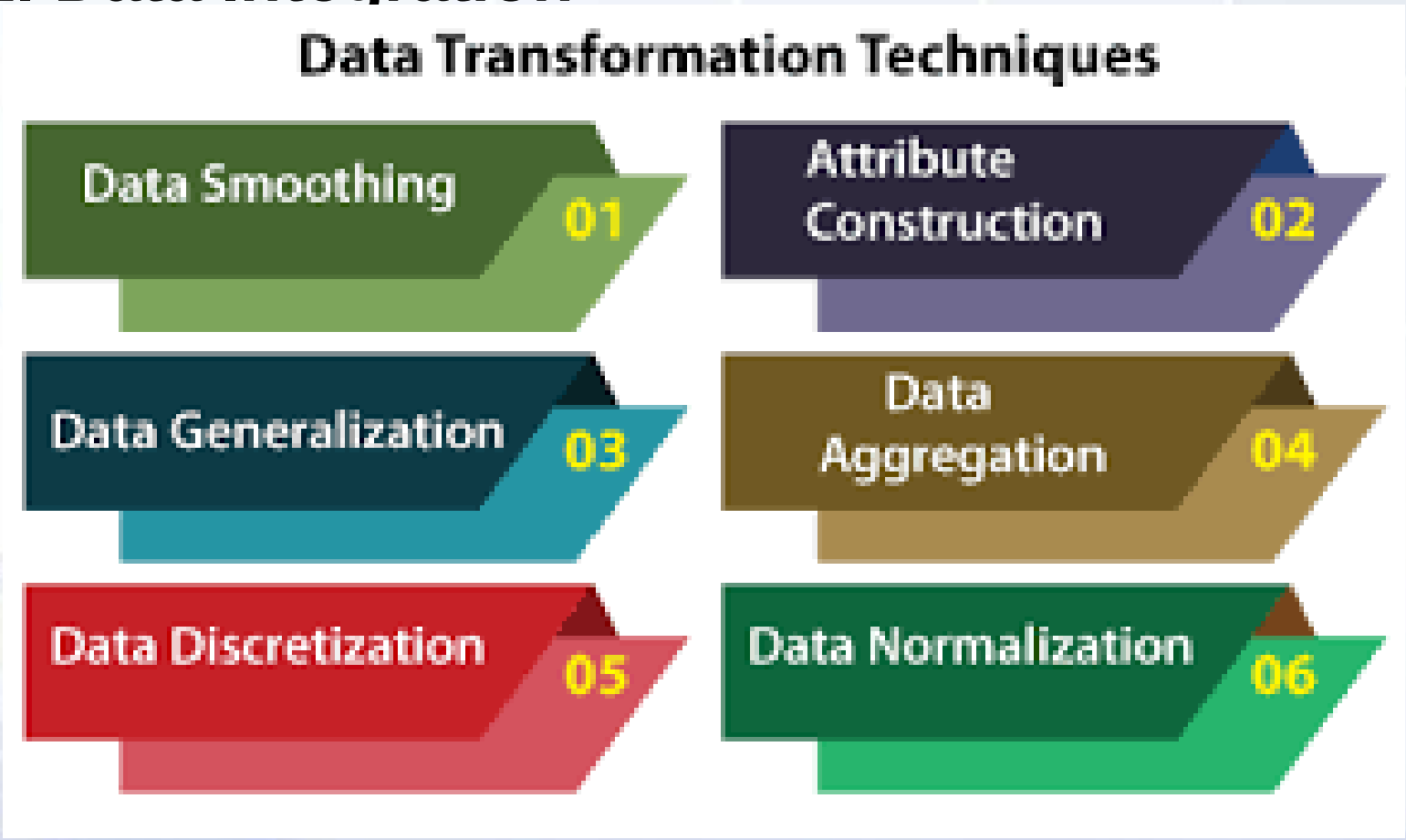
DATA MINING

PERTEMUAN Ke-6

Persiapan Data

I.2 Persiapan Data (Lanjutan)

1. Data Transformation and Data Discretization
2. Data Integration



Versi : 01

Data Transformation

- Sebuah fungsi yang memetakan seluruh himpunan nilai atribut tertentu ke himpunan nilai pengganti yang baru
- Setiap nilai lama dapat diidentifikasikan dengan salah satu nilai baru

Transformation Methods:

