CII-2M3 Pengantar Kecerdasan Buatan

k-Nearest Neighbour

Suyanto

S1 Informatika – Fakultas Informatika

k-Nearest Neighbour (kNN)

- *Instance-Based Learning* (IBL)
- *Lazy Learner* (pembelajar malas)
- Tidak melakukan proses belajar (dari data latih)
- Klasifikasi secara langsung berdasarkan **tetangga terdekat**

k-Nearest Neighbour (kNN)

- *Instance-Based Learning* (IBL)
- *Lazy Learner* (pembelajar malas)
- Tidak melakukan proses belajar (dari data latih)
- Klasifikasi secara langsung berdasarkan **tetangga terdekat**
- Bekerja secara lokal
  - kNN bekerja secara lokal sehingga cocok digunakan untuk himpunan data yang mengandung *outlier* atau pencilan
- Bisa digunakan untuk data apapun
  - Numerik maupun non-numerik. Diskrit maupun kontinu.
- Formula jarak atau *dissimilarity*?

k-Nearest Neighbour (kNN)

- *Instance-Based Learning* (IBL)
- *Lazy Learner* (pembelajar malas)
- Tidak melakukan proses belajar (dari data latih)
- Klasifikasi secara langsung berdasarkan **tetangga terdekat**
- Bekerja secara lokal kNN bekerja secara lokal
  - sehingga cocok digunakan untuk himpunan data yang mengandung *outlier* atau pencilan
- Bisa digunakan untuk data apapun
  - Numerik maupun non-numerik. Diskrit maupun kontinu.
- Formula jarak atau *dissimilarity*?

CII-2M3 Pengantar Kecerdasan Buatan

Jarak Atribut Numerik

Suyanto

S1 Informatika – Fakultas Informatika

Jarak Atribut Numerik: *Euclidean Distance*

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

Nama	Pulsa (ribu)	Internet (ribu)
Andi	100	200
Budi	400	600
Citra	100	100
Dedi	150	200
Evan	700	400

$$d(1,2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2} = \sqrt{90000 + 160000} = 500$$

Jarak Atribut Numerik: *Manhattan Distance*

$$d(i, j) = \left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \ldots + \left|x_{ip} - x_{jp}\right|$$

Nama	Pulsa (ribu)	Internet (ribu)
Andi	100	200
Budi	400	600
Citra	100	100
Dedi	150	200
Evan	700	400

$$d(1,2) = \left|x_{11} - x_{21}\right| + \left|x_{12} - x_{22}\right| = \left|100 - 400\right| + \left|200 - 600\right| = 300 + 400 = 700$$

Jarak Atribut Numerik: *Minkowski Distance*

$$d(i,j) = \sqrt[h]{\left|x_{i1} - x_{j1}\right|^h + \left|x_{i2} - x_{j2}\right|^h + \ldots + \left|x_{ip} - x_{jp}\right|^h}$$

Nama	Pulsa (ribu)	Internet (ribu)
Andi	100	200
Budi	400	600
Citra	100	100
Dedi	150	200
Evan	700	400

$$d(1,2) = \sqrt[1,5]{\left|x_{11} - x_{21}\right|^{1,5} + \left|x_{12} - x_{22}\right|^{1,5}} = \sqrt[1,5]{5196{,}15 + 8000} = 558{,}42$$

Jarak Atribut Numerik: *Supremum Distance*

$$d(i,j) = \lim_{h\to\infty}\left(\sum_{f=1}^{p}\left|x_{if} - x_{jf}\right|^{h}\right)^{\frac{1}{h}} = \max_{f}^{p}\left|x_{if} - x_{jf}\right|$$

Nama	Pulsa (ribu)	Internet (ribu)
Andi	100	200
Budi	400	600
Citra	100	100
Dedi	150	200
Evan	700	400

$d(1,2) =$ max(|100-400|, |200-600|) = 400

CII-2M3 Pengantar Kecerdasan Buatan

Jarak Atribut Non Numerik

Suyanto

S1 Informatika – Fakultas Informatika

Jarak Atribut Non Numerik: Nominal

$$d(i, j) = \frac{p - m}{p}$$

Nama	Pekerjaan	Lokasi Rumah
Andi	Analis	A
Budi	Dokter	A
Citra	Guru	B
Dedi	Analis	A
Evan	Dokter	C

$$d(1,2) = d(2,1) = \frac{2-1}{2} = 0{,}5$$

$$\begin{bmatrix} 0 & & & & \\ 0.5 & 0 & & & \\ 1 & 1 & 0 & & \\ 0 & 1 & 1 & 0 & \\ 1 & 0{,}5 & 1 & 1 & 0 \end{bmatrix}$$

