan nama kategori (kelompok/kelas/label) dari data sudah ditentukan
- Data dibagi menjadi dua:
  - Data latih: diketahui labelnya
  - Data uji: belum diketahui labelnya
- Algoritma (Matriks & Vektor):
  - K-Nearest Neighbors
  - Support Vector Machines
- Algoritma lainnya:
  - Naïve Bayes (statistik)
  - Decision Trees (ID3, C4.5, CART)

Clustering (Unsupervised Learning)

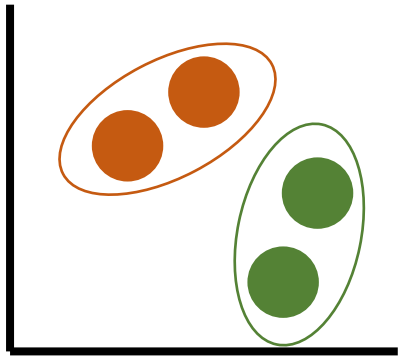
- Jumlah kelompok/kelas/label tidak diketahui
- Tidak ada konsep data pelatihan dan pengujian
- Algoritma (Matriks dan Vektor)
  - K-means
  - Agglomerative Hierarchical Clustering
- Algoritma lain:
  - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Mean-Shift
  - Gaussian Mixture Model (GMM)

ICAEW CHARTERED ACCOUNTANTS

Ilustrasi Classification vs Clustering



Classification



Clustering

https://www.educative.io/answers/classification-vs-clustering

UNTAR
Universitas Tarumanagara
UNTAR untuk INDONESIA

Data dan Atributnya

- Data:
  - Data Tabel: laporan
  - Teks: dokumen, postingan media sosial, log files, email
  - Visual: foto, video
  - Audio: musik
- Atribut atau variabel bebas:
  - karakteristik dari data
  - Satu buah data bisa memiliki lebih dari satu karakteristik
    - Membentuk matriks atribut

Classification

K-Nearest Neighbors

UNTAR
Universitas Tarumanagara
UNTAR untuk INDONESIA

Kategori

- Contoh nama kategori:
  - Biner: “spam” atau “no spam”
  - Topik: “programming”, “law”, atau “finance”
  - Opini: “like”, “dislike”, atau “neutral”

K-Nearest Neighbors (KNN)

Diketahui $\boldsymbol{n}$ data latih $\boldsymbol{X} = \{\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{n-1}}\}$ yang sudah memiliki label dan sebuah data uji $\mathbf{y}$:

1. Menentukan satu rumus perhitungan nilai kemiripan (metrik jarak)
   - Euclidean distance (cek materi vektor: norm), atau
   - Cosine Similarity (cek materi vektor: norm dan dot product)
2. Menentukan nilai K
3. Menghitung kemiripan $\mathbf{y}$ dengan setiap data latih di $\boldsymbol{X} = \{\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_{n-1}}\}$
4. Urutkan menaik nilai kemiripan dari $\mathbf{y}$ dengan setiap data latih di $\boldsymbol{X}$
5. Pilih K data latih pertama
6. Kategorikan data uji $\mathbf{y}$ menggunakan mayoritas label dari K data latih pertama tersebut.

UNTAR
Universitas Tarumanagara
UNTAR untuk INDONESIA

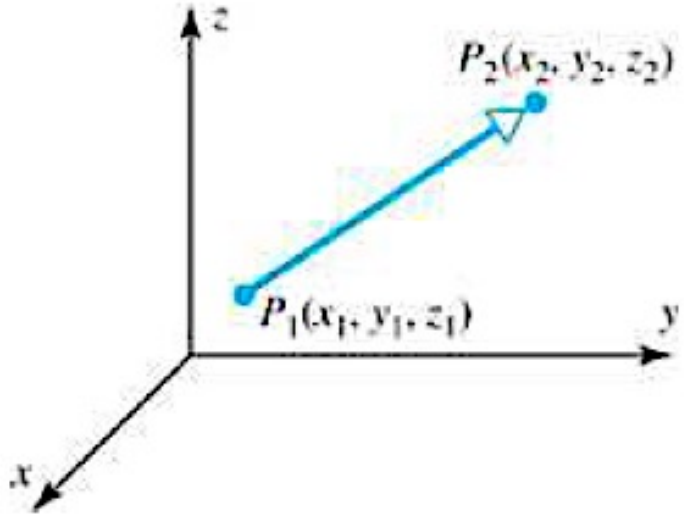
Ingat: Norm – Jarak dari Sebuah Vektor!

