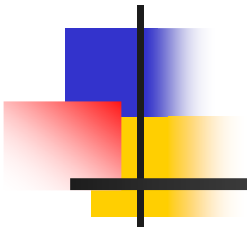


Data Preprocessing

Atribut Data

- Mencerminkan karakteristik obyek data
- Tipe atribut menentukan himpunan nilai yang diperbolehkan
  - Nominal
  - Binary
  - Ordinal
  - Numerik
  - Diskret atau Continue

Mengapa Perlu Data Preprocessing?

- Data dalam dunia nyata "dirty"
  - Tidak komplet: berisi data yang hilang/kosong, kekurangan atribut yang sesuai, hanya berisi data aggregate
    - e.g., occupation=" "
  - Banyak "noise": berisi data yang outlier atau error
    - e.g., Salary="-10"
  - Tidak konsisten: berisi nilai yang berbeda dalam suatu kode atau nama
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records

Mengapa Data Preprocessing Penting?

- Data yang tidak berkualitas, akan menghasilkan kualitas mining yang tidak baik pula.
- Data Preprocessing, cleanning, dan transformasi merupakan pekerjaan mayoritas dalam aplikasi data mining (90%).

Ukuran Kualitas

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Value added
- Interpretability
- Accessibility

Data Cleaning

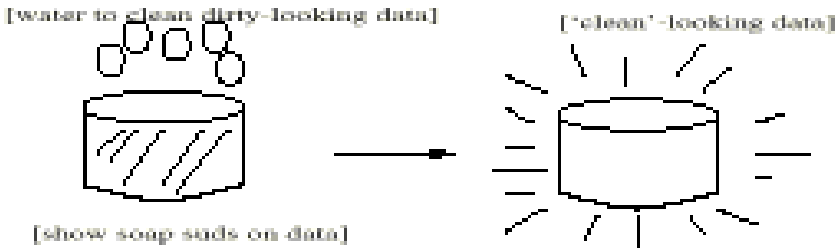
- Proses untuk membersihkan data dengan beberapa teknik
  - Memperkecil noise
  - Membetulkan data yang tidak konsisten
  - Mengisi missing value
  - Mengidentifikasi atau membuang outlier

Teknik Data Preprocessing

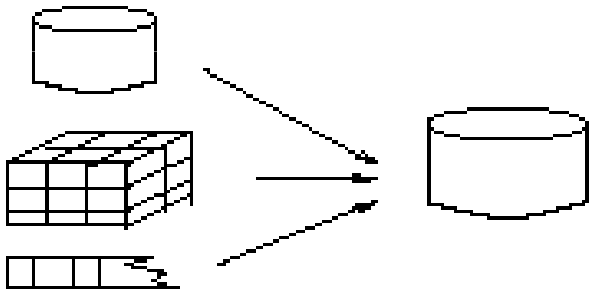
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation

Bentuk dari Data Preprocessing

Data Cleaning



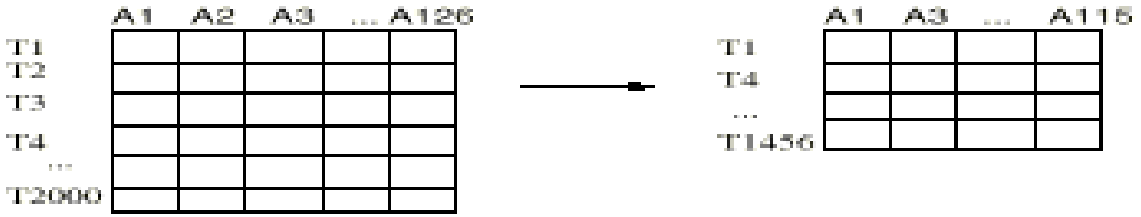
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Cleaning: Missing Values

- Mengabaikan record
  - Biasanya untuk label klasifikasi yang kosong
- Mengisikan secara manual
- Menggunakan mean/median dari atribut yang mengandung missing value
  - Mean dapat dipakai jika distribusi data normal
  - Median digunakan jika distribusi data tidak normal (condong)
- Menggunakan nilai global
- Menggunakan nilai termungkin
  - Menerapkan regresi

