K-Means Clustering and Gaussian Mixture Model

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Weekly Objectives

- Understand the clustering task and the K-means algorithm
  - Know what the unsupervised learning is
  - Understand the K-means iterative process
  - Know the limitation of the K-means algorithm
- Understand the Gaussian mixture model
  - Know the multinomial distribution and the multivariate Gaussian distribution
  - Know why mixture models are useful
  - Understand how the parameter updates are derived from the Gaussian mixture model
- Understand the EM algorithm
  - Know the fundamentals of the EM algorithm
  - Know how to derive the EM updates of a model

K-MEANS ALGORITHM

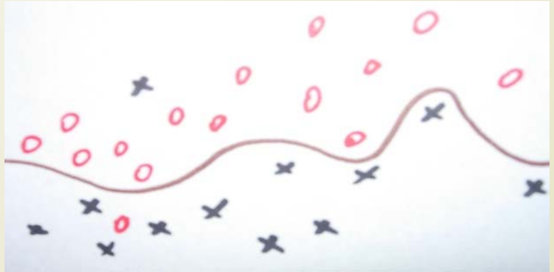
Types of Machine Learning

Machine Learning

.....

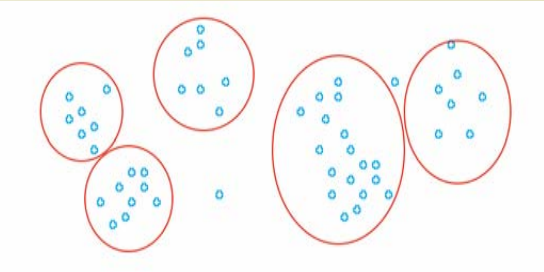
Supervised Learning

You know the true answers of some of instances



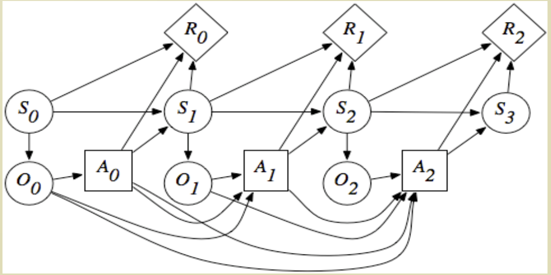
Unsupervised Learning

You do not know the true answers of instances



Reinforcement Learning

You do know the objective, but you do not know how to achieve



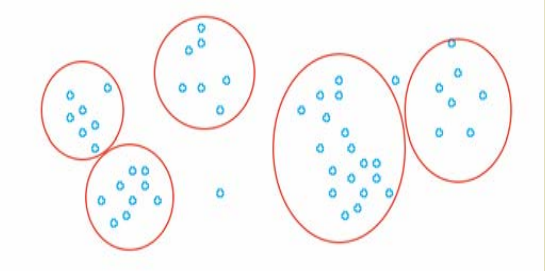
- *You* can
  - Machine learning
  - Dataset provider
  - Machine learning users
  - etc

- Various classifications by different professors
  - Purpose, data types, etc
- Other learning classifications also exist