- Jarak dari sebuah vektor jika diketahui titik awal dan akhir dari vektor $\overrightarrow{P_1P_2}$ di ruang R^3 yaitu $P_1(x_1, y_1, z_1)$ dan $P_2(x_2, y_2, z_2)$:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

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

- Contoh: Jarak d antara titik $P_1(2, -1, -5)$ dan $P_2(4, -3, 1)$
  - $d = \sqrt{(4-2)^2 + (-3+1)^2 + (1+5)^2} = \sqrt{44} = 2\sqrt{11}$
- Disebut **Eucliean Distance!**



UNTAR Universitas Tarumanagara
UNTAR untuk INDONESIA

Contoh KNN: Data

- Diketahui data pengajuan pinjaman dari 5 nasabah yang disetujui dan 5 nasabah yang tidak disetujui.
- K = 5

Data Latih X

ID Nasabah	Umur	Pinjaman (juta)	Keputusan
1	25	40	Tolak
2	35	60	Tolak
3	45	80	Tolak
4	20	20	Tolak
5	35	120	Tolak
6	52	18	Setujui
7	23	95	Setujui
8	40	62	Setujui
9	60	100	Setujui
10	48	220	Setujui
11	33	150	?

UNTAR Universitas Tarumanagara
UNTAR untuk INDONESIA

Contoh KNN: Kemiripan

- Euclidean Distance:

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

- Contoh:
  - Atribut nasabah dengan ID= 1:
    - (25,40)
  - Atribut nasabah dengan ID = 11:
    - (33, 150)
  - $d(\mathbf{x}, \mathbf{y}) = \sqrt{(25 - 33)^2 + (40 - 150)^2}$
  - $d(\mathbf{x}, \mathbf{y}) = 110.29$

Data Latih X

ID (i)	Umur (x_1)	Pinjaman (juta) (x_2)	Keputusan	Kemiripan ($d(\mathbf{x}, \mathbf{y})$)
1	25	40	Tolak	110.29
2	35	60	Tolak	90.02
3	45	80	Tolak	71.02
4	20	20	Tolak	130.65
5	35	120	Tolak	30.07
6	52	18	Setujui	133.36
7	23	95	Setujui	55.9
8	40	62	Setujui	88.28
9	60	100	Setujui	56.82
10	48	220	Setujui	71.59

Data Uji $\mathbf{y}$

ID	y_1	y_2	Keputusan
11	33	150	?

Contoh KNN: Urut Menaik

Data Latih X

ID (i)	Umur (x_1)	Pinjaman (juta) (x_2)	Keputusan	Kemiripan ($d(\mathbf{x}, \mathbf{y})$)
5	35	120	Tolak	30.07
7	23	95	Setujui	55.9
9	60	100	Setujui	56.82
3	45	80	Tolak	71.02
10	48	220	Setujui	71.59
8	40	62	Setujui	88.28
2	35	60	Tolak	90.02
1	25	40	Tolak	110.29
4	20	20	Tolak	130.65
6	52	18	Setujui	133.36

Data Uji $\mathbf{y}$

ID	y_1	y_2	Keputusan
11	33	150	?

Contoh KNN: Label

- Mengambil K = 5 nasabah dengan nilai kemiripan terkecil
- Mayoritas keputusan: **Setujui**

Data Latih X

ID (i)	Umur (x_1)	Pinjaman (juta) (x_2)	Keputusan	Kemiripan ($d(\mathbf{x}, \mathbf{y})$)
5	35	120	Tolak	30.07
7	23	95	Setujui	55.9
9	60	100	Setujui	56.82
3	45	80	Tolak	71.02
10	48	220	Setujui	71.59
8	40	62	Setujui	88.28
2	35	60	Tolak	90.02
1	25	40	Tolak	110.29
4	20	20	Tolak	130.65
6	52	18	Setujui	133.36

Data Uji $\mathbf{y}$

ID	y_1	y_2	Keputusan
11	33	150	Setujui

Clustering

K-Means

UNTAR
Universitas Tarumanagara
UNTAR untuk INDONESIA

K-Means Clustering

- Algoritma berbasis pengulangan,
  - Pada setiap pengulangan, sebuah dataset dikelompokkan dalam K sub-kelompok (K clusters)
    - Dataset terdiri dari n data points dengan setiap data point memiliki m atribut yang sama,
    - Setiap data point merupakan anggota hanya dari satu cluster (non-overlapping subgroups)
  - Satu kelompok memiliki tingkat kemiripan yang hampir sama (homogen)
    - Nilai kemiripan untuk sebuah data point dihitung berdasarkan jarak antara setiap data point di sebuah cluster c dengan centroid dari cluster tersebut.
    - Centroid dari sebuah cluster c adalah nilai rata-rata dari semua data points (anggota) dari cluster c.
  - Pengulangan berhenti jika semua anggota di setiap cluster c adalah anggota dari cluster c pada pengulangan sebelumnya.