Data Cleaning: Missing Values

- Mengabaikan record
  - Biasanya untuk label klasifikasi yang kosong
- Mengisikan secara manual
- Menggunakan mean/median dari atribut yang mengandung missing value
  - Mean dapat dipakai jika distribusi data normal
  - Median digunakan jika distribusi data tidak normal (condong)
- Menggunakan nilai global
- Menggunakan nilai termungkin
  - Menerapkan regresi

Metoda Binning: Diskritisasi Sederhana

- Partisi kedalaman sama (frekuensi):
  - Membagi range kedalam *N interval, masing-masing* memuat jumlah sampel yang hampir sama
  - Penskalaan data yang baik
  - Penanganan atribut yang bersifat kategori bisa rumit.

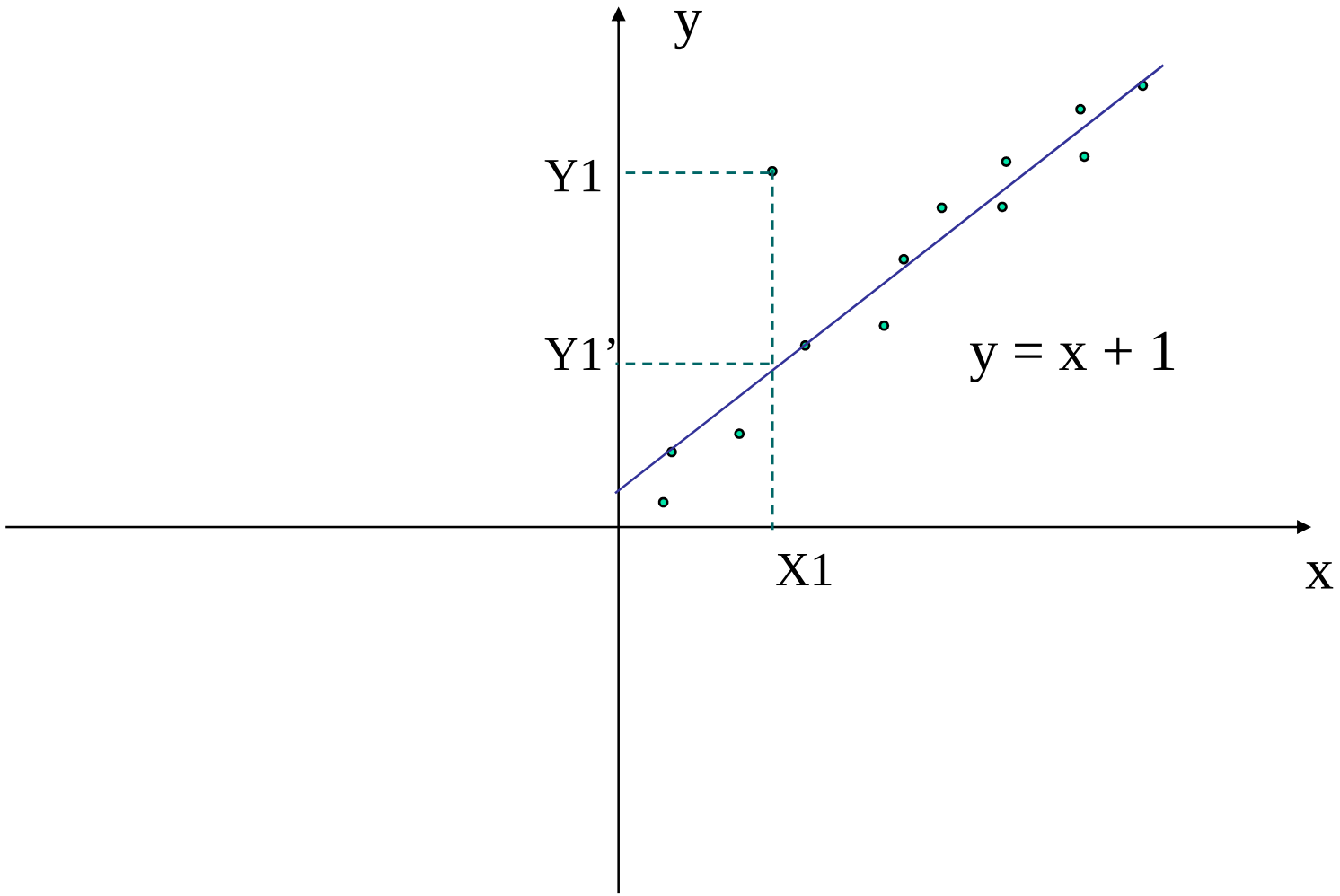
Metoda Binning: Diskritisasi Sederhana

- Data terurut untuk harga (dalam dollar): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
  - Partisi kedalam bin dengan kedalaman yang sama (misal, dalam bin-3):
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
  - Haluskan dengan rata-rata bins:
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29

