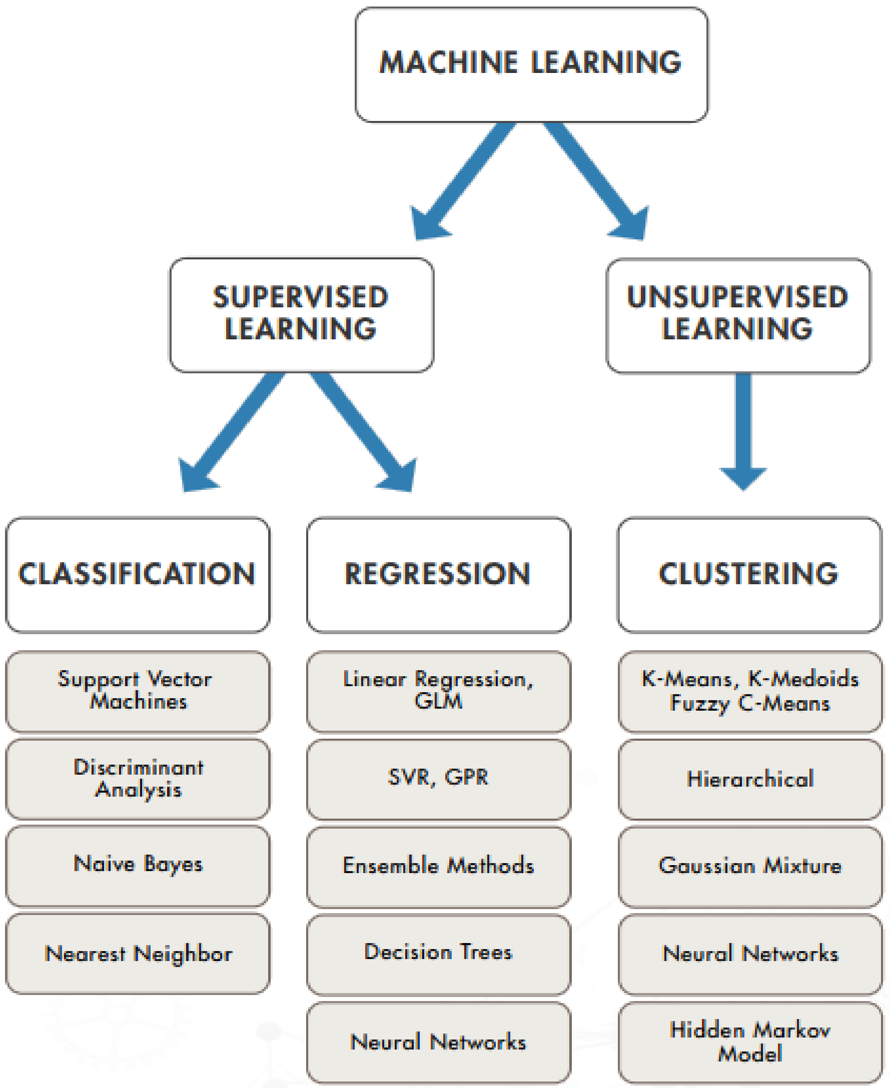
Clustering

Unsupervised Learning



https://www.datasciencecentral.com/profiles/blogs/machine-learning-summarized-in-one-picture?utm_content=bufferbd927&utm_medium=social&utm_source=facebook.com&utm_campaign=buffer)

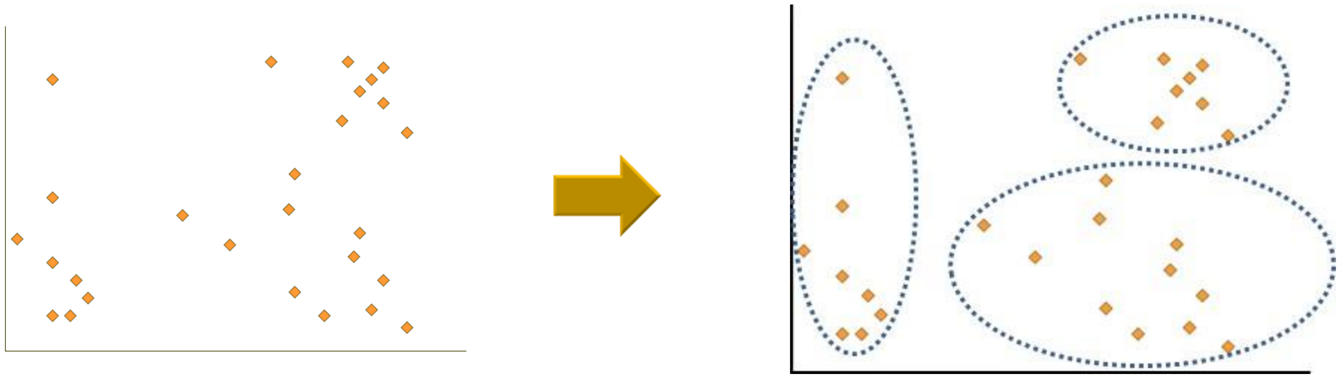
Unsupervised Learning

- Unsupervised learning
  - Finding a specific pattern for data without result information
  - Finding latent structures and hierarchies in the data
  - Finding hidden user groups
  - Structuring documents according to their subject matter
  - Using log information to find usage patterns

- Kinds of unsupervised learning
  - Clustering
  - Density estimation
  - Dimensionality reduction

Clustering

❖ Segmenting data based on similarity



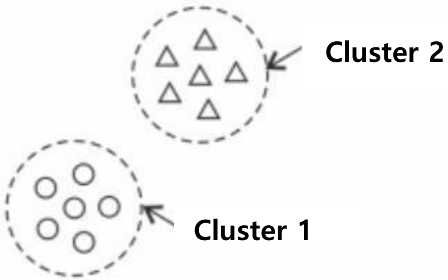
영상 분할(segmentation)

image : Pedro Felzenswalb

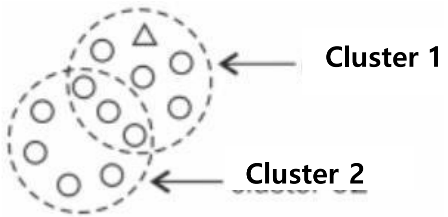
Clustering

❖ Kinds of clustering

- Hard clustering
  - Data belongs to only one cluster
    - e.g., k-means algorithm



- Fuzzy clustering
  - Data partially belongs to multiple clusters
  - The sum of the degree of belonging becomes 1
    - e.g., Fuzzy k-means algorithm



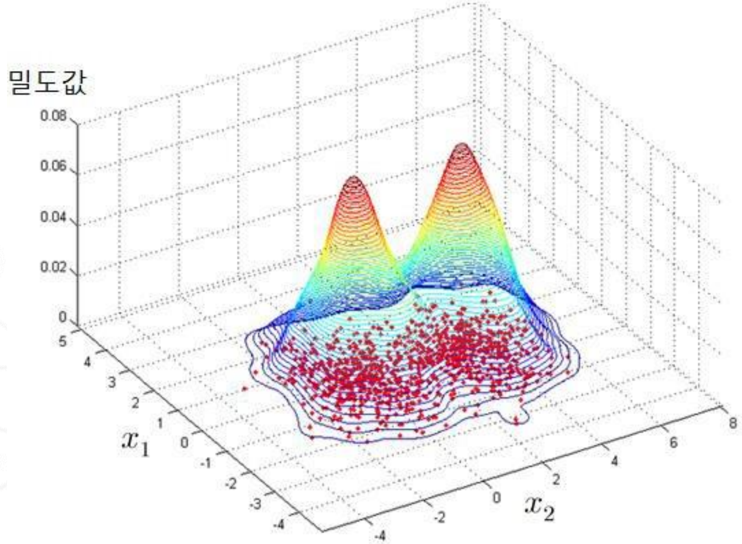
- Purpose
  - Estimate of underlying structure in data sample
  - Insight into the overall structure of the data
  - Hypothesis setting, outlier detection
  - Data compression: representing data in the same cluster with the same value
  - Data preprocessing operations