UNTAR Universitas Tarumanagara
UNTAR untuk INDONESIA

Implementasi K-Means

- Text Clustering
  - Analisis social media
  - Analisis document
- Segmentasi pasar (market segmentation)
- Segmentasi citra
- Kompresi gambar
- Klasifikasi citra penginderaan jauh

Algoritma K-Means

Diketahi sebuah dataset X yang terdiri dari n data points dengan setiap data point memiliki m atribut

1. Tentukan jumlah cluster, yaitu nilai K,
2. Inisialisasi dengan mengambil secara acak K data points (K centroids) dari dataset X,
3. Hitung nilai kemiripan (similarity/distance) antara setiap data point dengan setiap centroid,
   - **Cosine Similarity**
4. Tentukan cluster terdekat untuk dari setiap data point berdasarkan nilai kemiripan paling besar (nilai jarak yang terkecil) dengan salah satu centroid,
5. Hitung centroid baru (nilai rata-rata dari semua data points) dari setiap cluster,
6. Ulangi langkah 3 – 5 sampai
   a. tidak ada perubahan angota pada setiap cluster dibanding dengan pengulangan sebelumnya, atau
   b. total nilai varians dari setiap cluster mencapai batas paling rendah.

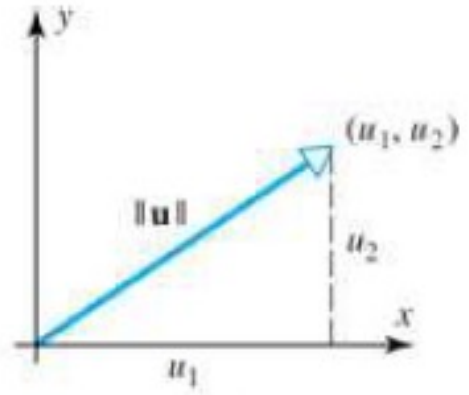
UNTAR Universitas Tarumanagara
UNTAR untuk INDONESIA

Ingat: Norm dari Sebuah Vektor

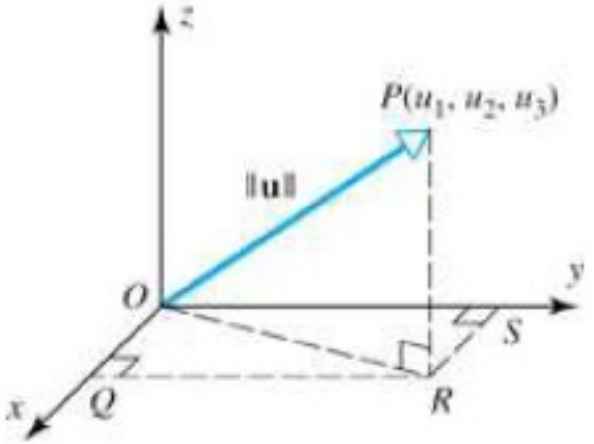
- Panjang dari sebuah vektor v disebut sebagai norm dari $\mathbf{v}$ yang dinotasikan sebagai $\|v\|$
- Untuk vector di ruang R^n:

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

- $\|k\mathbf{v}\| = |k|\|\mathbf{v}\|$
- Contoh: norm dari vektor $\mathbf{v} = (-3,2,1)$:
  - $\|\mathbf{v}\| = \sqrt{(-3)^2 + 2^2 + 1^2} = \sqrt{14}$



(a)



(b)

UNTAR
Universitas Tarumanagara

Ingat: Dot Product!

