

Data warehouse dan Business Intelligence Systems (9th Ed., Prentice Hall)

Pertemuan 6: Metode Learning Algoritma Data Mining

Konsep dan Definisi Machine Learning

- **Machine learning (ML)** adalah keluarga teknologi kecerdasan buatan yang terutama berkaitan dengan desain dan pengembangan algoritma yang memungkinkan komputer untuk "belajar" dari data historis
 - ML adalah proses dimana komputer belajar dari pengalaman
 - Ini berbeda dari perolehan pengetahuan di ES: bukannya mengandalkan para ahli (dan kemauan mereka) ML bergantung pada fakta sejarah
 - ML membantu dalam menemukan pola dalam data

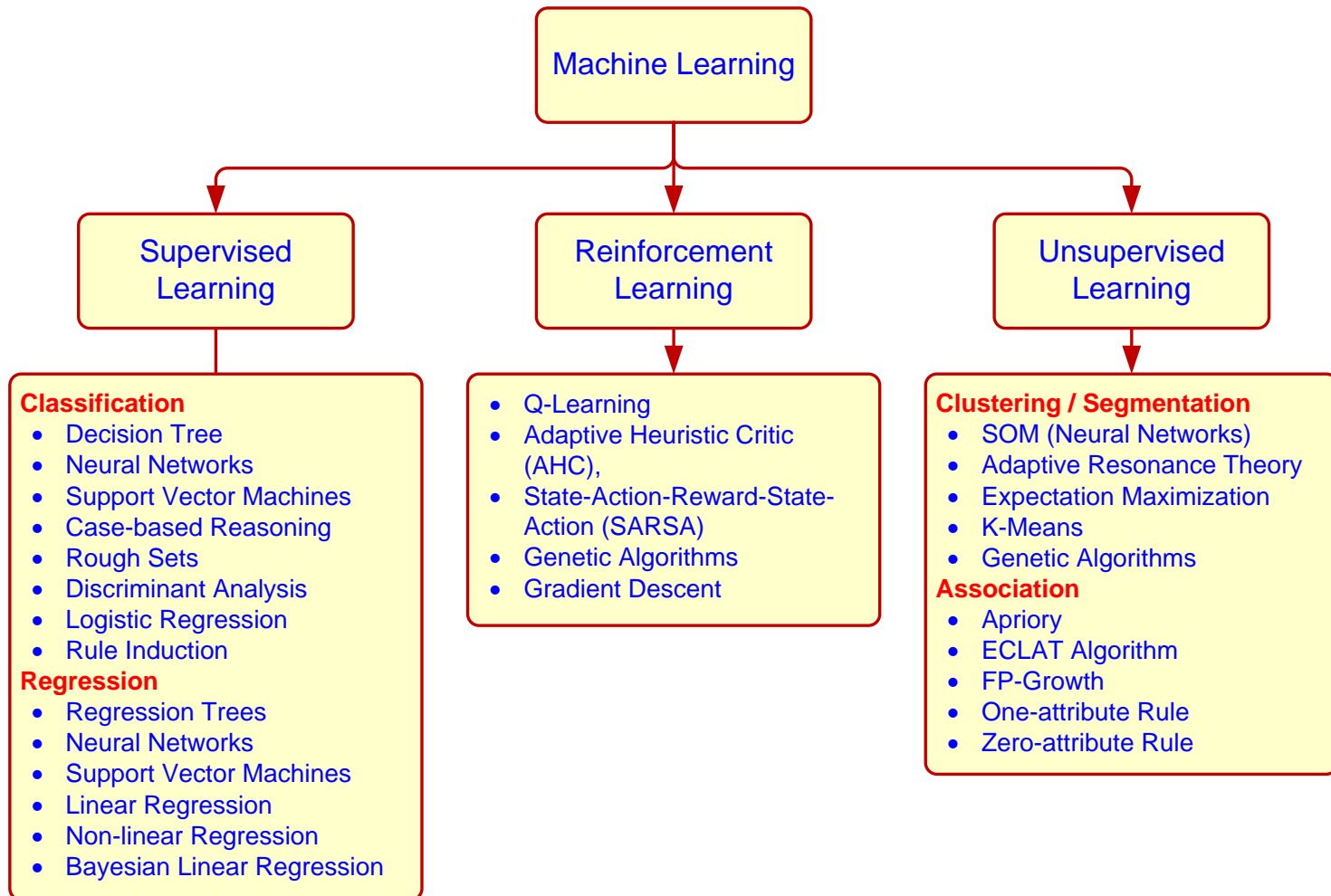
Konsep dan Definisi Machine Learning

- **Learning** adalah proses peningkatan diri, yang merupakan fitur penting dari perilaku cerdas
- Pembelajaran manusia adalah kombinasi dari banyak proses kognitif yang rumit, termasuk :
 - Induksi
 - Deduksi
 - Analogi
 - Prosedur khusus lainnya yang terkait dengan mengamati dan / atau menganalisis contoh