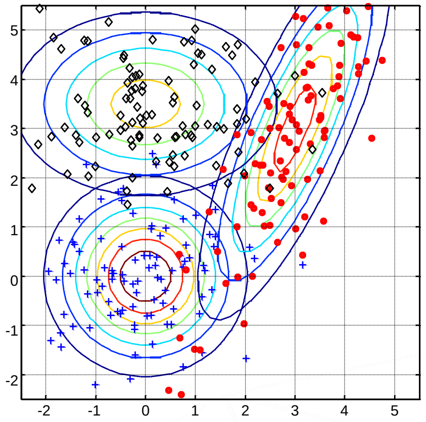
- Performance measurement
  - Variance within a cluster and distance between clusters

Density Estimation

❖ Density estimation

- Finding a probability distribution that is presumed to have generated class-specific data

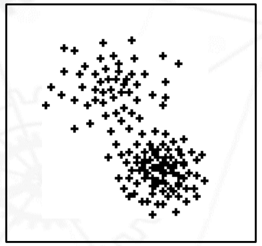


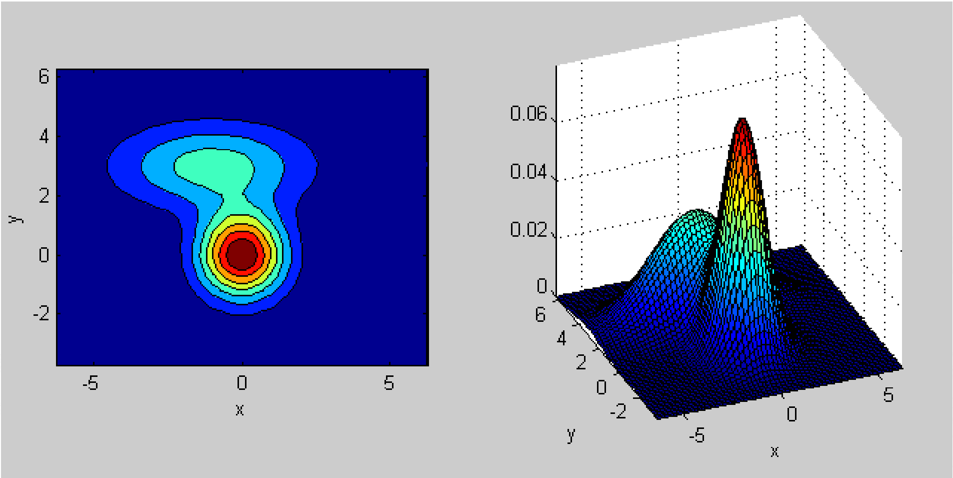


- Purpose
  - Calculate the probability of generating a given data for each class
  - Classify as the most likely class

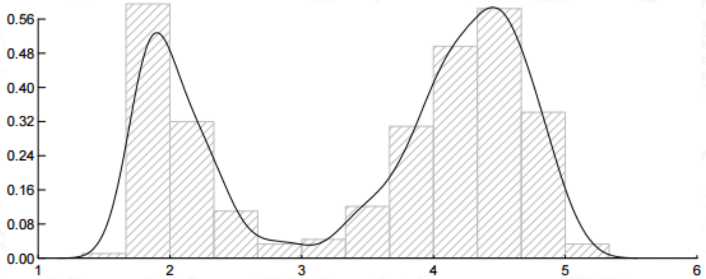
Density Estimation

- Parametric density estimation
  - Assume that the distribution has the form of a particular mathematical function
  - Determine the parameters of the function to best reflect the given data
    - Gaussian function or several Gaussian
      - Mixture of Gaussian





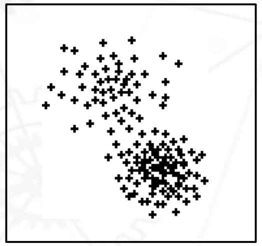
http://i.stack.imgur.com/pE0Xu.gif

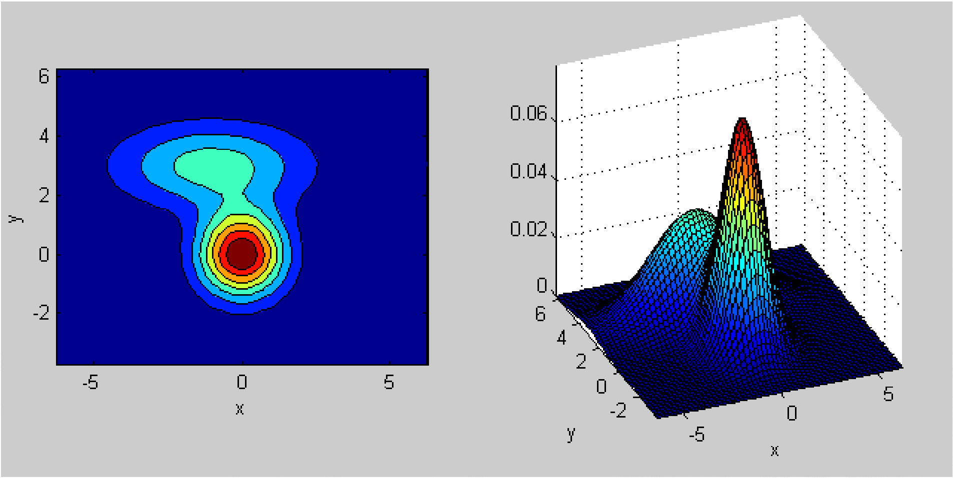


Density Estimation

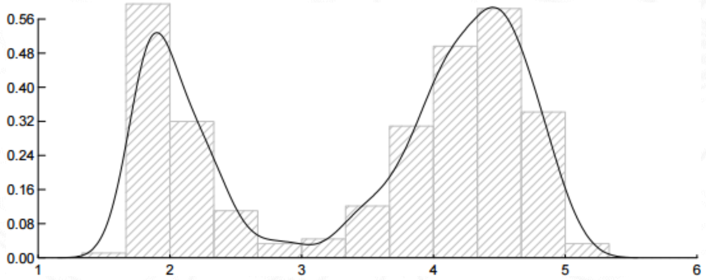
❖ Nonparametric density estimation

- Expression of the shape of the density function using the given data without assuming a specific function for the distribution
    - Histogram





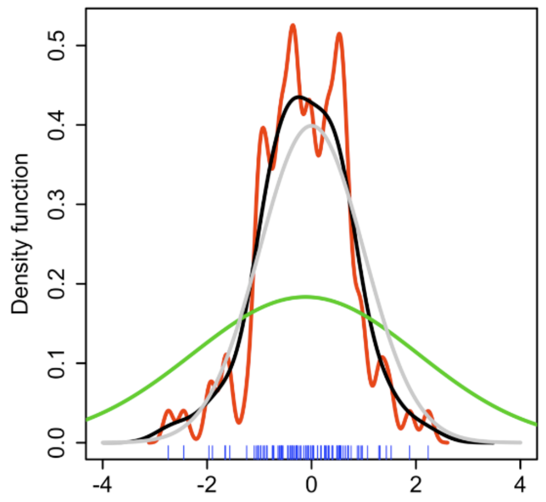
http://i.stack.imgur.com/pE0Xu.gif



Density Estimation

❖ Nonparametric density estimation

- Limitations of Histogram
  - 계급구간(bin)의 경계가 불연속적 (discrete)
  - 계급구간(bin)의 시작 위치에 따라 히스토그램(분포)가 달라짐
  - 계급구간(bin)의 크기에 따라 히스토그램(분포)가 달라짐
  - 그 외에도 고차원 데이터에 대해서는 메모리 문제 등으로 사용하기 어렵다는 단점이 있음