- Smoothing : Menghilangkan noise dari data
- Attribute/feature construction
  - Atribut baru dibangun dari atribut yang diberikan
- Aggregation : Peringkasan, konstruksi kubus data
- Normalization : Diskalakan agar berada dalam rentang yang lebih kecil dan ditentukan
  - normalisasi min-maks
  - normalisasi skor-z
  - normalisasi dengan skala desimal
- Normalization : Pendakian hierarki konsep

Normalization (1)

- Min-max normalization

Normalisasi min-max adalah metode dalam analisis data yang digunakan untuk mengubah nilai-nilai dalam dataset menjadi rentang nilai tertentu, biasanya antara 0 dan 1. Metode ini membantu dalam menyesuaikan skala nilai-nilai yang ada sehingga mereka memiliki rentang yang seragam dan dapat dibandingkan atau diproses dengan lebih baik.

Normalization (2)

- Z-score normalization, juga dikenal sebagai standard score normalization, adalah teknik statistik yang digunakan untuk mentransformasi dan mengubah distribusi dari variabel-variabel dalam dataset menjadi distribusi normal dengan mean (rata-rata) 0 dan standar deviasi 1. Teknik ini mengubah setiap nilai dalam dataset menjadi nilai yang menggambarkan berapa standar deviasi jauhnya nilai tersebut dari rata-rata.

Normalization (3)

- Normalization by decimal scaling adalah metode normalisasi lain yang digunakan untuk mengubah nilai-nilai dalam dataset ke rentang tertentu dengan membagi setiap nilai dengan faktor skalar yang sesuai. Teknik ini menggeser titik desimal nilai-nilai tersebut sehingga nilai terbesar pada dataset diperoleh dalam rentang [-1, 1], atau [-0.1, 0.1], atau rentang lainnya tergantung pada kebutuhan.

Normalization

- Min-max normalization: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that Max(|v'|) < 1

Versi : 01

Discretization (1)

- **Tiga jenis atribut**
- Nominal —nilai dari himpunan tak berurutan, misalnya warna kulit, profesi
- Ordinal —nilai dari himpunan terurut, misalnya pangkat militer atau akademis
- Numerik —bilangan real, misalnya bilangan bulat atau bilangan real

Discretization (2)

- Diskritisasi: Membagi rentang atribut kontinu menjadi beberapa interval
  - Interval labels kemudian dapat digunakan untuk menggantikan nilai data sebenarnya
  - Reduce data size dengan diskritisasi
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization dapat dilakukan secara rekursif pada suatu atribut
  - Persiapan untuk analisis lebih lanjut, misalnya klasifikasi

unsupervised

- "Unsupervised" mengacu pada jenis teknik atau pendekatan dalam pembelajaran mesin di mana algoritma atau model tidak menerima label atau informasi target yang terkait dengan data yang sedang diproses. Dalam skenario ini, model diberikan dataset tanpa informasi tentang output yang diharapkan.
- Unsupervised learning berguna ketika tidak ada label yang tersedia atau memungkinkan untuk mengeksplorasi struktur intrinsik dari data. Contohnya termasuk analisis pola pasar, segmentasi pelanggan, analisis teks untuk temuan pola topik, atau pengelompokkan gen dalam biologi.

Contoh teknik unsupervised learning termasuk:

1. **Clustering:** Algoritma seperti K-Means, Hierarchical Clustering, atau DBSCAN digunakan untuk mengelompokkan data ke dalam cluster berdasarkan kesamaan fitur atau atribut.
2. **Reduksi Dimensi:** Metode seperti Principal Component Analysis (PCA) atau t-SNE (t-Distributed Stochastic Neighbor Embedding) digunakan untuk mengurangi dimensi dari dataset dengan tetap mempertahankan informasi penting.
3. **Asosiasi:** Algoritma seperti Apriori digunakan untuk menemukan hubungan atau asosiasi antara item dalam dataset transaksional.

supervised

- "Supervised learning" adalah paradigma dalam pembelajaran mesin di mana model atau algoritma belajar dari data yang dilengkapi dengan label atau informasi target yang sesuai. Dalam supervised learning, data yang digunakan untuk melatih model terdiri dari pasangan input dan output yang sudah ditentukan sebelumnya.
- Tujuan utama dari supervised learning adalah untuk mempelajari hubungan atau pola yang ada antara input dan output agar model dapat membuat prediksi atau mengidentifikasi output yang tepat untuk input baru yang belum pernah dilihat.