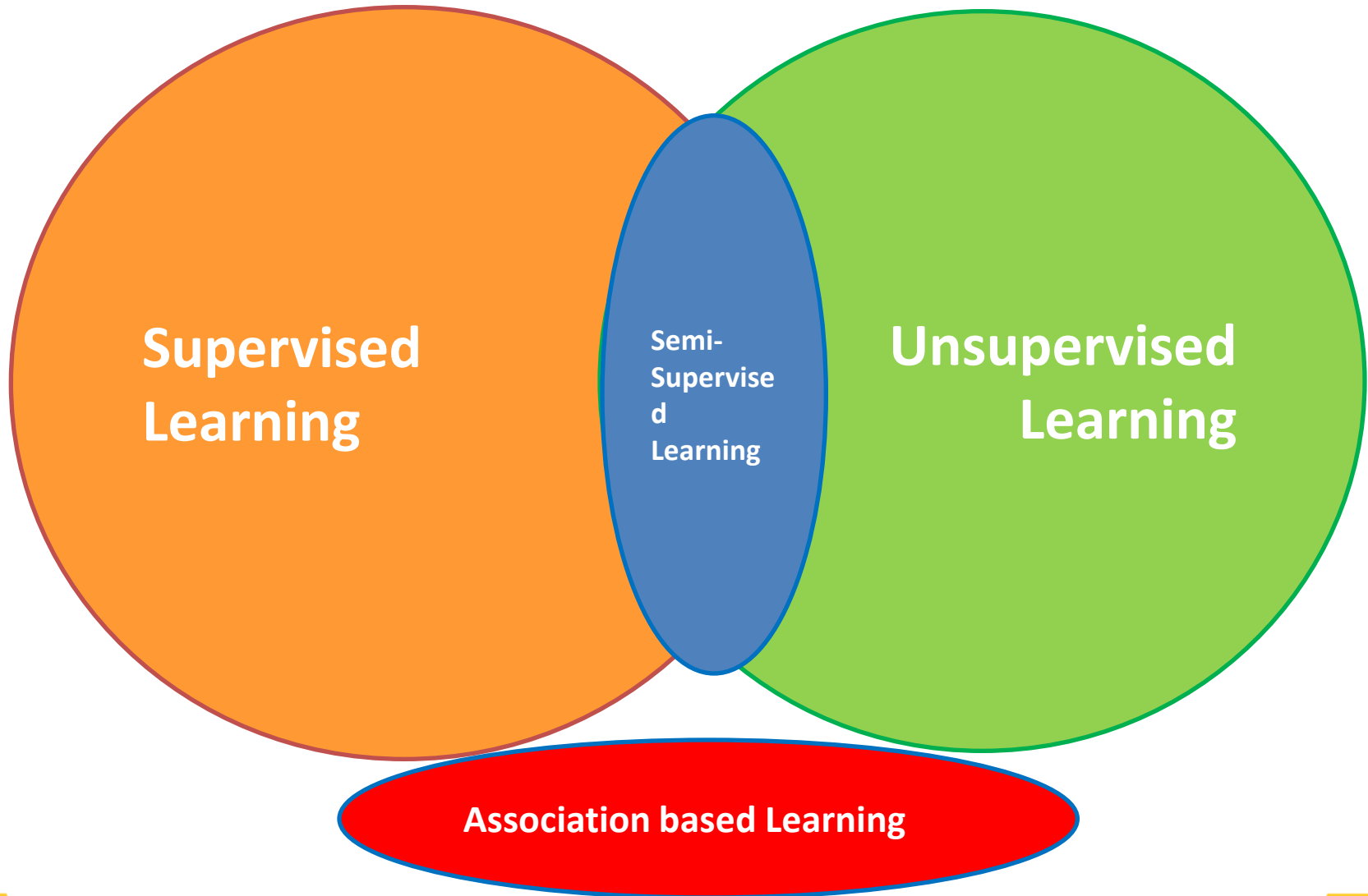
Konsep dan Definisi Machine Learning

- **Machine Learning versus Human Learning**
 - Beberapa perilaku ML dapat menantang kinerja pakar manusia (mis., Bermain catur)
 - Meskipun ML kadang-kadang sesuai dengan kemampuan belajar manusia, ia tidak dapat belajar sebaik manusia atau dengan cara yang sama seperti yang dilakukan manusia
 - Tidak ada klaim bahwa pembelajaran mesin dapat diterapkan dengan cara yang benar-benar kreatif
 - Sistem ML tidak berlabuh dalam teori formal apa pun (mengapa mereka berhasil atau gagal tidak jelas)
 - Keberhasilan ML sering dikaitkan dengan manipulasi simbol (bukan hanya informasi numerik)

Metode Machine Learning



Metode Learning Algoritma Data Mining



1. Supervised Learning

- Pembelajaran dengan **guru**, data set memiliki **target/label/class**
- **Sebagian besar** algoritma data mining (estimation, prediction/forecasting, classification) adalah supervised learning
- Algoritma melakukan proses belajar berdasarkan **nilai dari variabel target** yang terasosiasi dengan nilai dari variable prediktor

Dataset dengan Class

Attribute/Feature/Dimension

Class/Label/Target