Dissimilarity Matrix

Jarak Atribut Non Numerik: Ordinal

$$d(i, j) = \frac{p - m}{p}$$

Nama	Jumlah Anak	Kategori Pelanggan
Andi	0	Silver
Budi	2	Platinum
Citra	0	Silver
Dedi	3	Gold
Evan	4	Platinum

$$d(i, j) = \sqrt{(0 - 0{,}5)^2 + (0 - 1)^2} \quad = 1{,}12$$

Jarak Atribut Non Numerik: Biner Simetris

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

Nama	Gender	Kartu
Andi	Pria	Prabayar
Budi	Pria	Pascabayar
Citra	Wanita	Prabayar
Dedi	Pria	Prabayar
Evan	Pria	Pascabayar

$$d(1,2) = d(2,1) = \frac{1+0}{1+1+0+0} = 0{,}50$$

$$\begin{bmatrix} 0{,}00 & & & & \\ 0{,}50 & 0{,}00 & & & \\ 0{,}50 & 1{,}00 & 0{,}00 & & \\ 0{,}00 & 0{,}50 & 0{,}50 & 0{,}00 & \\ 0{,}50 & 0{,}00 & 1{,}00 & 0{,}50 & 0{,}00 \end{bmatrix}$$

Dissimilarity Matrix

Jarak Atribut Non Numerik: Biner Asimetris

$$d(i, j) = \frac{r + s}{q + r + s}$$

Nama	Rumah	Menikah
Andi	Kontrak	Tidak
Budi	Prbadi	Ya
Citra	Kontrak	Tidak
Dedi	Kontrak	Ya
Evan	Prbadi	Ya

$$d(1,2) = d(2,1) = \frac{0 + 2}{0 + 0 + 2} = 1{,}00$$

$$\begin{bmatrix} 0{,}00 & & & & \\ 1{,}00 & 0{,}00 & & & \\ 0{,}00 & 1{,}00 & 0{,}00 & & \\ 0{,}50 & 0{,}50 & 0{,}50 & 0{,}00 & \\ 1{,}00 & 0{,}00 & 1{,}00 & 0{,}50 & 0{,}00 \end{bmatrix}$$

Dissimilarity Matrix

CII-2M3 Pengantar Kecerdasan Buatan

Jarak Atribut Campuran

Suyanto

S1 Informatika – Fakultas Informatika

Jarak Atribut Campuran

Nama	Pekerjaan	Lokasi Rumah	Gender	Kartu	Rumah	Menikah	Pulsa (ribu)	Internet (ribu)	Jumlah Anak	Kategori Pelanggan
Andi	Analis	A	Pria	Prabayar	Kontrak	Tidak	100	200	0	Silver
Budi	Dokter	A	Pria	Pascabayar	Prbadi	Ya	400	600	2	Platinum
Citra	Guru	B	Wanita	Prabayar	Kontrak	Tidak	100	100	0	Silver
Dedi	Analis	A	Pria	Prabayar	Kontrak	Ya	150	200	3	Gold
Evan	Dokter	C	Pria	Pascabayar	Prbadi	Ya	700	400	4	Platinum

Jarak Atribut Campuran

Nama	Pekerjaan	Lokasi Rumah	Gender	Kartu	Rumah	Menikah	Pulsa (ribu)	Internet (ribu)	Jumlah Anak	Kategori Pelanggan
Andi	Analis	A	Pria	Prabayar	Kontrak	Tidak	100	200	0	Silver
Budi	Dokter	A	Pria	Pascabayar	Prbadi	Ya	400	600	2	Platinum
Citra	Guru	B	Wanita	Prabayar	Kontrak	Tidak	100	100	0	Silver
Dedi	Analis	A	Pria	Prabayar	Kontrak	Ya	150	200	3	Gold
Evan	Dokter	C	Pria	Pascabayar	Prbadi	Ya	700	400	4	Platinum

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}},$$

di mana $\delta_{ij}^{(f)} = 0$ jika salah satu kondisi ini dipenuhi: a) x_{if} atau x_{jf} tidak memiliki nilai alias kosong (*missing*); atau b) $x_{if} = x_{jf} = 0$ dan f adalah atribut biner asimetris dan $\delta_{ij}^{(f)} = 1$ untuk kondisi yang lain. Sementara itu, $d_{ij}^{(f)}$ adalah kontribusi atribut f terhadap *dissimilarity* antara objek data i dan objek data j, yang dihitung berdasarkan jenis atribut tersebut, yaitu:

- Jika f adalah atribut **nominal** atau **biner**: $d_{ij}^{(f)} = 0$ jika $x_{if} = x_{jf}$ dan $d_{ij}^{(f)} = 1$ untuk semua kondisi yang lain.
- Jika f adalah atribut **numerik**: $d_{ij}^{(f)} = \frac{\left| x_{if} - x_{jf} \right|}{\max_h x_{hf} - \min_h x_{hf}}$, di mana h didapat dari semua objek yang tidak kosong (*nonmissing*) untuk atribut f.
- Jika f adalah atribut **ordinal**: hitung ranking r_{if}, kemudian normalisasikan menggunakan $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, dan perlakukan z_{if} sebagai atribut numerik.