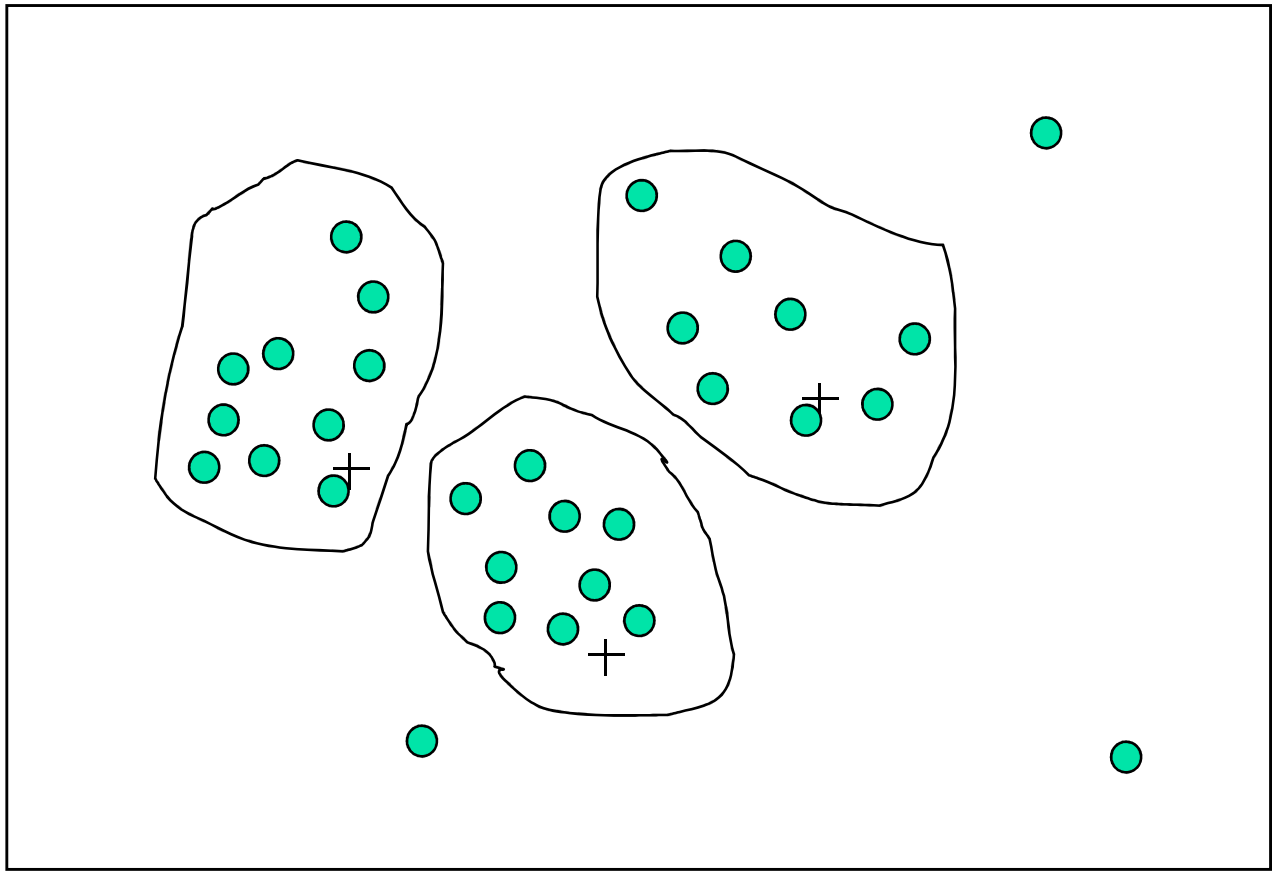
Metoda Binning: Diskritisasi Sederhana

- Penghalusan dengan batas bin:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Regression