Proses dalam supervised learning terdiri dari dua tahap utama:

1. **Training (Pelatihan):** Model atau algoritma dilatih menggunakan dataset yang memiliki pasangan input-output yang sudah diberi label. Model belajar untuk menghubungkan input dengan output yang sesuai.
2. **Testing (Pengujian):** Setelah model dilatih, model diuji menggunakan data yang tidak pernah dilihat sebelumnya untuk mengevaluasi seberapa baik model tersebut mampu membuat prediksi yang akurat.

Contoh-contoh aplikasi supervised learning meliputi:

- **Klasifikasi:** Memisahkan data ke dalam kategori atau kelas yang berbeda. Contohnya, mengklasifikasikan email sebagai spam atau bukan spam.
- **Regresi:** Memprediksi nilai kontinu berdasarkan input tertentu. Contohnya, memprediksi harga rumah berdasarkan fitur-fitur seperti luas tanah, jumlah kamar, dll.
- **Deteksi Objek:** Mengidentifikasi objek dalam gambar atau video, misalnya, mengenali mobil, orang, atau hewan.

Data Discretization Methods

- Metode umum : Semua metode dapat diterapkan secara rekursif
  - Binning: Top-down split, unsupervised
  - Histogram analysis: Top-down split, unsupervised
  - Clustering analysis: Unsupervised, top-down split or bottom-up merge
  - Decision-tree analysis: Supervised, top-down split
  - Correlation (e.g., χ^2) analysis: Unsupervised, bottom-up merge

Simple Discretization: Binning(1)

- Equal-width (distance) partitioning
  - Membagi rentang menjadi N interval dengan ukuran yang sama: kisi seragam
  - jika A dan B adalah nilai atribut terendah dan tertinggi, maka lebar intervalnya adalah: W = (B –A)/N.
  - Yang paling lugas, namun hal-hal lain mungkin mendominasi presentasi
  - Data yang miring tidak ditangani dengan baik

Simple Discretization: Binning(2)

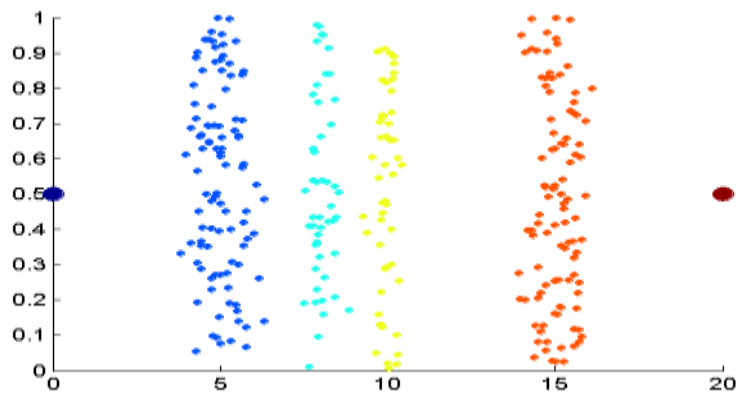
- Equal-depth (frequency) partitioning
  - Membagi rentang menjadi N interval, masing-masing berisi jumlah sampel yang kira-kira sama
  - Penskalaan data yang baik
  - Mengelola atribut kategorikal bisa jadi rumit