	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	<i>Iris setosa</i>
2	4.9	3.0	1.4	0.2	<i>Iris setosa</i>
3	4.7	3.2	1.3	0.2	<i>Iris setosa</i>
4	4.6	3.1	1.5	0.2	<i>Iris setosa</i>
5	5.0	3.6	1.4	0.2	<i>Iris setosa</i>
...					
51	7.0	3.2	4.7	1.4	<i>Iris versicolor</i>
52	6.4	3.2	4.5	1.5	<i>Iris versicolor</i>
53	6.9	3.1	4.9	1.5	<i>Iris versicolor</i>
54	5.5	2.3	4.0	1.3	<i>Iris versicolor</i>
55	6.5	2.8	4.6	1.5	<i>Iris versicolor</i>
...					
101	6.3	3.3	6.0	2.5	<i>Iris virginica</i>
102	5.8	2.7	5.1	1.9	<i>Iris virginica</i>
103	7.1	3.0	5.9	2.1	<i>Iris virginica</i>

Nominal

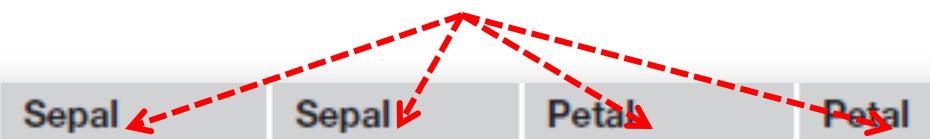
Numerik

2. Unsupervised Learning

- Algoritma data mining mencari pola dari **semua variable (atribut)**
- Variable (atribut) yang menjadi **target/label/class tidak ditentukan (tidak ada)**
- Algoritma **clustering** adalah algoritma unsupervised learning

Dataset tanpa Class

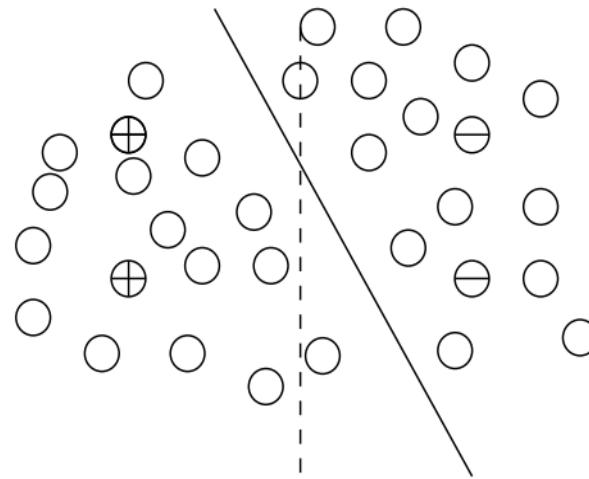
Attribute/Feature/Dimension



	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
...				
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
...				
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1

