Problem Solving

with data mining/analytics

4 Steps

1. Understand Business and determine data mining Framework
2. Determine Variables
    1. Not the only one formulation but many possible ones
    2. Each has Pros and Cons
    3. So, compromise
3. Find a model, clusters, or rules
4. If result is not satisfactory, go to 1.

Step 1: Framework

- Do you have something to predict?
  - Prediction/Classification
  - Exception: "normal vs abnormal"
    Conceptually supervised but practically unsupervised Anomaly Detection
- Do you want to partition object set?
  - Clustering
- Do you want to find item co-occurrence?
  - Association Rule Mining

Step 2: Variables (prediction)

- Choose y among candidates
  - Business impact, availability of x's
- Choose x's among candidates
  - Relevance to y,
  - cost and time to obtain:
    - Readily available, with minor efforts, with major investment
- (May repeat after modeling)

Step 2: Variables (anomaly detection)

- Definition of Anomaly detection
  - Detect "anomaly"
  - Problem: Normal data abundant, yet anomalies or abnormal data too few or non existent
- Conventional classification framework does not work
- Learn "normal" area with normal data with Unsupervised learning
- Identify variables to watch when you determine Anomaly

Step 2: Variables (clustering)

- Partition of object set based on what?
- Customers
  - SNS, Telco, Creditcard, Bank, Retail,
- Machines
  - IoT signals

Step 2: Distance (clustering)

- How similar is similar?
- Quantitative measure
  - Euclidean
  - Manhattan
  - Tanimoto
  - Etc

Step 2: Items (ARM)

- Granularity of Items (objects)
- Example: retail items
  - F&B
  - vs Beverage
  - vs Soda
  - vs Cola
  - vs Coca Cola
  - vs Diet Coca Cola
  - vs Diet Coca Cola 2 liter

Step 3

- Find predictor f such that y = f(x)
- Find cluster label for each data x
- Find co-occurrence rules from transaction data