Cluster Analysis



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*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Normalization

- adalah proses penskalaan nilai atribut dari data sehingga bisa jatuh pada range tertentu.
- Contoh: Misalnya berkenaan dengan pencatatan tingkat kematian penduduk di Indonesia per bulannya berdasarkan jenis umur. Secara sederhana, disana ada 3 dimensi data, yaitu bulan (1-12), umur (0-150 misalnya), dan jumlah kematian (0-jutaan). Kalau kita bentangkan range masing-masing dimensi, maka kita akan mendapatkan ketidak-seimbangan range pada dimensi yang ketiga, yaitu jumlah kematian.

Normalization methods

- Min-Max
- Z-Score
- Decimal Scaling
- Sigmoidal
- Softmax

Normalization method (Min-Max)

- Min-Max merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli.
- Rumus:
  newdata = (data-min)*(newmax-newmin)/(max-min)+newmin
- Keuntungan dari metode ini adalah keseimbangan nilai perbandingan antar data saat sebelum dan sesudah proses normalisasi. Tidak ada data bias yang dihasilkan oleh metode ini. Kekurangannya adalah jika ada data baru, metode ini akan memungkinkan terjebak pada "out of bound" error.

Normalization method (Z-Score)

- Z-score merupakan metode normalisasi yang berdasarkan mean (nilai rata-rata) dan standard deviation (deviasi standar) dari data.
- Rumus:

  newdata = (data-mean)/std

- Metode ini sangat berguna jika kita tidak mengetahui nilai aktual minimum dan maksimum dari data.

Normalization Method (Decimal Scaling)

- Metode ini melakukan normalisasi dengan menggerakkan nilai desimal dari data ke arah yang diinginkan.

- Rumus:

$$newdata = data / 10^{i}$$

dimana i adalah nilai integer untuk menggerakkan nilai desimal ke arah yang diinginkan.

Normalization method (Sigmoidal)

- Sigmoidal normalization melakukan normalisasi data secara nonlinier ke dalam range -1 - 1 dengan menggunakan fungsi sigmoid.

- Rumus:

  newdata = (1-e^(-x))/(1+ e^(-x))

  dimana:

  x = (data-mean)/std

  e = nilai eksponensial (2,718281828)

- Metode ini sangat berguna pada saat data-data yang ada melibatkan data-data outlier.

Normalization method (Softmax)

- Metode ini merupakan pengembangan transformasi secara linier. Output range-nya adalah 0-1.

- Rumus:

  newdata = 1/(1+e^(-transfdata))

  dimana:

  transfdata = (data-mean)/(x*(std/(2*3.14)))

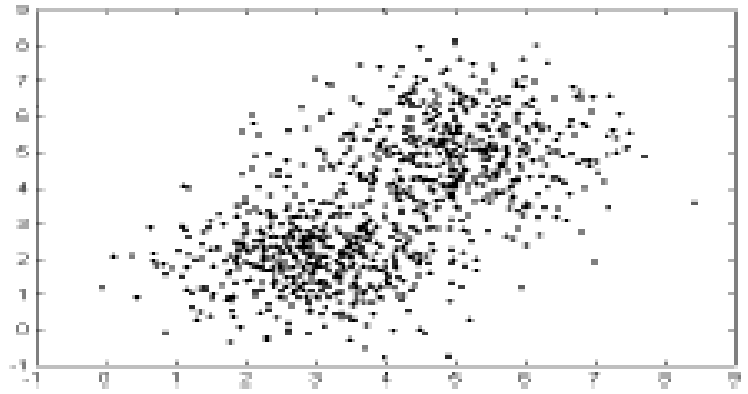
  x = respon linier di deviasi standar

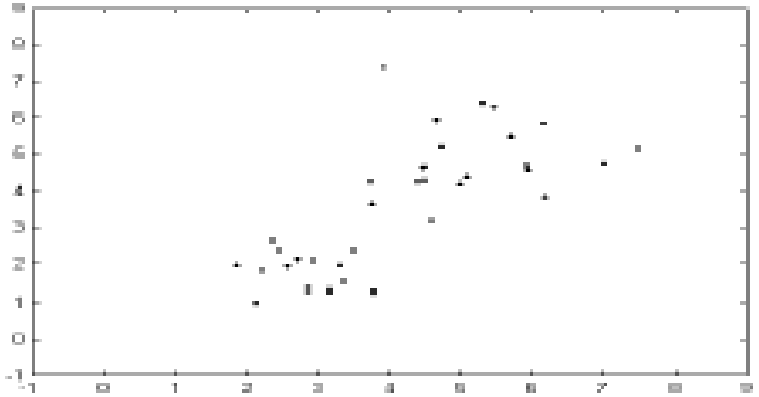
Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Discriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but data preprocessing still an active area of research

