DATAMINING B(3)

11-1 Clustering/Multi-Instance

Clustering

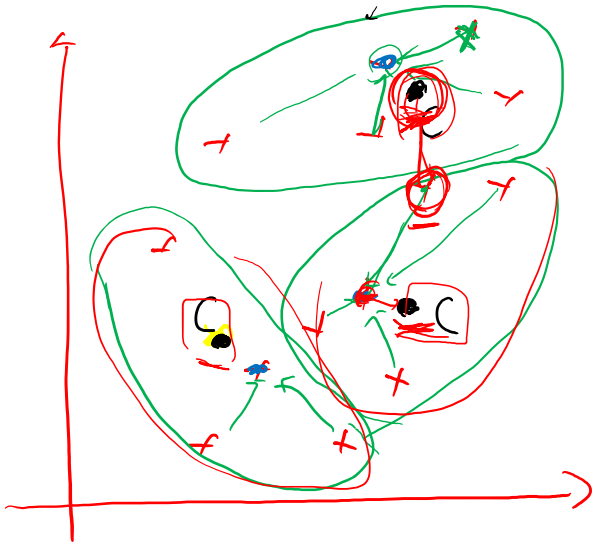
- We want to classify the data set which class is unknown.
  - Correct classes are not given. In other words, no training data.

教師なし unsupervised

- Makes cluster which is the group of instances similar each other.
  - "Similarity" again.

類似度 ⟷ Instance Based Learning

How to make clusters. K-means.



- K-means method.
  - The most famous method for hard clustering.

- Basic Procedure
  1. Define k, the number of clusters.
  2. Choose k instances randomly as centroids of clusters C_j.
  3. Find the nearest centroid for each instance.
  4. Let instance i into the cluster C_j which centroid is the nearest from i.
  5. Calculate centroid C_j by the mean of the member of the cluster.
  6. If at least one centroid is moved, return to 3., else finish this procedure.

Some points of K-means

- Works well. Simple and effective.
- It is not optimal result.
- Depends on the initial value.
  - Different initial value leads different result.
  - Sometimes, clustering is failed.
  - K-means++, which introduce some improvement on the selection of initial value.
    - Choose instance as far away from other instance already chose as centroids as possible.

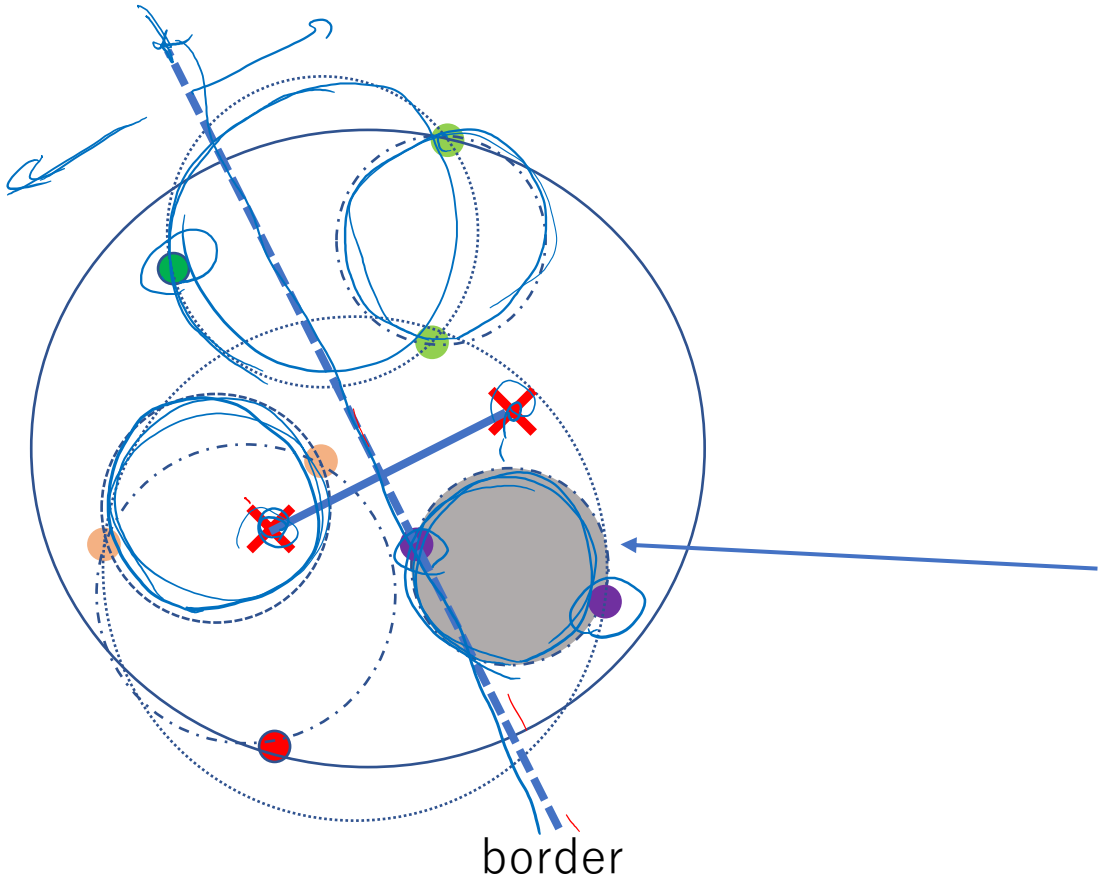
Other point of K-means.

- Find the nearest centroid effectively.

- It is same concept as Instance-based learning.
  - kD-Tree or ball-Tree is available for clustering.

How to use ball-tree for centroid search.

- Make ball-tree from data.



1. Put centroids ✕ on the space.
2. Divide space by perpendicular bisectors.

3. Balls completely contained in one area, the instance of the ball belong to the cluster corresponds to the centroid in the area.
4. If ball exists on the border line, the instance in the ball should be checked precisely.

In the left example, the grayed ball is on the border, purple instances should be checked.

Summary of Clustering

- K-means and K-means++
  - Algorithm of k-means.
  - Features and problems.
  - Improvement of k-means++

- First search
  - Same concept as Instance-based learning.
    - kD-Tree and ball-tree.

Multi-Instance Learning

- Chap.2: Multi-Instance as a knowledge representation.
  - Combination of instances makes another information.
  - Ex: sister-of relation

- Here, Learning from Multi-instance is topic of this section.

Aggregating the Input

- Input is multi-instance
  - Use simple way.
  - Convert multi-instance to single-instance: Aggregating
    - Mean, Mode, Minimum, Maximum, and so on.
  - Learn or predict from the single-instance.

  - It may mark almost same accuracy as a complex method.
    - Complex: handle multi-instance as multi-instance.

  - It need to try to choose aggregating way.
    - The cost will be offset by the reducing the number of instances.

Aggregate the Output

- Input is multi-instance.
  - Separate each instance in it.
  - Learn or predict from each instance and get result.
  - Results are aggregated.
    - Mean, Mode, and so on.
- Number of instances may be different.
  - Consider the weight of each instance in it.