



Unsupervised Learning



Review

- Problem Solving Agent
- Knowledge Based Agent
- Learning Agent
 - ☐ Supervised Learning
 - ☐ **Unsupervised Learning**
 - ☐ Reinforcement Learning

Unsupervised Learning

- Clustering
- Input: training set (data)
- Output: cluster → centroid, threshold
- Usage: using centroid to group data
- Process: calculate similarity of data

MLK/IF-ITB/2010

3

K-Means

- mempartisi koleksi vektor $\{x_i\}$ ke dalam set cluster $\{C_i\}$. Algoritma ini membutuhkan k cluster seeds atau nilai random untuk inisialisasi.
- K-Means populer karena sederhana dan efisien (kompleksitas setiap iterasi $O(kn)$ perbandingan, dan jumlah iterasi cukup kecil).

MLK/IF-ITB/2010

4

K-Means Algorithm

Inisialisasi:

k seeds sebagai centroid dari k clusters

Setiap vektor lain di-assign ke cluster dengan seed terdekat

Iterasi:

$$M_i \leftarrow |C_i|^{-1} \sum_{x \in C_i} x \quad \{\text{perhitungan centroid } M_i \text{ dari cluster}\}$$

Setiap vektor di-reassign lagi ke cluster dengan centroid terdekat

Stop: convergence {tidak terjadi perubahan lagi }

MLK/IF-ITB/2010

5

K-Means (2)

- Masalah utama K-Means adalah sensitif terhadap pemilihan seed inisialisasi sehingga hasilnya tidak optimal.

MLK/IF-ITB/2010

6

Abstraksi data

- Abstraksi data: memberi label yang bermakna pada cluster.
- Label yang baik akan singkat dan membedakan dengan tepat satu cluster dengan yang lain.

MLK/IF-ITB/2010

7

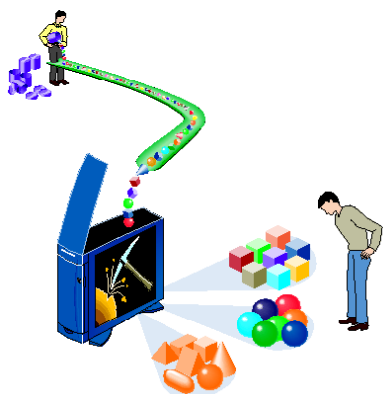
Evaluasi

- $\text{Purity}(C_i) = \max_j |L_j \cap C_i| / |C_i|$
 - $\{L_1, \dots, L_n\}$: dokumen berlabel
 - $\{C_1, \dots, C_m\}$: cluster yang dihasilkan clustering
- utility dari cluster yang dihasilkan untuk aplikasi yang akan menggunakan

MLK/IF-ITB/2010

8

Klasifikasi vs Clustering



	Klasifikasi	Clustering
Learning	supervised	unsupervised
Input	<data, label>	data
Output	Deskripsi kelas (pola/model)	Centroid, anggota kelas
Klasifikasi unseen data	Model → kelas	Kemiripan dengan kelas

MLK/IF-ITB/2006

9

Latihan

■ Diberikan data training sebagai berikut:

Data	Pengalaman	Programming	Bahasa Inggris	Buta warna	Menikah
A	Tidak	Ya	Ya	Tidak	Tidak
B	Tidak	Tidak	Ya	Tidak	Ya
C	Ya	Ya	Tidak	Ya	Tidak
D	Tidak	Tidak	Tidak	Ya	Ya
E	Ya	Tidak	Ya	Ya	Ya
F	Tidak	Ya	Tidak	Ya	Tidak

MLK/IF-ITB/2010

10

Latihan (2)

- Lakukan proses clustering K-Means dengan jumlah cluster (k) = 3:
 - Seed cluster 1: A; Seed cluster 2: B; Seed cluster 3: C
- Gunakanlah jumlah perbedaan sebagai jarak antar data. Jika jarak sama, data cenderung dimasukkan ke cluster 2.
- Centroid diambil sebagai mayoritas nilai atribut pada cluster. Jika tidak ada mayoritas (jumlah Ya = jumlah Tidak), ambil nilai centroid sebelumnya untuk atribut tersebut.
- Kondisi berhenti: pada setiap cluster, jarak terjauh anggota dengan centroid < 2 . Jarak terjauh dihitung setelah centroid terbentuk.

MLK/IF-ITB/2010

11

Solusi Latihan

- 1:A ; 2:B,D,E ; 3: C,F

Iterasi 1:

A	Tidak	Ya	Ya	Tidak	Tidak
---	-------	----	----	-------	-------

Centroid 1: $\langle T, Y, Y, T, T \rangle$

A: 0 \rightarrow jarak terjauh=0

B	Tidak	Tidak	Ya	Tidak	Ya
D	Tidak	Tidak	Tidak	Ya	Ya
E	Ya	Tidak	Ya	Ya	Ya

Centroid 2: $\langle T, T, Y, Y, Y \rangle$

B:1, D:1, E:1 \rightarrow jarak terjauh=1

Solusi Latihan (2)

- 1:A ; 2:B,D,E; 3: C,F

Iterasi 1:

C	Ya	Ya	Tidak	Ya	Tidak
F	Tidak	Ya	Tidak	Ya	Tidak

Centroid 3: <Y,Y,T,Y,T>

C: 0, F: 1 → jarak terjauh=1