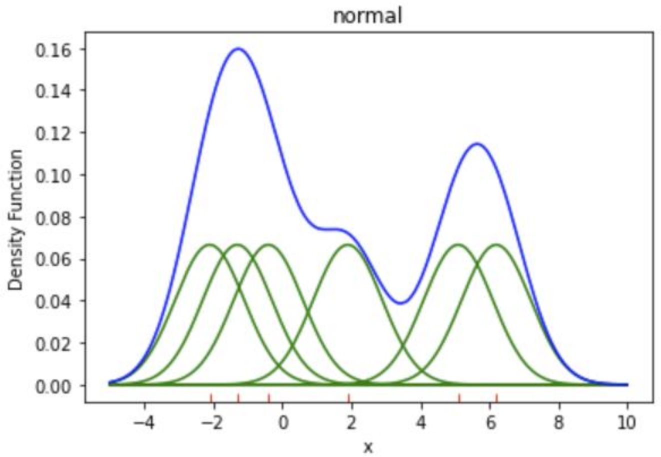
- Kernel Function

(1) $\int_{-\infty}^{\infty} K(u)du = 1$

(2) $K(-u) = K(u) for\ all\ values\ of\ u.$

(3) $Non\text{-}negative$

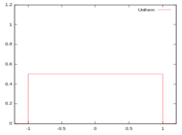
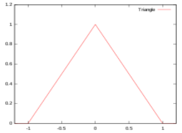
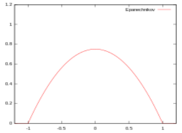
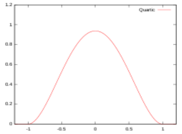
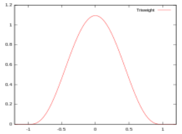
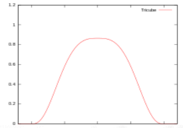
$$\widehat{f_h}(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathrm{x} - x_i) = \frac{1}{nh}K\left(\frac{\mathrm{x} - x_i}{h}\right)$$

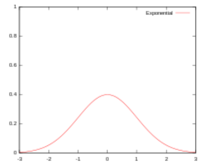
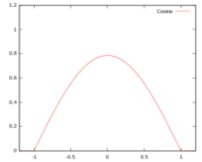
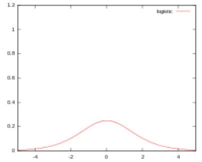
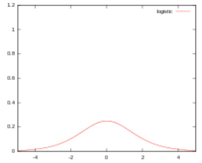
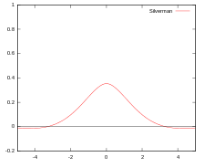


Density Estimation

❖ Nonparametric density estimation

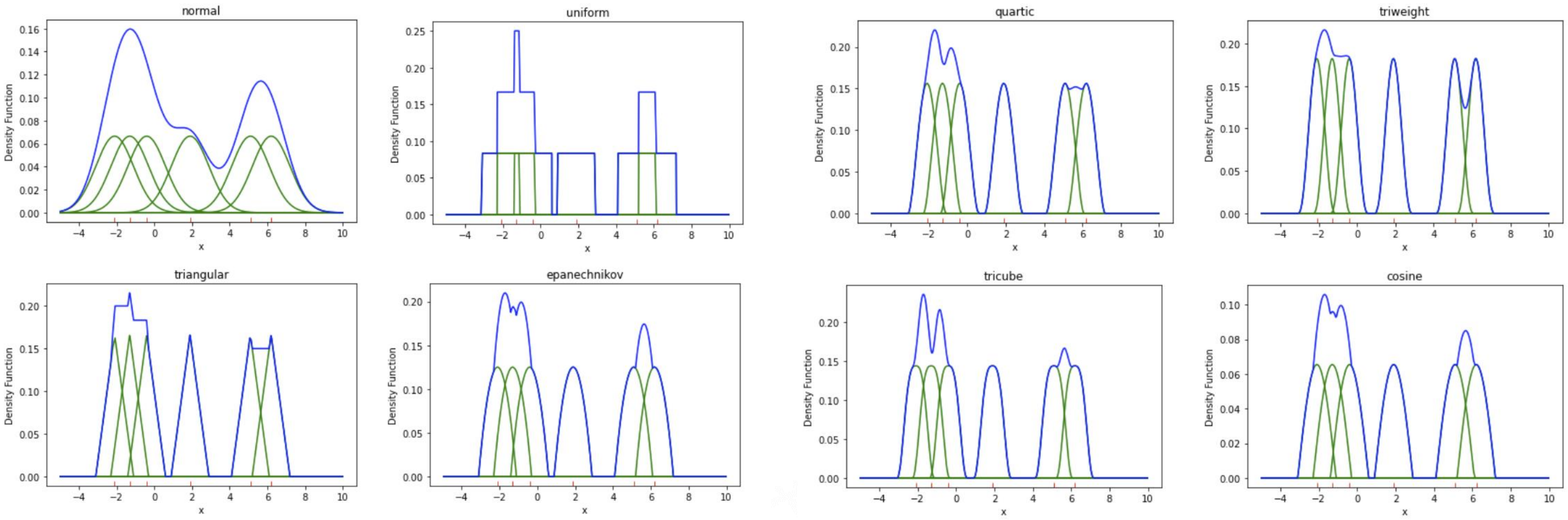
- Kernel Function

Uniform ("rectangular window")	$K(u) = \frac{1}{2}$ Support: $\lvert u \rvert \le 1$	 "Boxcar function"
Triangular	$K(u) = (1 - \lvert u \rvert)$ Support: $\lvert u \rvert \le 1$	
Epanechnikov (parabolic)	$K(u) = \frac{3}{4}(1 - u^2)$ Support: $\lvert u \rvert \le 1$	
Quartic (biweight)	$K(u) = \frac{15}{16}(1 - u^2)^2$ Support: $\lvert u \rvert \le 1$	
Triweight	$K(u) = \frac{35}{32}(1 - u^2)^3$ Support: $\lvert u \rvert \le 1$	
Tricube	$K(u) = \frac{70}{81}(1 - \lvert u \rvert^3)^3$ Support: $\lvert u \rvert \le 1$	

Gaussian	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$		1	$\frac{1}{2\sqrt{\pi}}$
Cosine	$K(u) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)$ Support: $\lvert u \rvert \le 1$		$1 - \frac{8}{\pi^2}$	$\frac{\pi^2}{16}$
Logistic	$K(u) = \frac{1}{e^u + 2 + e^{-u}}$		$\frac{\pi^2}{3}$	$\frac{1}{6}$
Sigmoid function	$K(u) = \frac{2}{\pi} \frac{1}{e^u + e^{-u}}$		$\frac{\pi^2}{4}$	$\frac{2}{\pi^2}$
Silverman kernel[6]	$K(u) = \frac{1}{2} e^{-\frac{\lvert u \rvert}{\sqrt{2}}} \cdot \sin\left(\frac{\lvert u \rvert}{\sqrt{2}} + \frac{\pi}{4}\right)$		0	$\frac{3\sqrt{2}}{16}$