3. Semi-Supervised Learning

- Semi-supervised learning adalah metode data mining yang menggunakan **data dengan label dan tidak berlabel sekaligus** dalam proses pembelajarannya
- Data yang memiliki kelas digunakan untuk **membentuk model** (pengetahuan), data tanpa label digunakan untuk **membuat batasan** antara kelas



⊕ Positive example - - - - Decision boundary without unlabeled examples
⊖ Negative example ——— Decision boundary with unlabeled examples
○ Unlabeled example

Proses Data Mining

No	Nama	Status	SP1	SP2	SP3
1	PERDANA	MAHASISWA	2.50	1.5	1.2
2	PERDANA	MAHASISWA	3.0	1.5	1.2
3	PERDANA	MAHASISWA	3.5	1.5	1.2
4	PERDANA	MAHASISWA	3.5	1.5	1.2
5	PERDANA	MAHASISWA	3.5	1.5	1.2
6	PERDANA	MAHASISWA	3.5	1.5	1.2
7	PERDANA	MAHASISWA	3.5	1.5	1.2
8	PERDANA	MAHASISWA	3.5	1.5	1.2
9	PERDANA	MAHASISWA	3.5	1.5	1.2
10	PERDANA	MAHASISWA	3.5	1.5	1.2
11	PERDANA	MAHASISWA	3.5	1.5	1.2
12	PERDANA	MAHASISWA	3.5	1.5	1.2
13	PERDANA	MAHASISWA	3.5	1.5	1.2
14	PERDANA	MAHASISWA	3.5	1.5	1.2
15	PERDANA	MAHASISWA	3.5	1.5	1.2
16	PERDANA	MAHASISWA	3.5	1.5	1.2
17	PERDANA	MAHASISWA	3.5	1.5	1.2
18	PERDANA	MAHASISWA	3.5	1.5	1.2
19	PERDANA	MAHASISWA	3.5	1.5	1.2
20	PERDANA	MAHASISWA	3.5	1.5	1.2

1. Himpunan Data

(Pemahaman dan Pengolahan Data)

DATA PRE-PROCESSING
Data Cleaning
Data Integration
Data Reduction
Data Transformation

$$f(x) \approx \lim_{n \rightarrow \infty} \frac{b-a}{n} \sum_{k=1}^n f\left(a + \frac{b-a}{n} \cdot k\right)$$

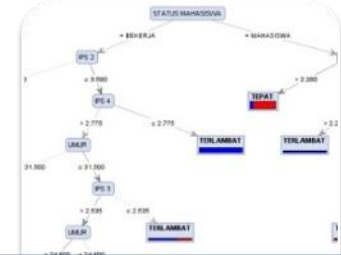
$$e^{i\theta} = \cos(\theta) + i\sin(\theta)$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \frac{x^7}{7!} + \frac{x^8}{8!} + \frac{x^9}{9!} + \frac{x^{10}}{10!} + \dots$$

2. Metode Data Mining

(Pilih Metode Sesuai Karakter Data)

Estimation
Prediction
Classification
Clustering
Association



3. Pengetahuan

(Pola/Model/Rumus/
Tree/Rule/Cluster)

Algoritma Data Mining

1. Estimation (Estimasi):

- Linear Regression, Neural Network, Support Vector Machine, etc

2. Prediction/Forecasting (Prediksi/Peramalan):

- Linear Regression, Neural Network, Support Vector Machine, etc

3. Classification (Klasifikasi):

- Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, Logistic Regression, etc

4. Clustering (Klastering):

- K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc

5. Association (Asosiasi):

- FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc

1. ALGORITMA KLASIFIKASI

- Klasifikasi (Han, 2006) adalah **proses penemuan model (atau fungsi)** yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui.
- Proses klasifikasi didasarkan pada empat komponen: (Gorunescu, 2011) yaitu :
 1. Kelas
 2. Predictor
 3. Training Dataset
 4. Testing Dataset

ALGORITMA KLASIFIKASI -2

- Kelas
- Variabel dependen yang berupa **kategorikal** yang merepresentasikan “label” yang terdapat pada objek.
- Contohnya: resiko penyakit jantung, resiko kredit, *customer loyalty, jenis gempa*.

ALGORITMA KLASIFIKASI -3

- Predictor
- Variabel independen yang direpresentasikan oleh **karakteristik** (atribut) data.
- Contohnya: Outlook, Temperature, Humidity dan Wind merupakan variabel untuk kelas play.