- Diketahui:
  - Dua vektor $\mathbf{u}$ dan $\mathbf{v}$ di ruang R^2 atau R^3,
  - Titik awal dari $\mathbf{u}$ dan $\boldsymbol{v}$ saling berhimpit, dan
  - θ adalah sudut antara $\mathbf{u}$ dan $\mathbf{v}$.
- **Dot product** berdasarkan komponen $\mathbf{u}$ dan $\mathbf{v}$ di R^2dan R^3:

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 \qquad \text{untuk } R^2$$

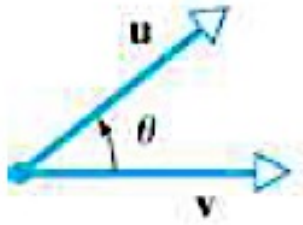
$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + u_3 v_3 \qquad \text{untuk } R^3$$

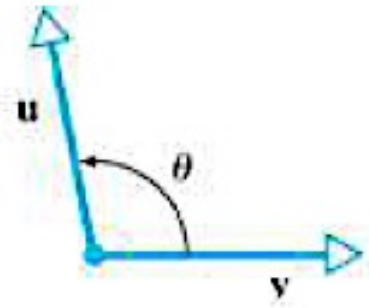
- Untuk mencari sudut antara dua vektor $\mathbf{u}$ dan $\mathbf{v}$ ($\mathbf{u} \neq 0, \mathbf{v} \neq 0$):

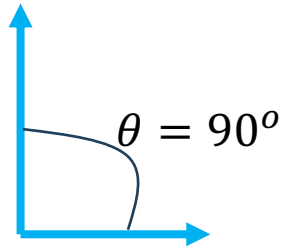
$$\cos\theta = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Jenis Sudut Hasil Dot Product

- θ adalah sudut tumpul ($90^o < \theta < 180^o$) => $\mathbf{u} \cdot \mathbf{v} < 0$
- θ adalah sudut lancip ($0^o < \theta < 90^o$) => $\mathbf{u} \cdot \mathbf{v} > 0$
- $\theta = \frac{\pi}{2}(90^o)$ => $\mathbf{u} \cdot \mathbf{v} = 0$
  - $\mathbf{u}$ dan $\mathbf{v}$ disebut vektor orthogonal ($\mathbf{u} \perp \mathbf{v}$)







UNTAR
Universitas Tarumanagara
UNTAR untuk INDONESIA

Contoh: Data

Diketahui data pengajuan pinjaman dari 5 nasabah yang disetujui dan 5 nasabah yang tidak disetujui.

1. K = 2 clusters
2. Random centroids:
   - Cluster 0: nasabah 3
     - $c_0 = (45, 80)$
   - Cluster 1: nasabah 8
     - $c_1 = (40, 62)$

Dataset

ID Nasabah	Atribut	
	Umur (tahun)	Pinjaman (juta)
1	25	40
2	35	60
3	45	80
4	20	20
5	35	120
6	52	18
7	23	95
8	40	62
9	60	100
10	48	220

Contoh: Cosine Similarity

3. $\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$

- Contoh nasabah 1 ke centroid c_0 :
  - $\mathbf{u}\cdot\mathbf{v} = u_1v_1 + u_2v_2 = 25\times45 + 40\times80$
  - $\mathbf{u}\cdot\mathbf{v} = 4325$
  - $\|\mathbf{u}\|\|\mathbf{v}\| = 4329.62181$
  - $\cos\theta = \frac{4325}{4329.62181} = 0.99893251$
- Contoh nasabah 1 ke centroid c_1 :
  - $\mathbf{u}\cdot\mathbf{v} = u_1v_1 + u_2v_2 = 25\times40 + 40\times62$
  - $\mathbf{u}\cdot\mathbf{v} = 3480$
  - $\|\mathbf{u}\|\|\mathbf{v}\| = 3480.35918$
  - $\cos\theta = \frac{3480}{3480.35918} = 0.9998968$

Dataset

ID	Atribut		$c_0 = (45, 80)$ $v_1 = 45$ $v_2 = 80$	$c_1 = (40, 62)$ $v_1 = 40$ $v_2 = 62$
	Umur (u_1)	Pinjaman (u_2)		
1	25	40	0.99893251	0.9998968
2	35	60	0.99987699	0.99899254
3	45	80	1	0.99816579
4	20	20	0.96296402	0.97752104
5	35	120	0.97398566	0.95848031
6	52	18	0.74839161	0.78717225
7	23	95	0.96246461	0.94426835
8	40	62	0.99816579	1
9	60	100	0.99960718	0.9994704
10	48	220	0.95605059	0.93654667