❖ Nonparametric density estimation

- Kernel Function



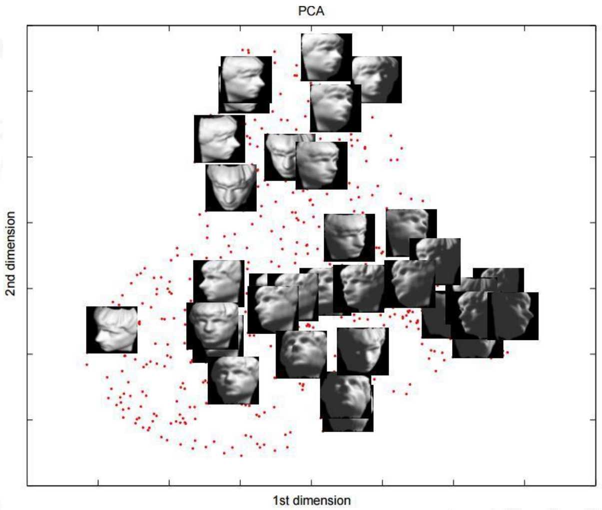
Dimension Reduction

❖ Dimension reduction

- Converting high-dimensional data to low dimensions with minimal loss of information

❖ Purpose

- 2D or 3D conversion and visualization enables intuitive data analysis
- Mitigate the issue of curses of dimensionality



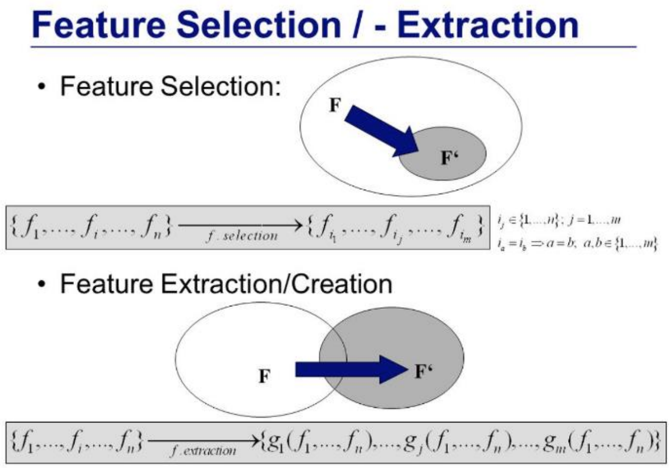
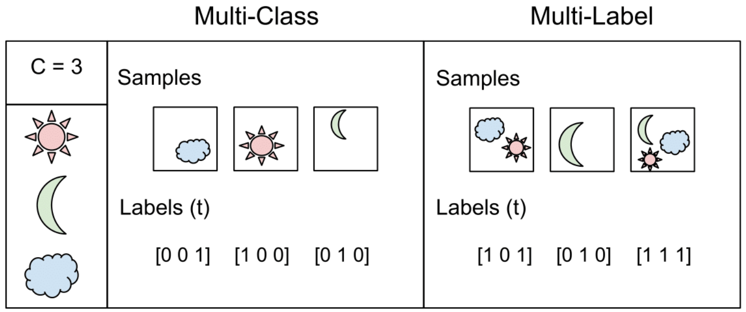


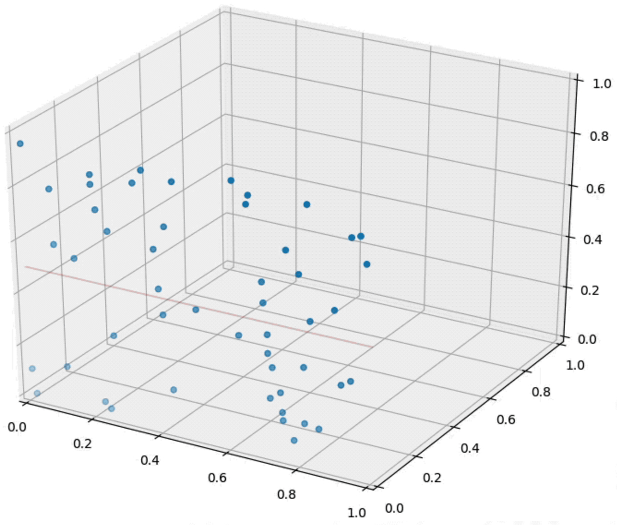
Image: Ghodsi 2006

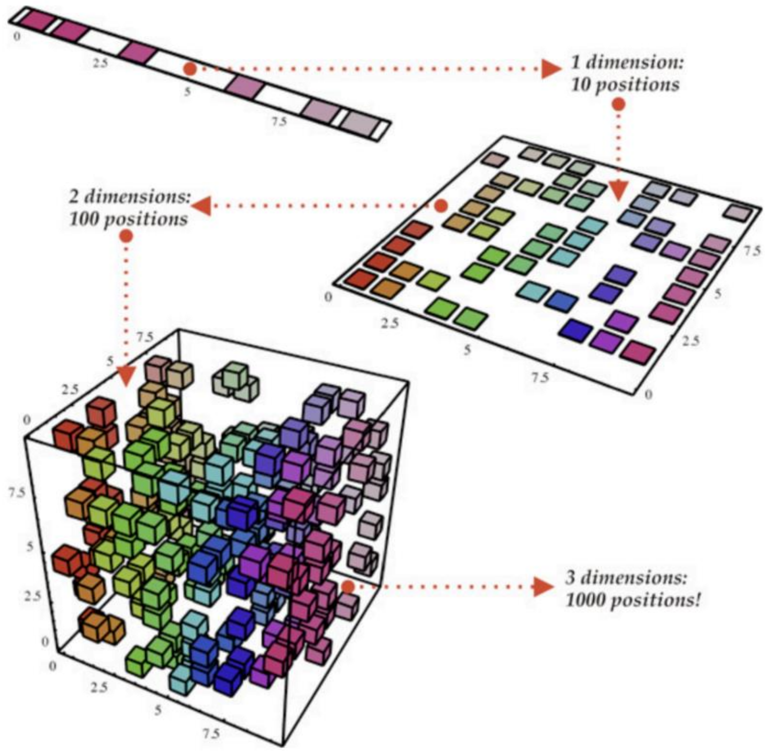
Dimension Reduction

❖ Curse of dimension

- The tendency of distance distribution to be constant as dimension increases
- As dimensions increase, the number of subspaces increases exponentially