Unsupervised Learning

Unsupervised Learning

You do not know the true answers of instances

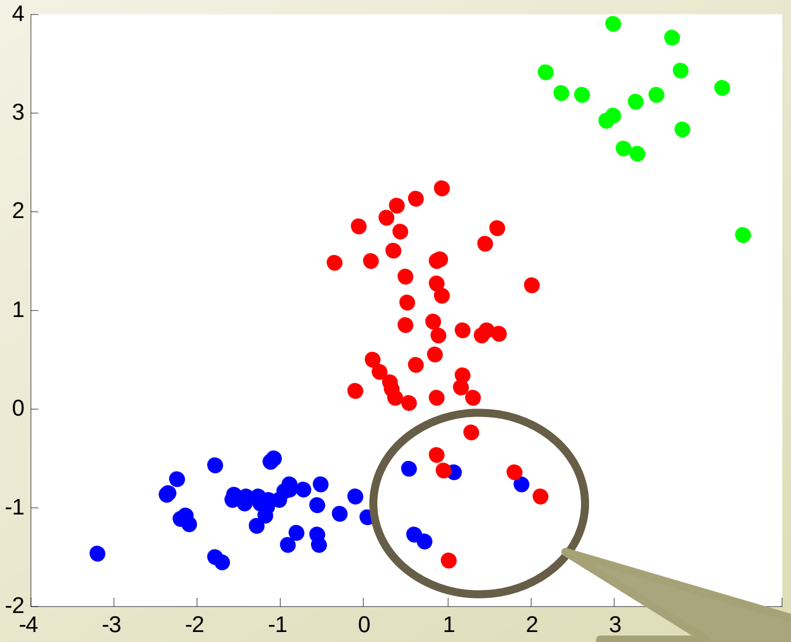


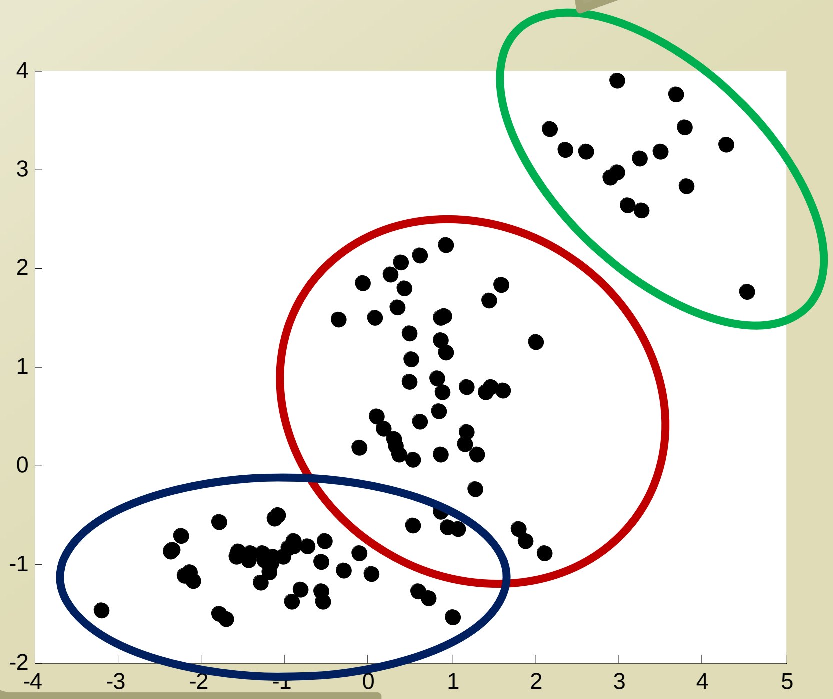
- **You don't know the true value, and you cannot provide examples of the true value.**
- Cases, such as
  - Discovering clusters
  - Discovering latent factors
  - Discovering graph structures
- Clustering or filtering or completing of
  - Finding **the representative topic words from text data**
  - Finding **the latent image from facial data**
  - Completing the incomplete **matrix of product-review scores**
  - Filtering the **noise from the trajectory data**
- Methodologies
  - Clustering: estimating sets and affiliations of instances to the sets
  - Filtering: estimating underlying and fundamental signals from the mixture of signals and noises

Clustering Problem

- How to cluster the unlabeled data points?
  - No concrete knowledge of their classes
  - Latent (hidden) variable of classes
  - Optimal assignment to the latent classes

How to assign data points to classes?
→ Clustering
(here classes == clusters)



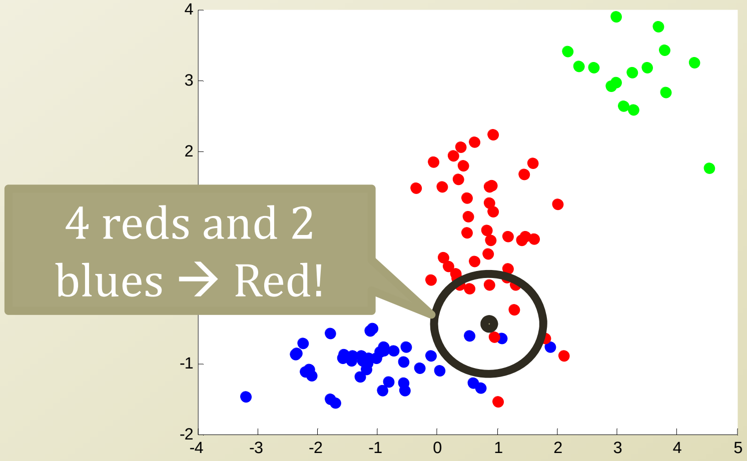


Uncertain area of clustering

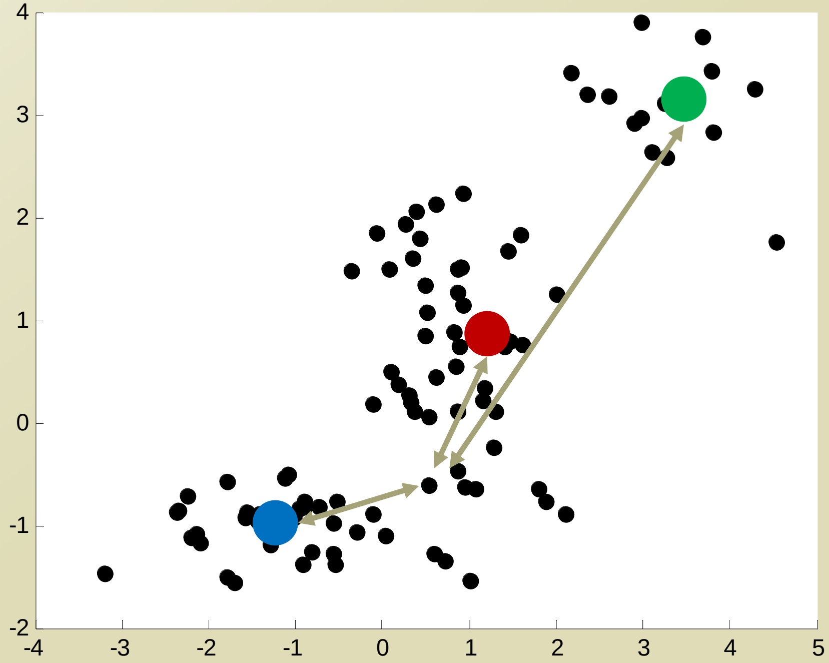
K-Means Algorithm

- K-Means algorithm
  - Setup K number of centroids (or prototypes) and cluster data points by the distance from the points to the nearest centroid
- Formally,
  - $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$
  - Minimize J by optimizing
    - r_{nk}: the assignment of data points to clusters
    - μ_k: the location of centroids
  - Iterative optimization
    - Why?
    - Two variables are interacting

K-Means! = K-Neareat Neighbor



4 reds and 2 blues → Red!


4
3
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3
4
5