Contoh: Anggota Cluster

4. Cluster terdekat per nasabah (lihat tabel)
5. Centroid baru untuk setiap cluster

Cluster c_0

35	60
45	80
35	120
23	95
60	100
48	220
Rata-rata	**41**	**112.5**

Cluster c_1

25	40
20	20
52	18
40	62
34.25	**35**

Dataset

ID	Atribut		$c_0 = (45, 80)$	$c_1 = (40, 62)$
	Umur (u_1)	Pinjaman (u_2)		
1	25	40	0.99893251	0.9998968
2	35	60	0.99987699	0.99899254
3	45	80	1	0.99816579
4	20	20	0.96296402	0.97752104
5	35	120	0.97398566	0.95848031
6	52	18	0.74839161	0.78717225
7	23	95	0.96246461	0.94426835
8	40	62	0.99816579	1
9	60	100	0.99960718	0.9994704
10	48	220	0.95605059	0.93654667

UNTAR Universitas Tarumanagara
UNTAR untuk INDONESIA

Contoh: Loop ke-2

3. $\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$
4. Cluster terdekat per nasabah (lihat tabel)
5. Centroid baru untuk setiap cluster
   - cluster c_0: (38.71, 102.14)
   - cluster c_1: (37.33, 33.33)

Dataset

ID	Atribut		c_0 $v_1 = 41$ $v_2 = 112.5$	c_1 $v_1 = 34.25$ $v_2 = 35$
	Umur (u_1)	Pinjaman (u_2)		
1	25	40	0.97821506	0.97676921
2	35	60	0.98409516	0.96977391
3	45	80	0.98676031	0.96582756
4	20	20	0.90648465	0.99994136
5	35	120	0.99784316	0.88196826
6	52	18	0.63091254	0.89472387
7	23	95	0.99374015	0.85923002
8	40	62	0.97513174	0.97974703
9	60	100	0.98182723	0.97271219
10	48	220	0.99094581	0.84738636

UNTAR Universitas Tarumanagara
UNTAR untuk INDONESIA

Contoh: Loop ke-3

3. $\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$
4. Cluster terdekat per nasabah (lihat tabel)
5. Centroid baru untuk setiap cluster
   - cluster c_0: (38.875,97.125)
   - cluster c_1: (36, 19)

Dataset

ID	Atribut		c_0 $v_1 = 38.71$ $v_2 = 102.14$	c_1 $v_1 = 37.33$ $v_2 = 33.33$
	Umur (u_1)	Pinjaman (u_2)		
1	25	40	0.98078802	0.96012492
2	35	60	0.98628488	0.9511451
3	45	80	0.9887523	0.94618565
4	20	20	0.9118064	0.99840154
5	35	120	0.99692283	0.84823486
6	52	18	0.64077563	0.92276355
7	23	95	0.99223139	0.82283622
8	40	62	0.97788427	0.96404225
9	60	100	0.98417221	0.95488399
10	48	220	0.98914916	0.80971413

UNTAR Universitas Tarumanagara
UNTAR untuk INDONESIA

Contoh: Loop ke-4

3. $\cos\theta = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$
4. Cluster terdekat per nasabah (lihat tabel)

Anggota setiap cluster sama dengan Loop ke-3!

SELESAI…

Dataset

ID	Atribut		c_0 $v_1 = 38.875$ $v_2 = 97.125$	c_1 $v_1 = 36$ $v_2 = 19$
	Umur (u_1)	Pinjaman (u_2)		
1	25	40	0.98422261	0.8645335
2	35	60	0.98916418	0.8487921
3	45	80	0.99134518	0.84039496
4	20	20	0.91923241	0.95540264
5	35	120	0.99530556	0.69571601
6	52	18	0.65484125	0.98841295
7	23	95	0.9897653	0.66175463
8	40	62	0.98157913	0.87166444
9	60	100	0.98727643	0.85525442
10	48	220	0.98626801	0.64455224

UNTAR Universitas Tarumanagara
UNTAR untuk INDONESIA

Soal Latihan

Lakukan perhitungan untuk kedua soal berikut di excel:

1. (50 poin) Gunakan dataset latih dan data uji di “datasets.xlsx” sheet “KNN” untuk mensimulasikan algoritma KNN.
   a. Gunakan Eucliean Distance
   b. Set k = 7
2. (50 poin) Gunakan dataset di “datasets.xlsx” sheet “KMeans” yang terdiri dari 4 variable bebas (client, rate of return, sales, years) untuk di-cluster oleh K-Means menjadi 3 kelompok.
   a. Gunakan Cosine Similarity
   b. Lakukan clustering dengan K-Means sampai dengan konvergen atau tidak melebihi dari 10 iterasi