CII-2M3 Pengantar Kecerdasan Buatan

Algoritma kNN

Suyanto

S1 Informatika – Fakultas Informatika

Algoritma kNN

1. Untuk setiap pola latih <x,$f(x)$>, tambahkan pola tersebut ke dalam **Daftar Pola Latih**

2. Untuk sebuah pola masukan x_q

   - Misalkan x_1, x_2, ..., x_k adalah k pola yang memiliki jarak terdekat (tetangga) dengan x_q

   - Kembalikan kelas yang memiliki jumlah pola paling banyak di antara k pola tersebut sebagai kelas keputusan

Algoritma kNN

1. Untuk setiap pola latih <x,$f(x)$>, tambahkan pola tersebut ke dalam **Daftar Pola Latih**

2. Untuk sebuah pola masukan x_q

   - Misalkan x_1, x_2, ..., x_k adalah k pola yang memiliki jarak terdekat (tetangga) dengan x_q

   - Kembalikan kelas yang memiliki jumlah pola paling banyak di antara k pola tersebut sebagai kelas keputusan

Algoritma kNN

1. Untuk setiap pola latih <x,$f(x)$>, tambahkan pola tersebut ke dalam **Daftar Pola Latih**

2. Untuk sebuah pola masukan x_q

   - Misalkan x_1, x_2, ..., x_k adalah k pola yang memiliki jarak terdekat (tetangga) dengan x_q

   - Kembalikan kelas yang memiliki jumlah pola paling banyak di antara k pola tersebut sebagai kelas keputusan

Kelemahan kNN

1. Sensitif terhadap fitur-fitur yang kurang relevan;
2. Sensitif terhadap ukuran ketetanggaan *k*;
3. Sensitif terhadap data berderau maupun data pencilan;
4. Kompleksitas waktu yang relatif tinggi untuk mencari tetangga terdekat di antara semua data latih setiap kali melakukan klasifikasi; dan
5. Kompleksitas memori yang relatif besar untuk menyimpan semua data latih.

CII-2M3 Pengantar Kecerdasan Buatan

Optimasi *k*

Suyanto

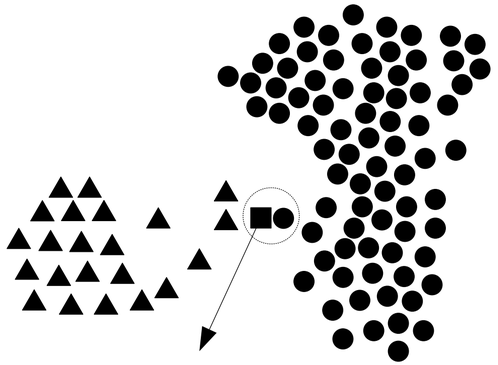
S1 Informatika – Fakultas Informatika

Berapa *k* yang optimum*?*

- Strategi menemukan *k* optimum
- Pahami data latih dengan baik
- Gunakan data latih dan data validasi untuk menghindari *overfit*

■ = ●

kNN dengan $k = 1$

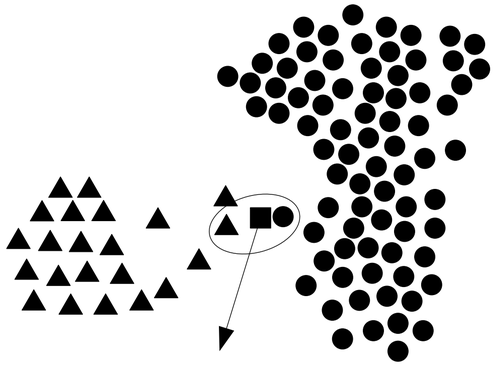


Diklasifikasi ke lingkaran
(jumlah pola terbanyak)