ALGORITMA KLASIFIKASI -4

- Training Dataset
- Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk menentukan **kelas yang cocok** berdasarkan *predictor*.
- Testing Dataset
- Berisi **data baru** yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi

ALGORITMA KLASIFIKASI - 5

- Hal-hal yang berhubungan dengan klasifikasi adalah :
 - ✓ Meramalkan kategori label kelas (nominal atau terpisah)
 - ✓ Menggolongkan data (membangun suatu model) yang didasarkan pada pelatihan, menetapkan nilai-nilai (label kelas) di (dalam) suatu penggolongan atribut dan penggunaan di dalam penggolongan data baru

ALGORITMA KLASIFIKASI -6

- Sedangkan aplikasi umum untuk Klasifikasi adalah :
 - Persetujuan kredit
 - Target marketing
 - Diagnosa medis
 - Analisis keefektifan tindakan

Algoritma Decision Tree

- **Algoritma Dasar** (a greedy algorithm)
 1. Tree dibangun dengan cara top-down recursive divide-and-conquer manner
 2. Pada awalnya, semua contoh training adalah akar (root)
 3. Atribut bersifat kategoris (jika dinilai terus-menerus, merek didiskualifikasi sebelumnya)
 4. Contoh dispartisi secara rekursif berdasarkan atribut yang dipilih
 5. Atribut uji dipilih berdasarkan ukuran heuristik atau statistik (mis., Perolehan informasi, rasio gain, indeks gini)

Tahapan Algoritma Decision Tree (ID3)

1. Siapkan **data training**
2. Pilih **atribut sebagai akar**

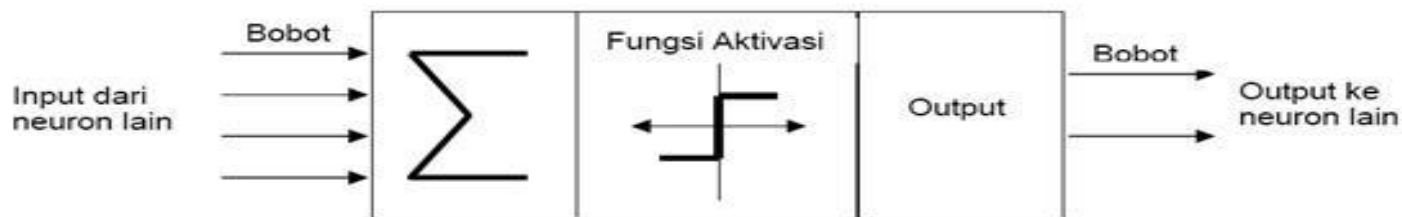
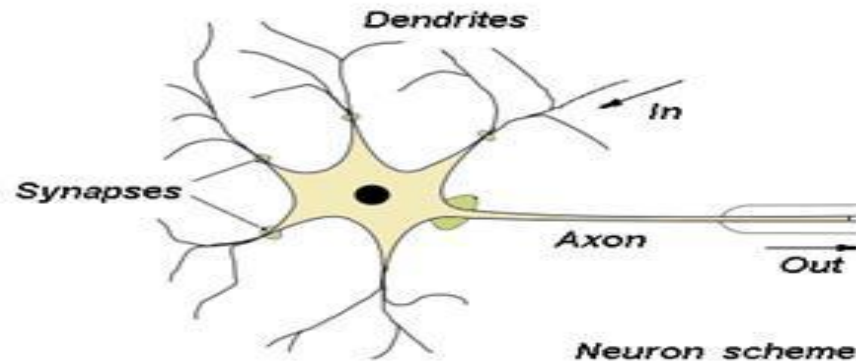
$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

3. Buat **cabang untuk tiap-tiap nilai**
4. **Ulangi proses** untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yg sama

Neural Network

- Neural Network adalah suatu model yang dibuat untuk **meniru fungsi belajar yang dimiliki otak manusia** atau jaringan dari sekelompok unit pemroses kecil yang dimodelkan berdasarkan jaringan saraf manusia