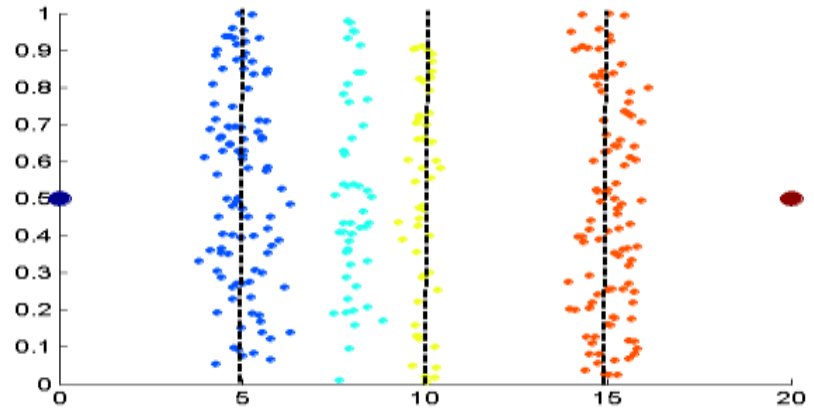
Binning Methods for Data Smoothing

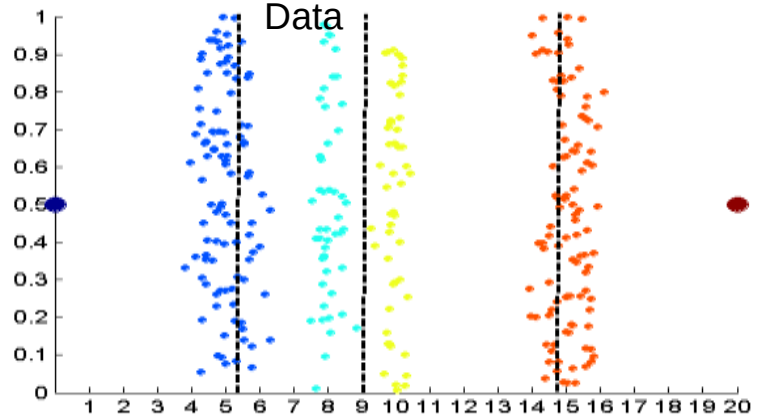
Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

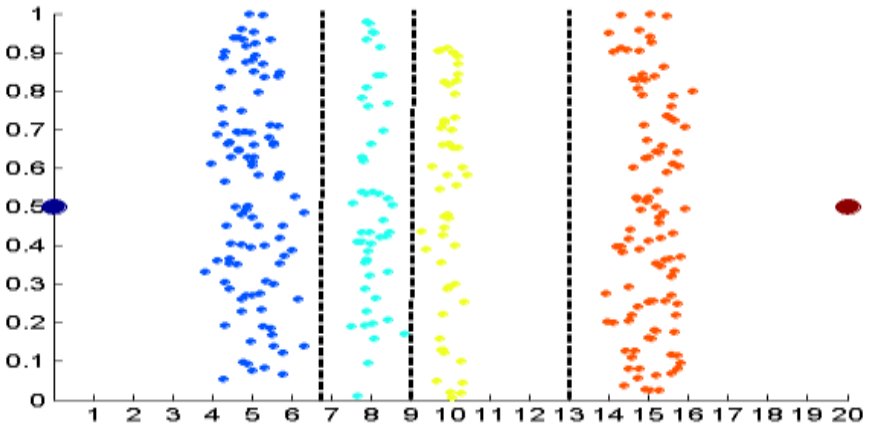
Discretization Without Using Class Labels (Binning vs. Clustering)







Equal frequency (binning)