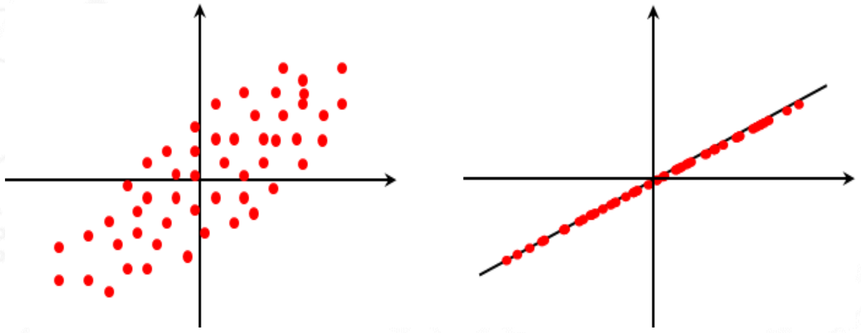


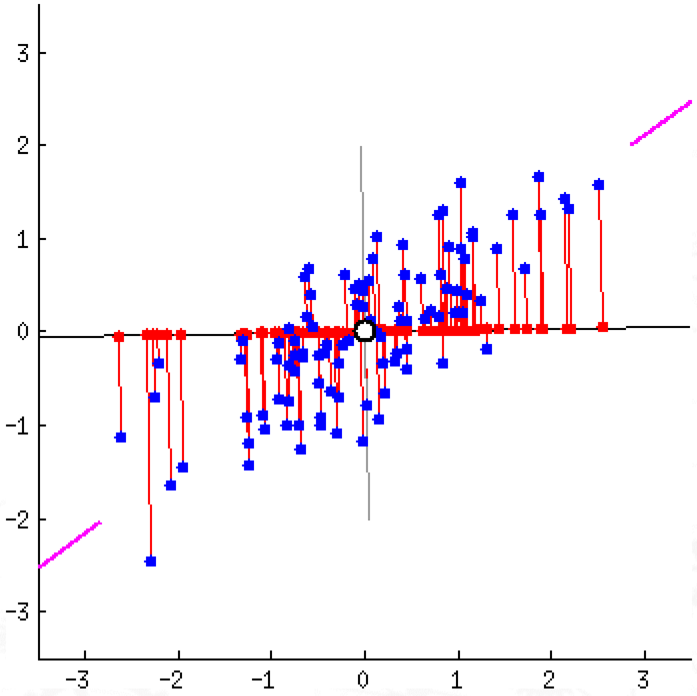
KDD 2014 tutorial, Yoshua Bengio

Dimension Reduction

❖ Principle Component Analysis, PCA

- Project data based on several large distributed axes and convert it into low dimensions
- Choose a few eigenvectors with a large eigenvalue for the covariance matrix of the data as the event axis





https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Outlier Detection

- Outlier
  - Data that is significantly different from other data and questionable whether it was generated by other mechanisms
  - Subject of interest

- Noise
  - Observation errors, random errors in the system
  - Destinations you are not interested in to remove

- Related to novelty detection

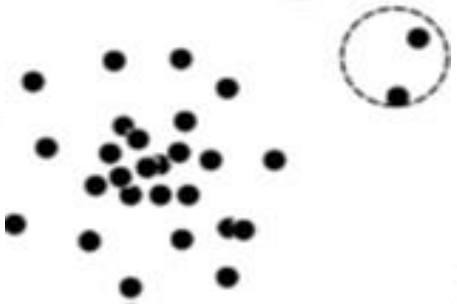
Outlier Detection

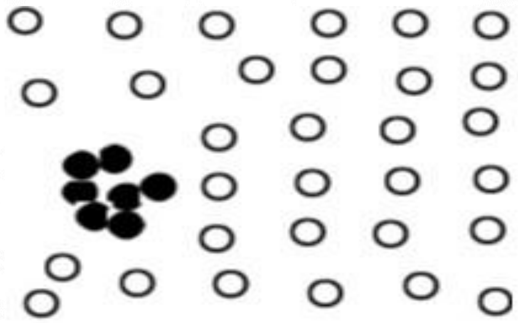
❖ Example of outlier detection

Outlier Detection

❖ Example of outlier

- Point outlier
  - Data that differs significantly compared to other data
- Contextual outlier
  - Uncontextual Data
- Collective outlier
  - A collection of data that looks abnormal when you collect multiple data





Outlier Detection

❖ Example

- 부정사용감지 시스템(fraud detection system, FDS)
  - 이상한 거래 승인 요청 시에 카드 소유자에게 자동으로 경고 메시지 전송
- 침입탐지 시스템(intrusion detection system, IDS)
  - 네트워크 트래픽을 관찰하여 이상 접근 식별
- 자율주행 시스템
- 시스템의 고장 진단
- 임상에서 질환 진단 및 모니터링
- 공공보건에서 유행병의 탐지
- 스포츠 통계학(Sabermetrics)에서 특이 사건 감지
- 관측 오류의 감지

Clustering Algorithm

- Clustering
  - Gathering data among similar things
  - The similarity between clusters is large, and the similarity within clusters is small
- Hierarchical clustering
  - To allow clusters to have hierarchical structures as a result of clustering
  - Agglomerative hierarchical clustering
    - Starting with each data forming one cluster, the process of combining nearby clusters is repeated to form hierarchical clusters
  - Divisive hierarchical clustering
    - Start with a cluster containing all the data and separate clusters based on similarity to form an increasingly hierarchical structure

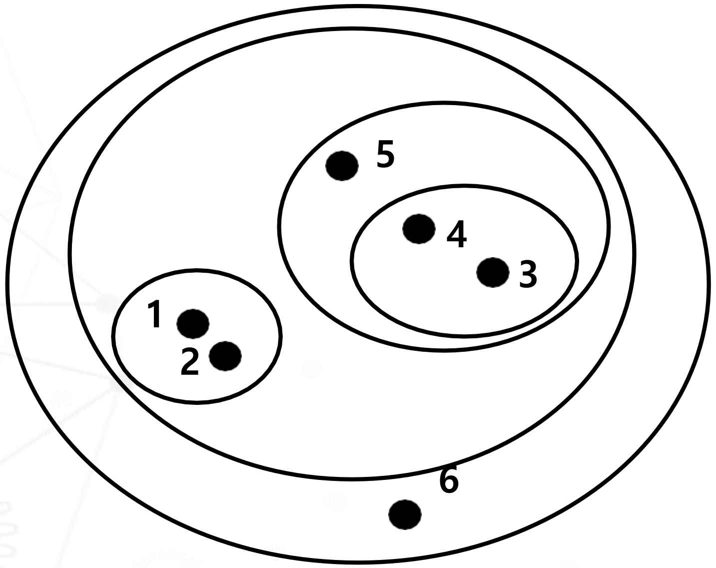
Clustering Algorithm

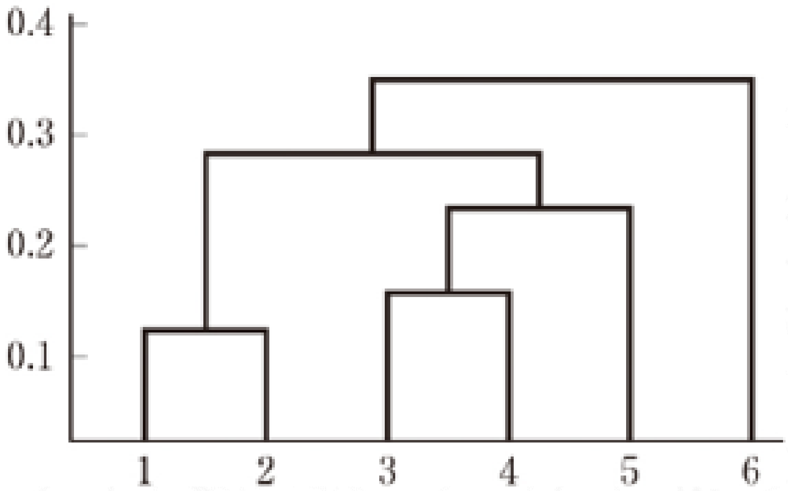
❖ Partitioning clustering

- Grouping the entire data into similar pieces without creating a hierarchical structure
- e.g., k-means algorithm