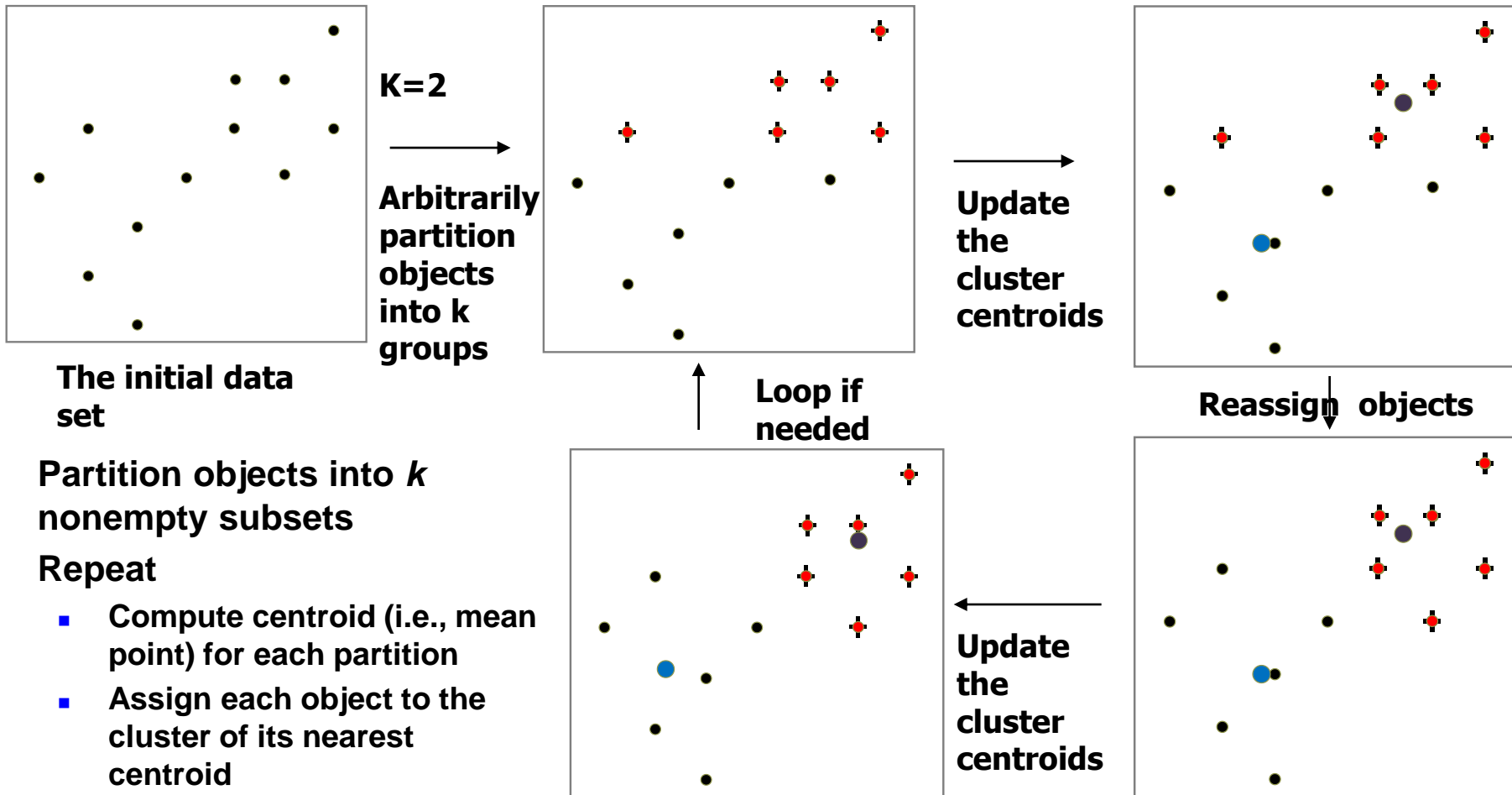
2. Analisis Cluster

- **Cluster**: Kumpulan Objek Data
 - serupa (atau terkait) satu sama lain dalam kelompok yang sama
 - berbeda (atau tidak terkait) dengan objek dalam kelompok lain
- **Cluster analysis** (atau clustering, segmentasi data, ...)
 - Menemukan kesamaan antara data sesuai dengan karakteristik yang ditemukan dalam data dan mengelompokkan objek data serupa ke dalam kelompok
- **Unsupervised learning**: tidak ada kelas yang telah ditentukan (mis., *learning by observations* vs. *learning by examples*: supervised)
- Typical applications
 - Sebagai alat yang berdiri sendiri untuk mendapatkan wawasan tentang distribusi data
 - Sebagai langkah preprocessing untuk algoritma lain