K-means clustering leads to better results

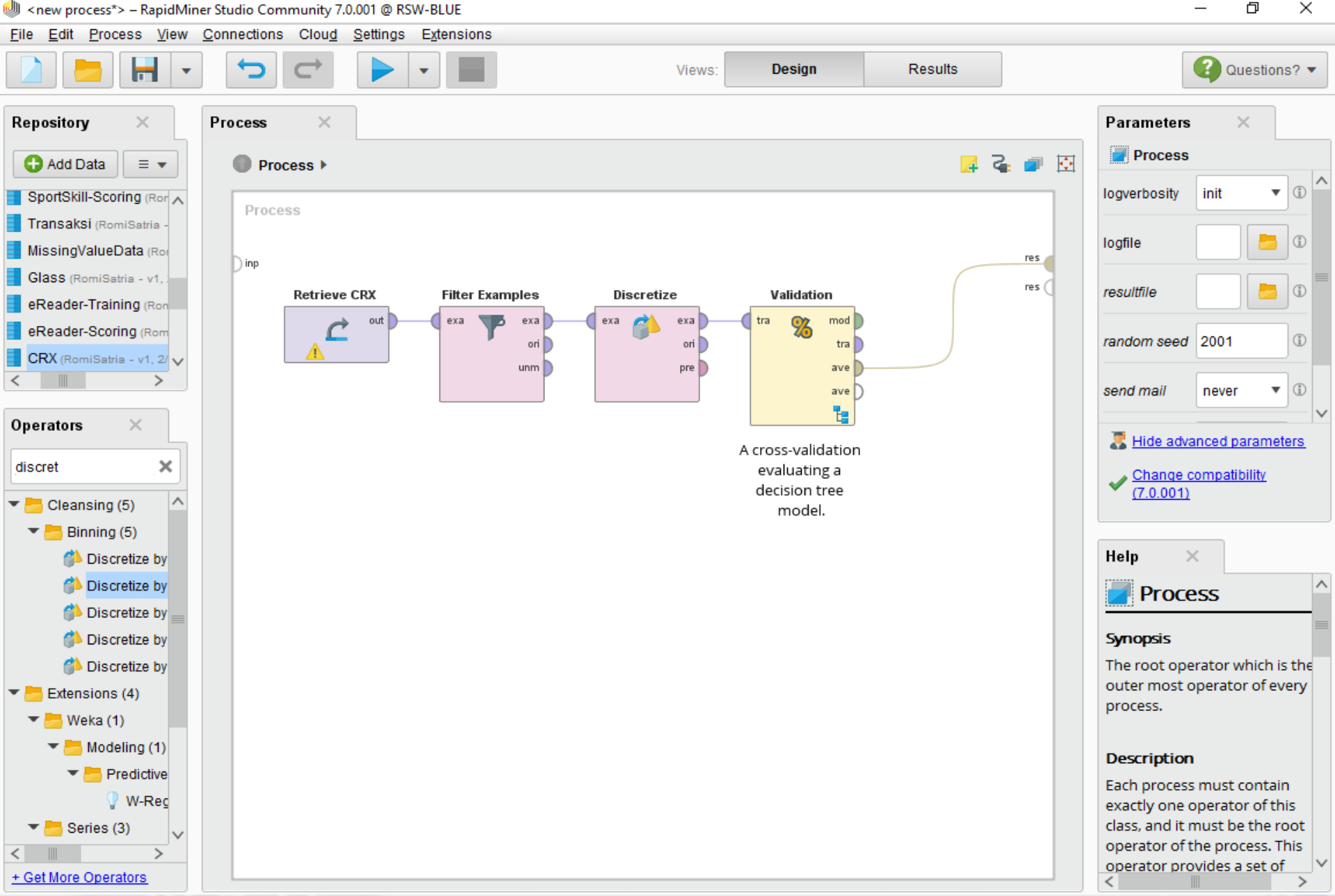
Versi : 01

Discretization by Classification & Correlation Analysis(1)

- Classification (e.g., decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Gunakan *entropy* untuk menentukan *split point* (discretization point)
  - Top-down, recursive split
- Correlation analysis (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge : temukan interval best neighboring (yang memiliki distribusi kelas serupa, yaitu nilai $\chi 2$ rendah) untuk digabungkan
  - Penggabungan dilakukan secara rekursif, hingga kondisi penghentian yang telah ditentukan

Latihan

- Lakukan eksperimen mengikuti buku Markus Hofmann (Rapid Miner - Data Mining Use Case) Chapter 5 (Naïve Bayes Classification I)
- Dataset: crx.data
- Analisis metode preprocessing apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut!
- Bandingkan akurasi model apabila tidak menggunakan filter dan diskretisasi
- Bandingkan pula apabila digunakan feature selection (wrapper) dengan Backward Elimination


<new process*> – RapidMiner Studio Community 7.0.001 @ RSW-BLUE
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Questions?
Repository
Add Data
SportSkill-Scoring
Transaksi (RomiSatria
MissingValueData
Glass (RomiSatria - v1,
eReader-Training
eReader-Scoring
CRX (RomiSatria - v1,
Operators
discret
Cleansing (5)
Binning (5)
Discretize by
Discretize by
Discretize by
Discretize by
Discretize by
Extensions (4)
Weka (1)
Modeling (1)
Predictive
W-Reg
Series (3)
+ Get More Operators
Process
Process
Process
inp
Retrieve CRX
out
Filter Examples
exa
exa
ori
unm
Discretize
exa
exa
ori
pre
Validation
tra
mod
tra
ave
ave
res
res
A cross-validation evaluating a decision tree model.
Parameters
Process
logverbosity
init
logfile
resultfile
random seed
2001
send mail
never
Hide advanced parameters
Change compatibility (7.0.001)
Help
Process
Synopsis
The root operator which is the outer most operator of every process.
Description
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of