Hierarchical Clustering

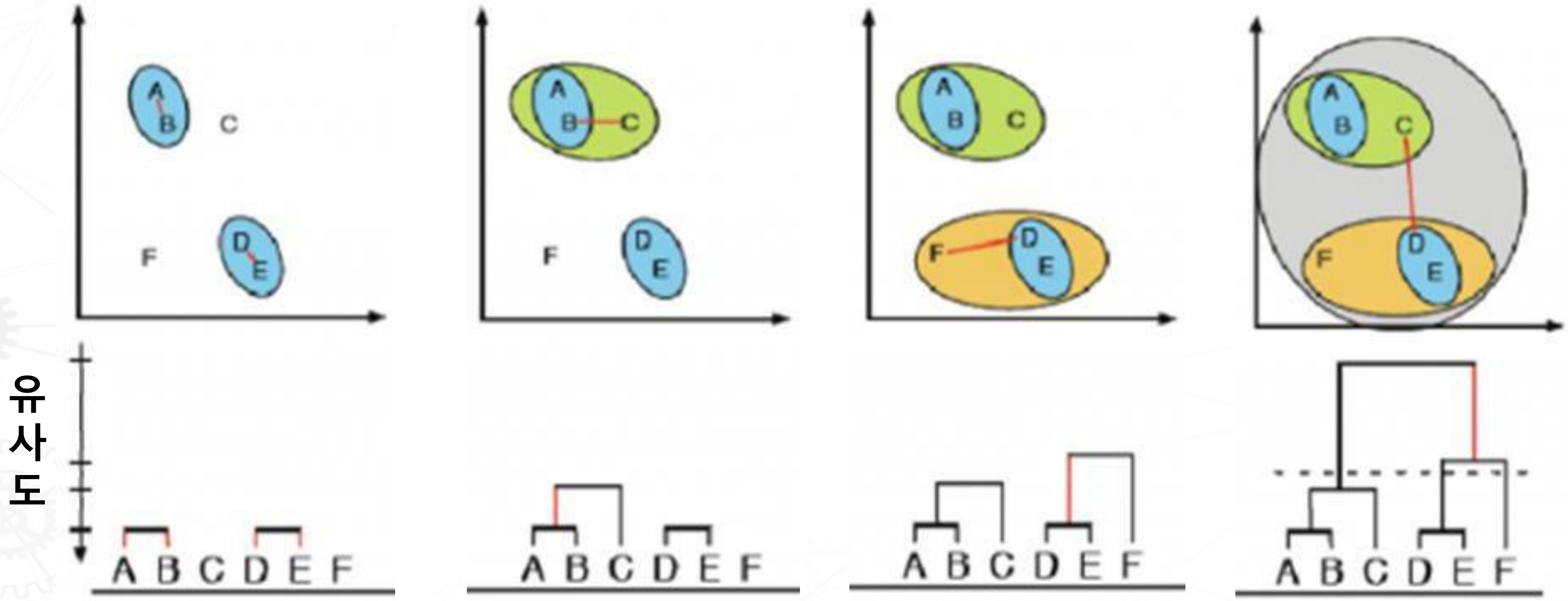
- ❖ Agglomerative hierarchical clustering
  - Dendrogram





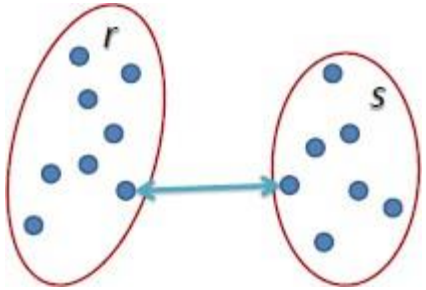
Hierarchical Clustering

❖ Agglomerative hierarchical clustering



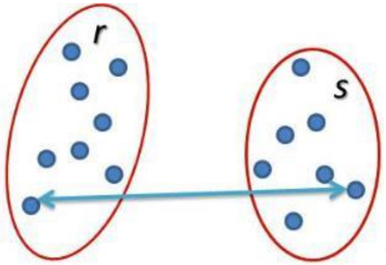
Hierarchical Clustering

- To calculate the inter-cluster distance
  - Simple linkage



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

  - Complete linkage



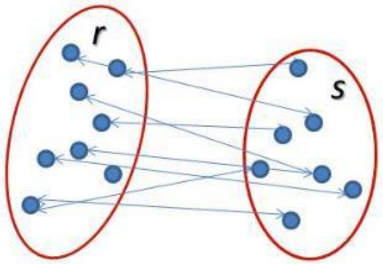
$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019

Hierarchical Clustering

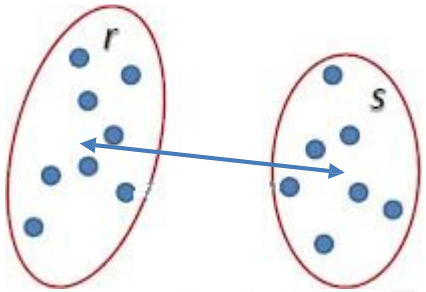
❖ To calculate the inter-cluster distance

- Average linkage



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

- Centroid linkage



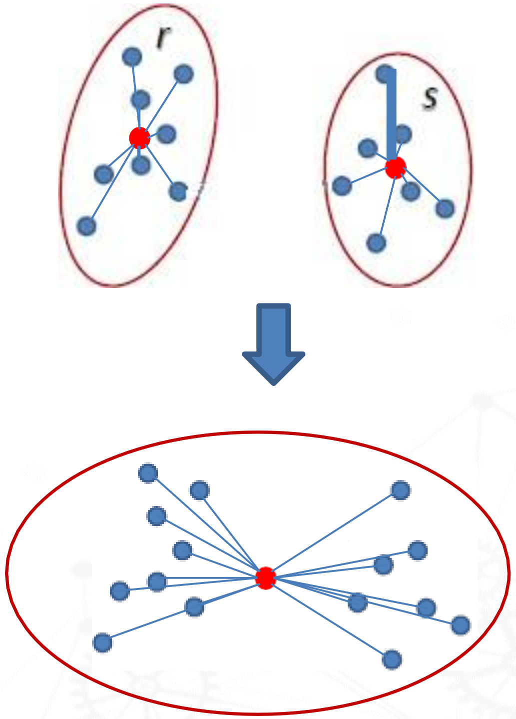
$$L(r,s) = D\left(\frac{1}{n_r}\sum_{i=1}^{n_r} x_{ri}, \frac{1}{n_s}\sum_{j=1}^{n_s} x_{sj}\right)$$

https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019

Hierarchical Clustering

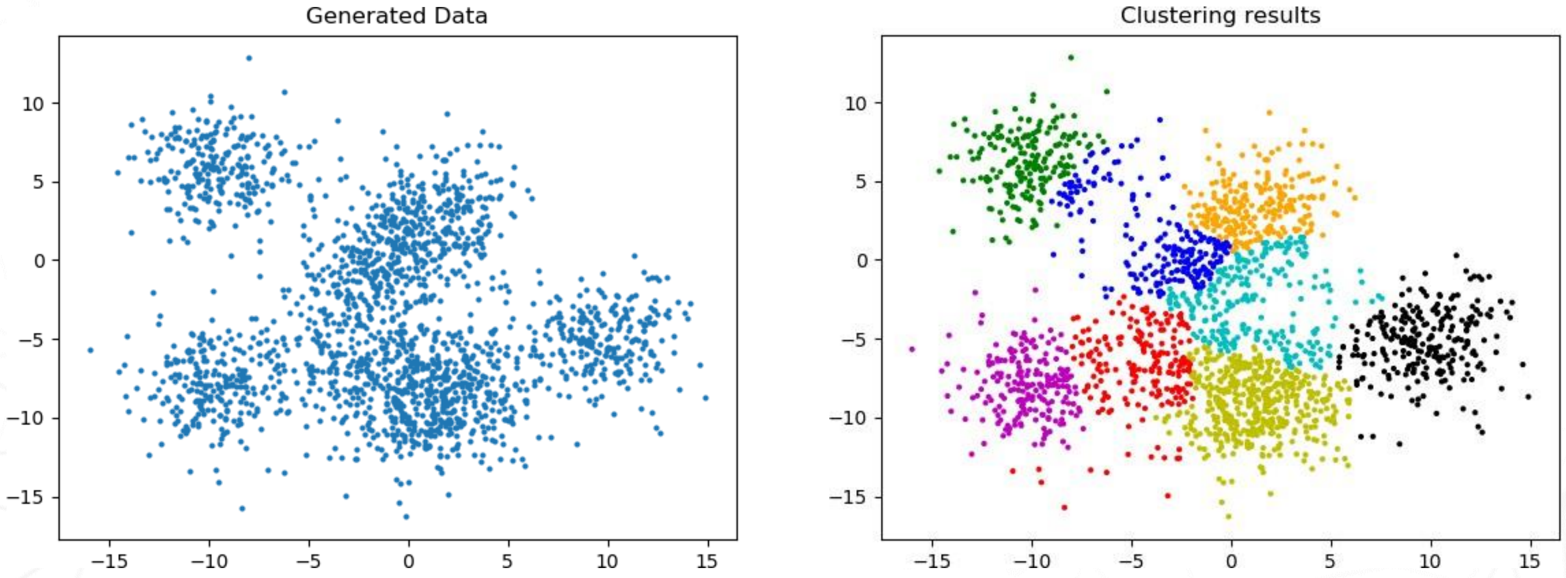
❖ To calculate the inter-cluster distance