K-Means

- Diberikan k , algoritma k-means diimplementasikan dalam empat langkah :
 1. Partisi objek menjadi himpunan bagian nonempty
 2. Hitung titik seed sebagai centroid dari cluster dari partisi saat ini (centroid adalah pusat, mis., Titik rata-rata, dari cluster)
 3. Tetapkan setiap objek ke cluster dengan titik benih terdekat
 4. Kembali ke Langkah 2, berhenti ketika tugas tidak berubah

Contoh K-Means Clustering



- Partition objects into k nonempty subsets
- Repeat
 - Compute centroid (i.e., mean point) for each partition
 - Assign each object to the cluster of its nearest centroid
- Until no change

Tahapan Algoritma k-Means

1. Pilih **jumlah kluster k** yang diinginkan
2. **Inisialisasi k pusat kluster** (centroid) secara random
3. **Tempatkan setiap data atau objek ke kluster terdekat**. Kedekatan dua objek ditentukan berdasar jarak. Jarak yang dipakai pada algoritma k-Means adalah *Euclidean distance* (d)

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

– $x = x_1, x_2, \dots, x_n$, dan $y = y_1, y_2, \dots, y_n$ merupakan banyaknya n atribut(kolom) antara 2 record

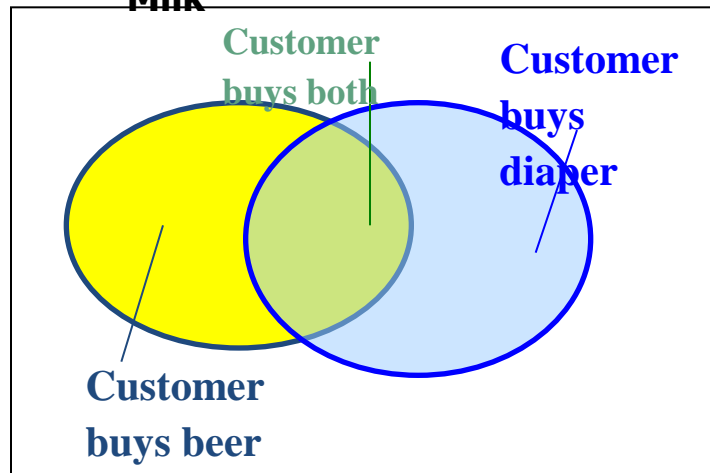
4. **Hitung kembali pusat kluster** dengan keanggotaan kluster yang sekarang. Pusat kluster adalah rata-rata (mean) dari semua data atau objek dalam kluster tertentu
5. **Tugaskan lagi setiap objek dengan memakai pusat kluster yang baru**. Jika **pusat kluster sudah tidak berubah lagi**, maka proses pengklasteran selesai. Atau, **kembali lagi ke langkah nomor 3** sampai pusat kluster tidak berubah lagi (stabil) atau tidak ada penurunan yang signifikan dari nilai SSE (*Sum of Squared Errors*)

3. ASOSIASI

- **Frequent pattern:** sebuah pola (satu set item, berikutnya, substruktur, dll.) Yang sering terjadi dalam kumpulan data
- **Pertama kali diusulkan Agrawal**, Imielinski, dan Swami [AIS93] dalam konteks frequent itemset dan asosiasi rule mining
- **Motivation:** Menemukan keteraturan yang melekat dalam data
 - Produk apa yang sering dibeli bersama? - Bir dan popok ?!
 - Apa pembelian selanjutnya setelah membeli PC?
 - Jenis DNA apa yang sensitif terhadap obat baru ini?
 - Bisakah kita secara otomatis mengklasifikasikan dokumen web?
- **Applications**
 - Analisis data keranjang, pemasaran silang, desain katalog, analisis kampanye penjualan, analisis log Web (aliran klik), dan analisis urutan DNA.

Basic Concepts: Association Rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- Find all the rules $X \rightarrow Y$ with minimum support and confidence
 - support, s , probability that a transaction contains $X \cup Y$
 - confidence, c , conditional probability that a transaction having X also contains Y

Let $minsup = 50\%$, $minconf = 50\%$

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

- Association rules: (many more!)
 - $Beer \rightarrow Diaper$ (60%, 100%)
 - $Diaper \rightarrow Beer$ (60%, 75%)