■ = ● √

■ = ●

kNN dengan $k = 2$

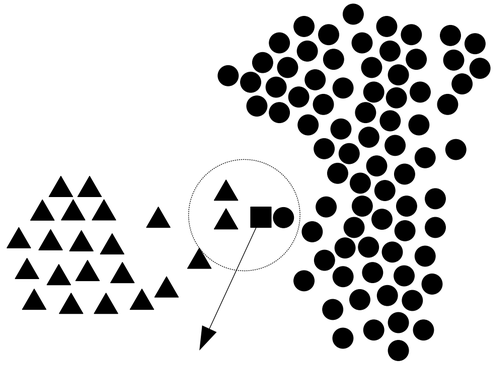


Diklasifikasi ke kelas apa?
(jumlah pola sama banyak)

■ = **?**

■ = ●

kNN dengan $k = 3$

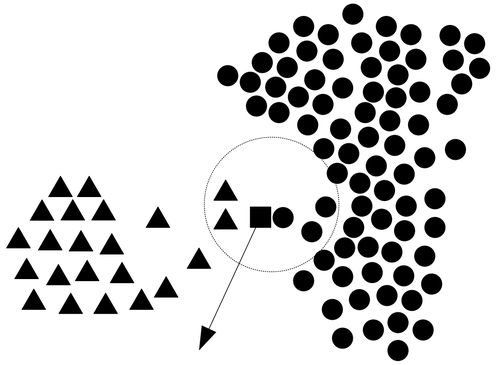


Diklasifikasi ke segitiga
(jumlah pola terbanyak)

■ = ▲ ×

■ = ●

kNN dengan $k = 5$



Diklasifikasi ke lingkaran
(jumlah pola terbanyak)