- Ward linkage
  - The difference between the sum of squares of the deviations in the cluster before and after cluster coupling



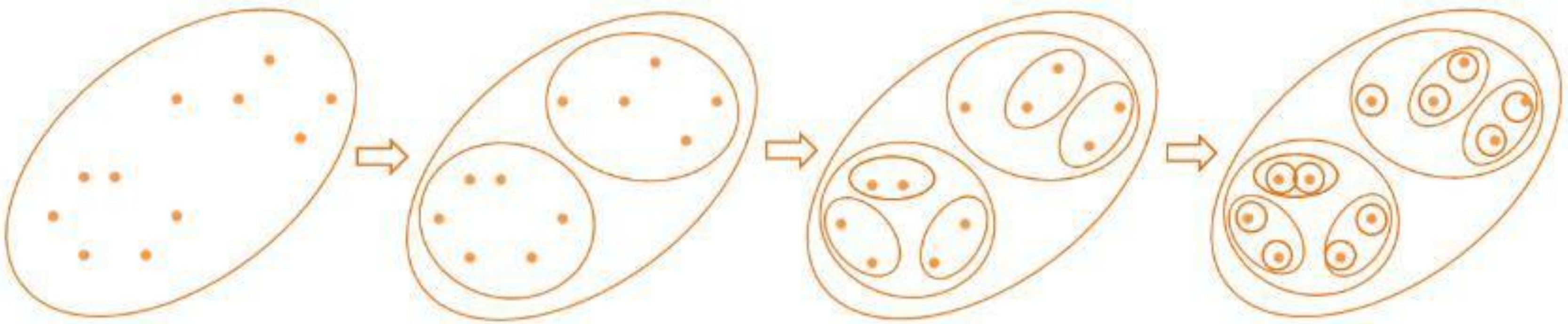
Hierarchical Clustering

❖ Visualization



Hierarchical Clustering

❖ Divisive hierarchical clustering

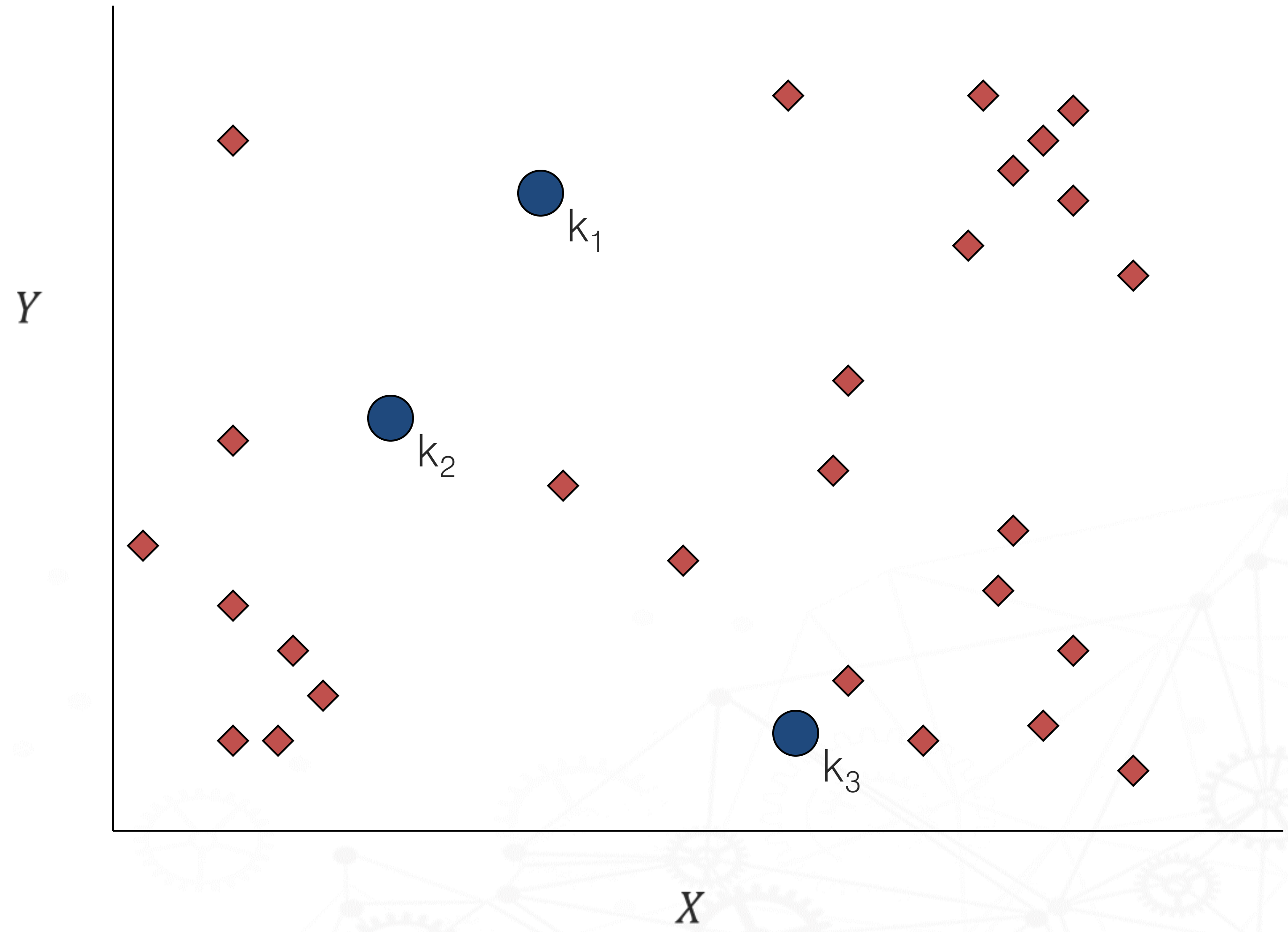


https://towardsdatascience.com/https-towardsdatascience-com-hierarchical-clustering-6f3c98c9d0ca

