Preparing Data

Representation

- Data usually has many features
- Feature type
  - Numerical (has order and distance relation)
  - Categorical/symbolic
  - Continuous (quantitative)
    - Interval scale (zero is places arbitrarily)
      - Ex: temperature
    - Ratio Scale (zero has absolut position)
      - Ex: height, weight

Representation

- Discrete (qualitative)
  - Nominal scale (order-less scale)
    - Ex: customer type
  - Ordinal scale (ordered, discrete graduation)
    - Ex: rank
  - Periodic variable (has order but no distance)
    - Ex: days of the week, month or year
  - Static Data (does not change with time)
  - Dynamic Data (change with time)

Raw Data for Mining

- Usually large in quantities
- Problems:
  - missing,
  - wrong (misrecorded)
- Needs transformation to be usable and useful for data mining application

Transformation

- Normalization
  - To a specific range such as [-1,1]
- Decimal Scaling
  - A typical scaling is [-1,1]
  - $V'(i) = V(i) / 10^k$

    for the smallest k such that $|V'(i)| < 1$
  - What is k if the largest value in the data set is 445 and the smallest value is -834 ?

Transformation

- Min-max normalization
  - V'(i) = (V(i) – min(V(i)) / (max(V(i))-min(V(i)))
  - Supose the data is [25,67,150, 200]
    - Min(V(i)) = 25
    - Max(V(i)) = 200
    - 25 → (25 – 25) / (200 – 25) = 0
    - 200 → (200 – 25) / (200 – 25) = 1

Transformation

- Standard Deviation Normalization
  - V'(i) = (V(i) – mean(V)) / sd(V)
  - Supose the data is [1, 2, 3]
    - Mean (V) = 2
    - Sd (V) = 1
    - Transformed data → [-1, 0, 1]

Data Smoothing

- Minor differences sometimes is not significant and will not degrade the performance and result of data mining
- F = { 0.93, 1.01, 1.001, 3.02, 4.98}
- $F_{smoothed}$ = {1.0, 1.0, 1.0, 3.0, 5.0}

Differences and Ratios

- Performances are often measured in differences or ratio rather then the output value.
- Instead of stating v(t+1) as the result, we can show the performance as v(t+1) – v(t) or v(t +1)/vt(t)

Missing Data

- Replace with a single constant
- Replace using feature mean
- Replace using feature mean for the given class
- Replace using the possible combination of values
- X = {1, ?, 3, 4, 5}
  - ? = 0
  - ? = 3
  - ? = 1, 2, 3, 4, 5 → X_1, X_2, X_3, X_4, X_5

Time-dependent Data

- X = {t(1), t(2), t(3), …., t(n)}
- What is t(n+1) ?
- Or sometimes at some time lag, j, t(n+j)
- X= {t(0),t(1),t(2),t(3),t(4),t(5),t(6),t(7),t(8),t(9),t(10)}

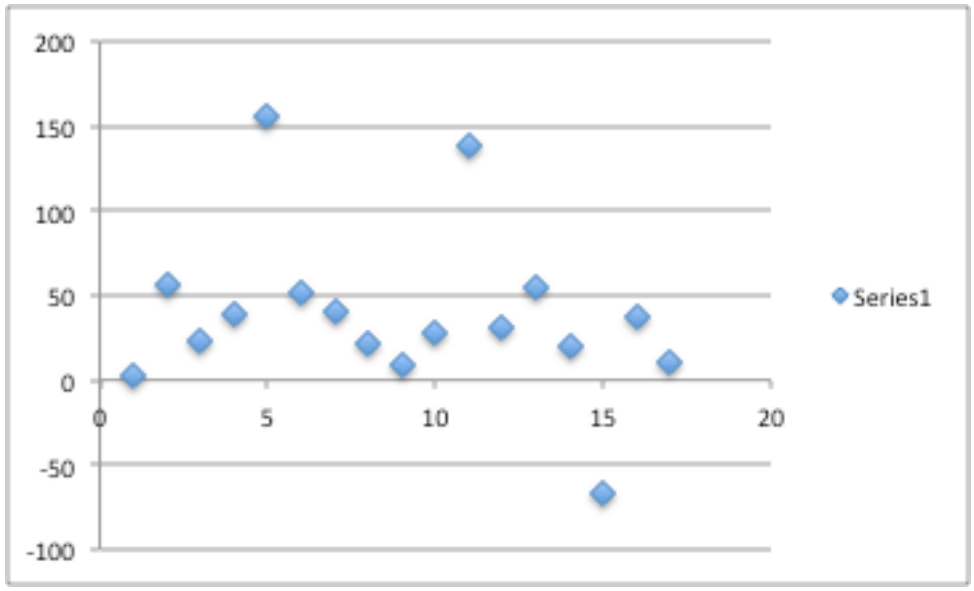
Time-dependent Data

Sample			Window			Next Value
	M1	M2	M3	M4	M5	
1	t(0)	t(1)	t(2)	t(3)	t(4)	t(5)
2	t(1)	t(2)	t(3)	t(4)	t(5)	t(6)
3	t(2)	t(3)	t(4)	t(5)	t(6)	t(7)
4	t(3)	t(4)	t(5)	t(6)	t(7)	t(8)
5	t(4)	t(5)	t(6)	t(7)	t(8)	t(9)
6	t(5)	t(6)	t(7)	t(8)	t(9)	t(10)

Sample			Window			Next Value
	M1	M2	M3	M4	M5	
1	t(0)	t(1)	t(2)	t(3)	t(4)	t(7)
2	t(1)	t(2)	t(3)	t(4)	t(5)	t(8)
3	t(2)	t(3)	t(4)	t(5)	t(6)	t(9)
4	t(3)	t(4)	t(5)	t(6)	t(7)	t(10)

Outlier Analysis

- Age = {3, 56, 23, 39, 156, 52, 41, 22, 9, 28, 139, 31, 55, 20, -67, 37, 11, 55, 45, 37}



Mean = 39.9
SD = 45.65
Threshold = Mean $\pm$ 2 * SD

Threshold = [-54.1, 131.2]