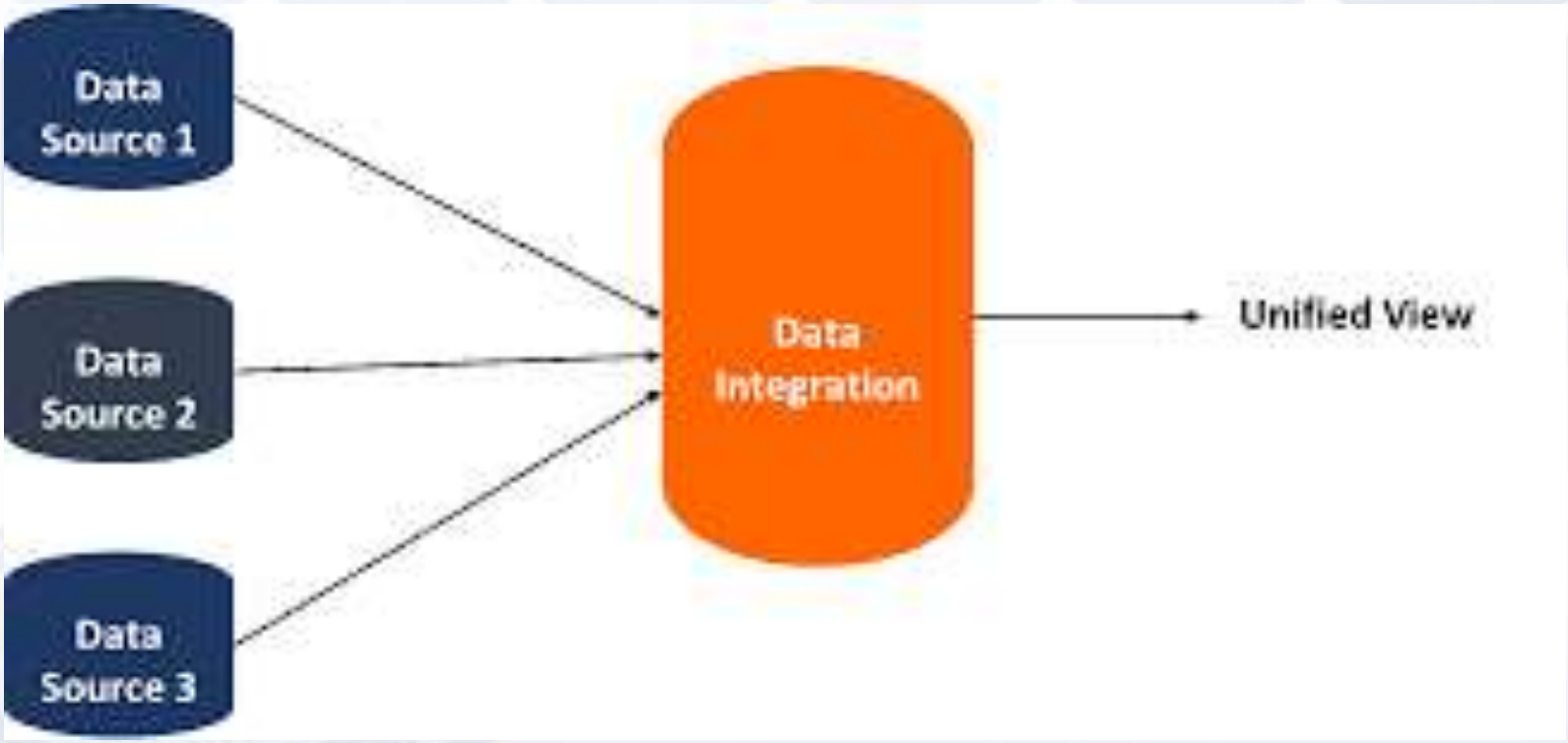
Versi : 01

Hasil

	NB	NB+ Filter	NB+ Discretization	NB+ Filter+ Discretization	NB+ Filter+ Discretization + Backward Elimination
Accuracy				85.79	86.26
AUC					

2. Persiapan Data

2. Data Integration



Data Integration

- Data integration:
  - Menggabungkan data dari berbagai sumber menjadi penyimpanan yang koheren
- Schema Integration: e.g., A.cust-id ≡ B.cust-#
  - Integrasikan metadata dari berbagai sumber
- Entity Identification Problem:
  - Identifikasi entitas dunia nyata dari berbagai sumber data, misalnya Bill Clinton = William Clinton
- Detecting and Resolving Data Value Conflicts
  - Untuk entitas dunia nyata yang sama, nilai atribut dari sumber berbeda berbeda
  - Kemungkinan alasannya: representasi berbeda, skala berbeda, misalnya satuan metrik vs. Inggris

Data Integration

- Data integration adalah proses menggabungkan data dari berbagai sumber yang berbeda menjadi satu set data yang lebih terpadu, lengkap, dan bermakna. Tujuan utamanya adalah untuk menciptakan gambaran yang lebih komprehensif dan dapat dipercaya dari data, yang bisa digunakan untuk analisis, pengambilan keputusan, pelaporan, dan aplikasi lainnya.