■ = ● √

CII-2M3 Pengantar Kecerdasan Buatan

Perbaikan kNN

Suyanto

S1 Informatika – Fakultas Informatika

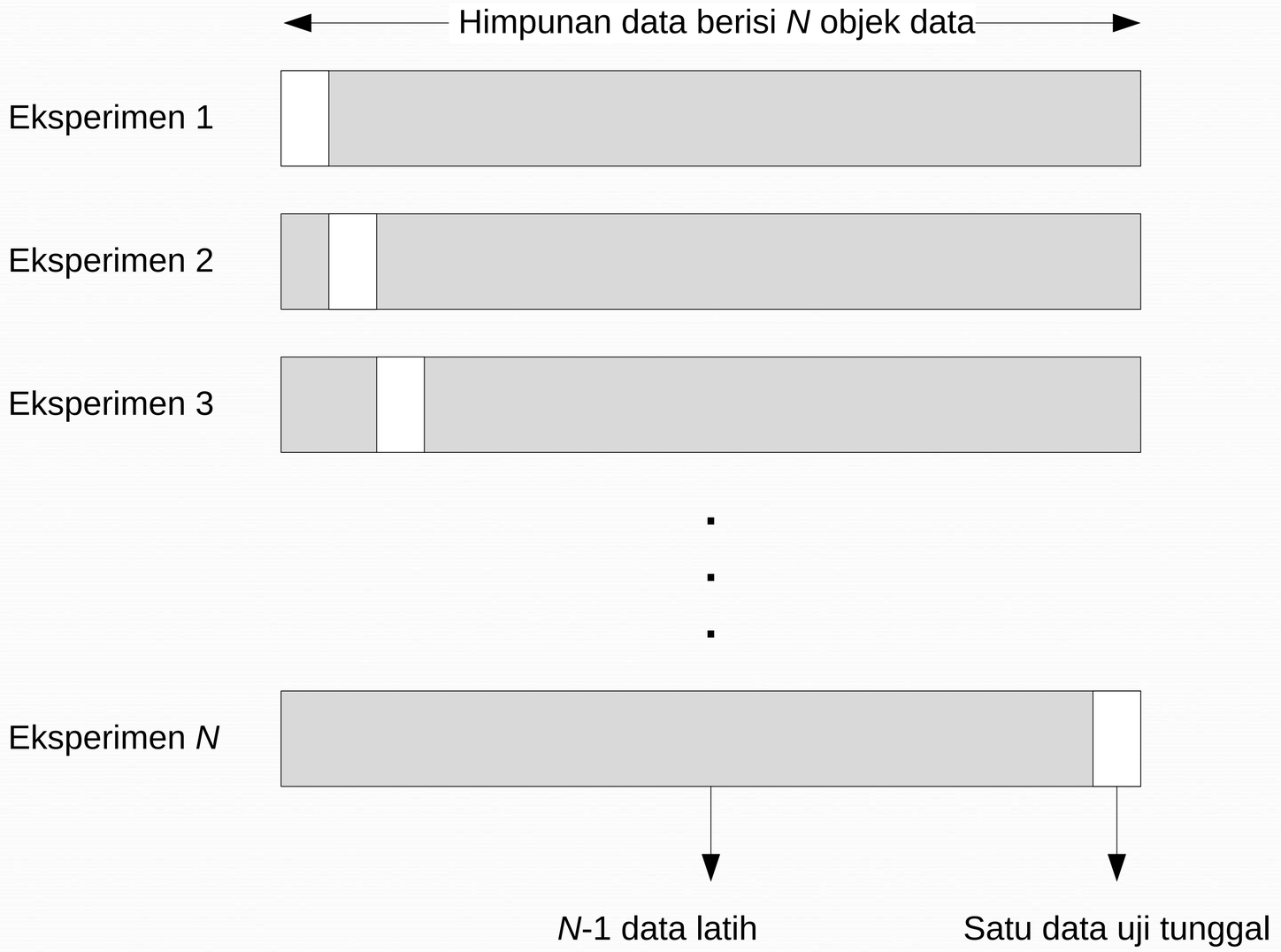
Perbaikan kNN

1. Perbaikan dengan fungsi jarak
2. Perbaikan dengan ukuran ketetanggaan
3. Perbaikan dengan estimasi probabilitas kelas
4. Perbaikan dengan struktur data

Perbaikan kNN dengan Fungsi Jarak

$$d(x,y) = \sqrt{\sum_{i=1}^{n} w_i^2 \left(a_i(x) - a_i(y)\right)^2}$$

Perbaikan kNN dengan ukuran ketetanggaan



Perbaikan dengan estimasi probabilitas kelas

1. *Locally Weighted Naïve Bayes* (LWNB)
2. *Instance Cloning Local Naïve Bayes* (ICLNB)
3. *k-local Hyperplane and Convex Distance Nearest Neighbor Algorithms* (HKNN)
4. *Fuzzy k-Nearest Neighbour* (FkNN)
5. *Fuzzy k-Nearest Neighbour in Every Class* (FkNNC)
6. *Pseudo Nearest Neighbour Rule* (PNNR)
7. Locl Mean PNNR

Perbaikan dengan struktur data

1. *Ball-Tree*
2. *kd-Tree*
3. *B-Tree*
4. *R*-Tree*
5. *R+-Tree*
6. *TPR-Tree*
7. *X-Tree*
8. *A-Tree*
9. *BD-Tree*
10. *SS-Tree*
11. *SR-Tree*
12. *Locality Sensitive Hashing* (LSH)

Terima Kasih

Contoh Soal 1

Diberikan **data latih** sebagai berikut. **x1** dan **y1** adalah data input, sedangkan **y** adalah output (kelas).

No	x1	x2	y
1	0.4	4	0
2	0.1	7	1
3	0.9	2	0
4	0.7	10	1
5	0.6	8	1

Berdasarkan data di atas, jika digunakan metode k-NN dengan **k = 3**, masuk ke **kelas** mana **data uji** berikut?

No	x1	x2	y
1	0.8	6	??

Berikut rumus untuk menghitung jarak antar vektor data.

$$d_{Euclidean} = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Jawaban Contoh Soal 1

Contoh Soal 2

Diketahui sebuah data latih kondisi cuaca beserta target data berupa review dari seorang atlet untuk berlatih pada kondisi cuaca tersebut.

Data	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Rainy	Cold	High	Strong	Warm	Change	No
3	Sunny	Warm	High	Strong	Cool	Change	Yes
4	Sunny	Warm	High	Strong	Warm	Same	Yes

Berdasarkan data di atas, jika digunakan metode k-NN dengan **k = 3**, bagaimana review seorang atlet jika berlatih dengan kondisi cuaca berikut?

Data	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Cold	Normal	Strong	Cool	Same	???

Berikut rumus untuk menghitung jarak antar vektor data.

$$d_{Euclidean} = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Jawaban Contoh Soal 2

Preprocessing data latih dari kategori (string) menjadi numerik:

Data	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	1	1	0	1	1	0	1
2	0	0	1	1	1	1	0
3	1	1	1	1	0	1	1
4	1	1	1	1	1	0	1

Preprocessing data uji dari kategori (string) menjadi numerik:

Data	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	1	0	0	1	0	0	???

Hitung jarak tetangga:

Data1=2 1.414213562
Data2=4 2
Data3=3 1.732050808
Data4=3 1.732050808

Sorting ascending:

Data1=2 1.414213562
Data3=3 1.732050808
Data4=3 1.732050808
Data2=4 2

Tentukan kelas output dari data uji beserta alasannya:

[Data1, Data3, Data4] = [1, 1, 1] = 1