Algoritma Reduksi Data

- Pemilihan sub sampel dari data merupakan hal yang biasa dilakukan dalam proses data mining.
- Pemilihan sub sampel dilakukan pada data yang mendekati pada model data sesungguhnya

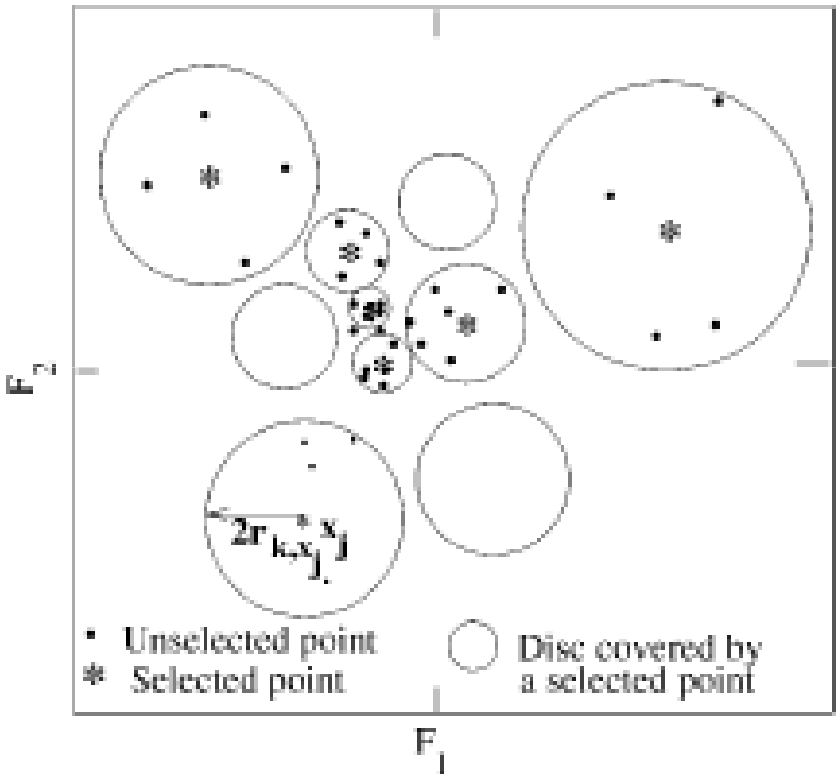




Density-Based Multiscale Data Condensation

- Prinsip dasar algoritma ini adalah mengurutkan titik-titik berdasarkan estimated densities, memilih titik-titik yang padat, dan menghapus titik lain yang berada dalam jarak tertentu dari titik yang dipilih sebagai bagian dari sampel data.
- Metode non-parametrik dalam memperkirakan probability density function adalah metode k-nearest-neighbor.
- Pada metode k-NN kepadatan titik dihitung berdasarkan area pada suatu lingkaran yang berisi sejumlah k titik yang berketetanggaan.

Density-Based Multiscale Data Condensation (cont.)



Density-Based Multiscale Data Condensation (cont.)

Himpunan $B_N = \{x_1, x_2, \ldots, x_N\}$ sebagai data set inputan. Pilih nilai integer positif k

- Untuk tiap titik $x_1 \in B_N$ hitung jarak k^{th} nearest neighbor dari x_1 pada B_N. Tandai dengan $r_{k,xi}$.
- Pilih $x_j \in B_N$ yang mempunyai $r_{k,xj}$ terkecil dan letakkan pada himpunan E.
- Hapus semua titik dari B_N yang berada dalam lingkaran radius 2 $r_{k,xj}$ yang berpusat di x_j dan titik-titik yang tersisa di himpunan sebagai B_N.