Proses integrasi data

1. **Ekstraksi:** Mengumpulkan data dari berbagai sumber, termasuk basis data, aplikasi, file teks, sensor, dan lainnya.
2. **Transformasi:** Menyelaraskan format, struktur, dan nilai data agar sesuai dengan format yang konsisten dan standar, serta memastikan integritasnya.
3. **Pembersihan (Cleansing):** Mengidentifikasi dan memperbaiki data yang tidak akurat, tidak lengkap, atau tidak konsisten.
4. **Penggabungan (Merging):** Menggabungkan data dari berbagai sumber menjadi satu set data tunggal.
5. **Pemuatan (Loading):** Menyimpan data yang telah diintegrasikan ke dalam penyimpanan data yang sesuai, seperti data warehouse atau data lake, untuk digunakan oleh sistem analisis atau aplikasi lainnya.

I.29 Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - Object identification: The same attribute or object may have different names in different databases
  - Derivable data: One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by correlation analysis and covariance analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Versi : 01

Correlation Analysis (Nominal Data)

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- Semakin besar nilai *X2* maka semakin besar kemungkinan variabel-variabel tersebut berhubungan
- Sel-sel yang memberikan kontribusi paling besar terhadap nilai X2 adalah sel-sel yang jumlah aktualnya sangat berbeda dengan jumlah yang diharapkan
- Korelasi tidak berarti kausalitas
  - Jumlah rumah sakit dan jumlah pencurian mobil di suatu kota berkorelasi
  - Keduanya secara kausal terkait dengan variabel ketiga: populasi

Correlation Analysis (Nominal Data)

- Analisis korelasi adalah metode statistik yang digunakan untuk mengukur hubungan antara dua variabel.
- Tujuan utamanya adalah untuk menentukan sejauh mana variabel-variabel tersebut bergerak bersamaan atau berhubungan satu sama lain.
- Metode ini mengukur kekuatan dan arah hubungan antara variabel-variabel tersebut.

Korelasi diukur dalam rentang antara -1 hingga 1:

- Jika korelasi mendekati 1, itu menunjukkan hubungan positif yang kuat. Artinya, ketika satu variabel meningkat, kemungkinan besar variabel lainnya juga meningkat.

- Jika korelasi mendekati -1, itu menunjukkan hubungan negatif yang kuat. Artinya, ketika satu variabel meningkat, kemungkinan besar variabel lainnya akan menurun.

- Jika korelasi mendekati 0, itu menunjukkan hubungan yang lemah atau tidak ada hubungan sama sekali antara kedua variabel.

Chi-Square Calculation(1)

- Perhitungan *Chi-Square* adalah teknik statistik yang digunakan untuk menentukan apakah ada hubungan antara dua variabel kategorikal (non-numerik). Ini adalah metode yang berguna untuk menentukan apakah hubungan antara dua variabel tersebut bersifat independen atau saling terkait.
- Perhitungan *Chi-Square* melibatkan pembuatan tabel kontingensi, yang menunjukkan frekuensi pengamatan untuk kombinasi kategori dari kedua variabel yang sedang dianalisis. Chi-Square dihitung dengan rumus:

Chi-Square Calculation(2)

- Chi-Square dihitung dengan rumus:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- Dalam rumus tersebut:
- χ^2 adalah nilai Chi-Square yang dihitung.
- O_i adalah frekuensi pengamatan aktual untuk setiap sel dalam tabel kontingensi.
- E_i adalah frekuensi yang diharapkan untuk setiap sel jika kedua variabel adalah independen.

Chi-Square Calculation(3)

Langkah-langkah perhitungan Chi-Square melibatkan:

1. Membuat tabel kontingensi dari data yang diamati.
2. Menghitung nilai yang diharapkan (E_i)untuk setiap sel dalam tabel jika variabel-variabel tersebut adalah independen. Ini sering dilakukan dengan menggunakan rumus (E_i =total baris×total kolom/total data) untuk distribusi probabilitas.
3. Menghitung selisih antara frekuensi pengamatan aktual (*Oi*) dan frekuensi yang diharapkan (*Ei*).
4. Mengkuadratkan selisih tersebut, membaginya dengan frekuensi yang diharapkan (*Ei*), dan menjumlahkannya untuk mendapatkan nilai Chi-Square total.

Chi-Square Calculation: An Example

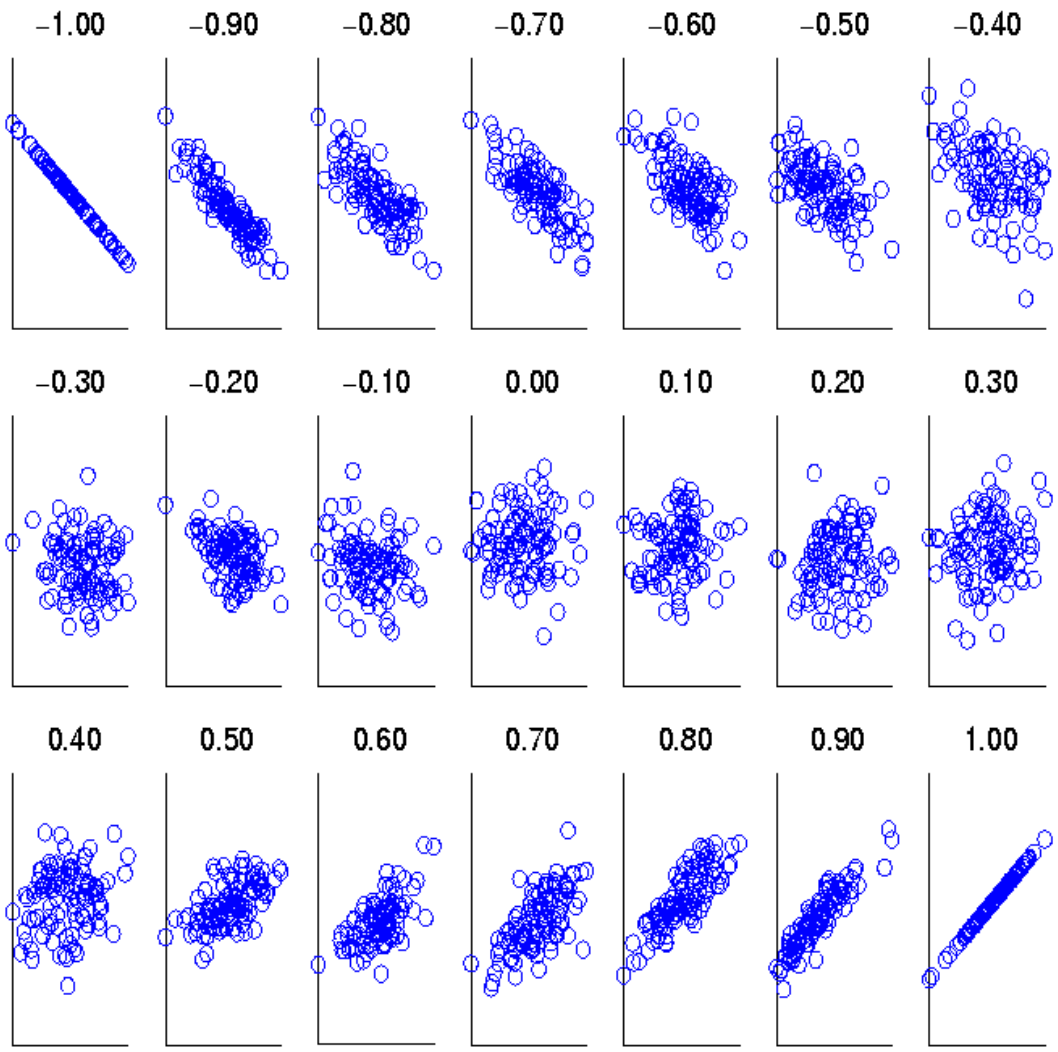
	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 perhitungan (Chi-square) (angka dalam tanda kurung adalah jumlah yang diharapkan dihitung berdasarkan sebaran data pada dua kategori)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Hal ini menunjukkan bahwa like_science_fiction dan play_chess berkorelasi dalam grup

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1

Versi : 01

Correlation

- Korelasi mengukur hubungan linier antar objek
- Untuk menghitung korelasi, kita menstandarisasikan objek data, A dan B, lalu mengambil perkalian titiknya

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A' \bullet B'$$

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Correlation coefficient: $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Covariance: An Example

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Pertanyaan: Jika saham dipengaruhi oleh tren industri yang sama, apakah harganya akan naik atau turun secara bersamaan?
- E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0

Rangkuman

1. Data quality: accuracy, completeness, consistency, timeliness, believability, interpretability
2. Data cleaning: e.g. missing/noisy values, outliers
3. Data reduction
   - Dimensionality reduction
   - Numerosity reduction
4. Data transformation and data discretization
   - Normalization
5. Data integration from multiple sources:
   - Entity identification problem
   - Remove redundancies
   - Detect inconsistencies

Review dan Latihan

☺ END ☺