Partitioning Clustering

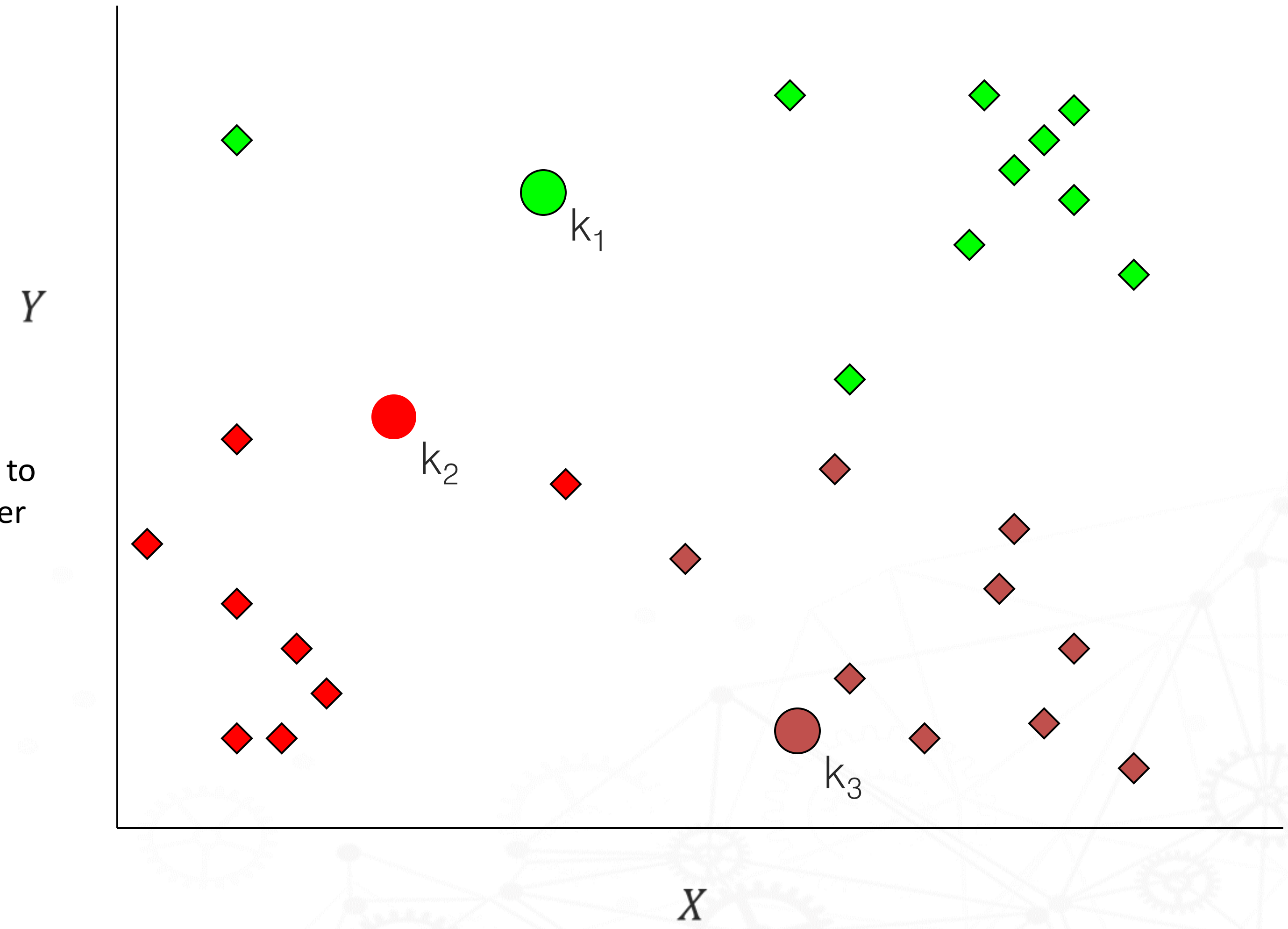
- k-means clustering
  - Partitioning cluster algorithm
  - Process
    1. Central positioning of clusters
    2. Reconfigure clusters based on cluster center
    3. Determining the mean position by cluster
    4. Adjust cluster center to cluster mean position
    5. Repeat steps 2-4 until convergence occurs

K-means Clustering 1



Randomly select three cluster centroid locations

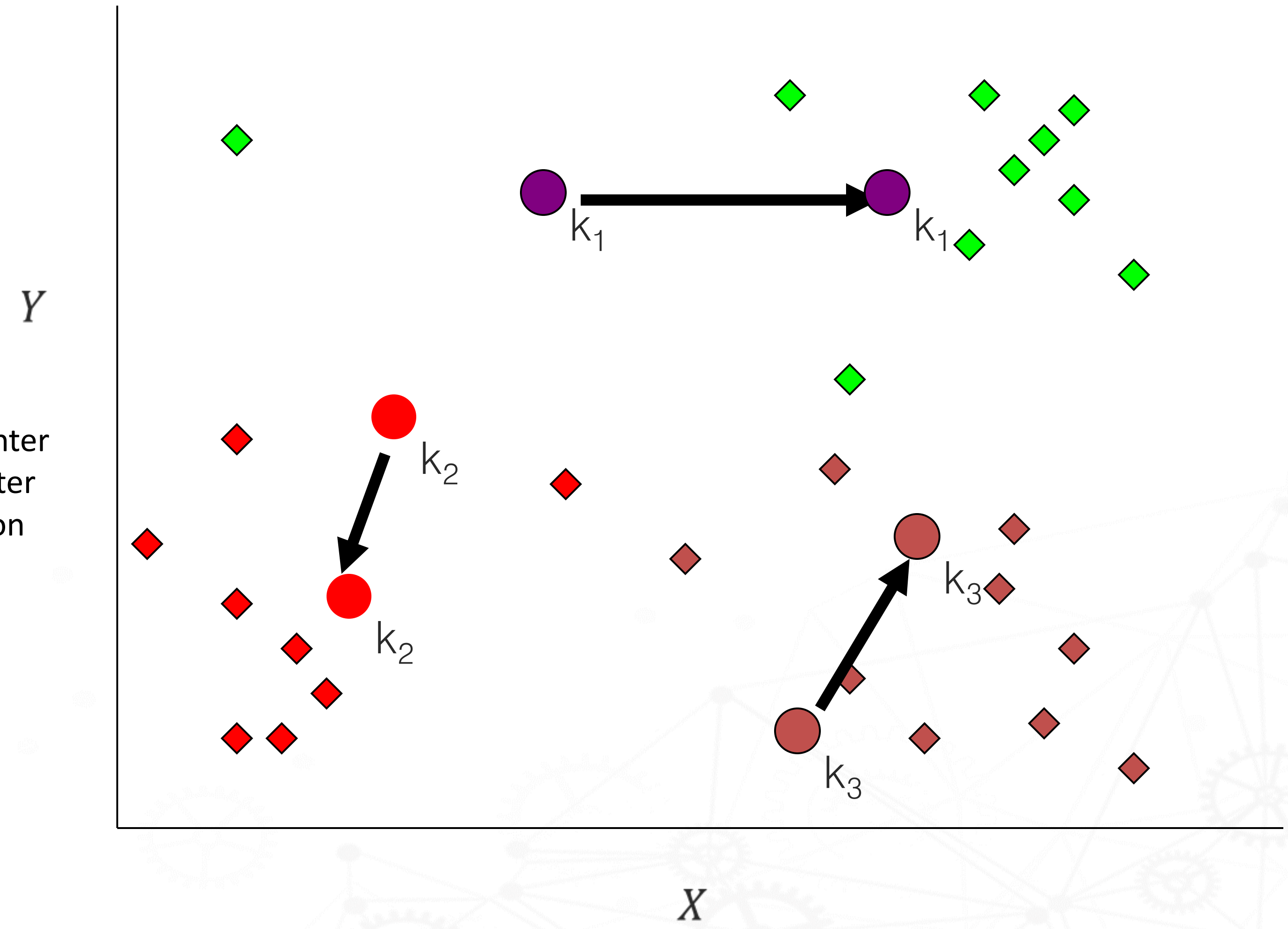
K-means Clustering 2



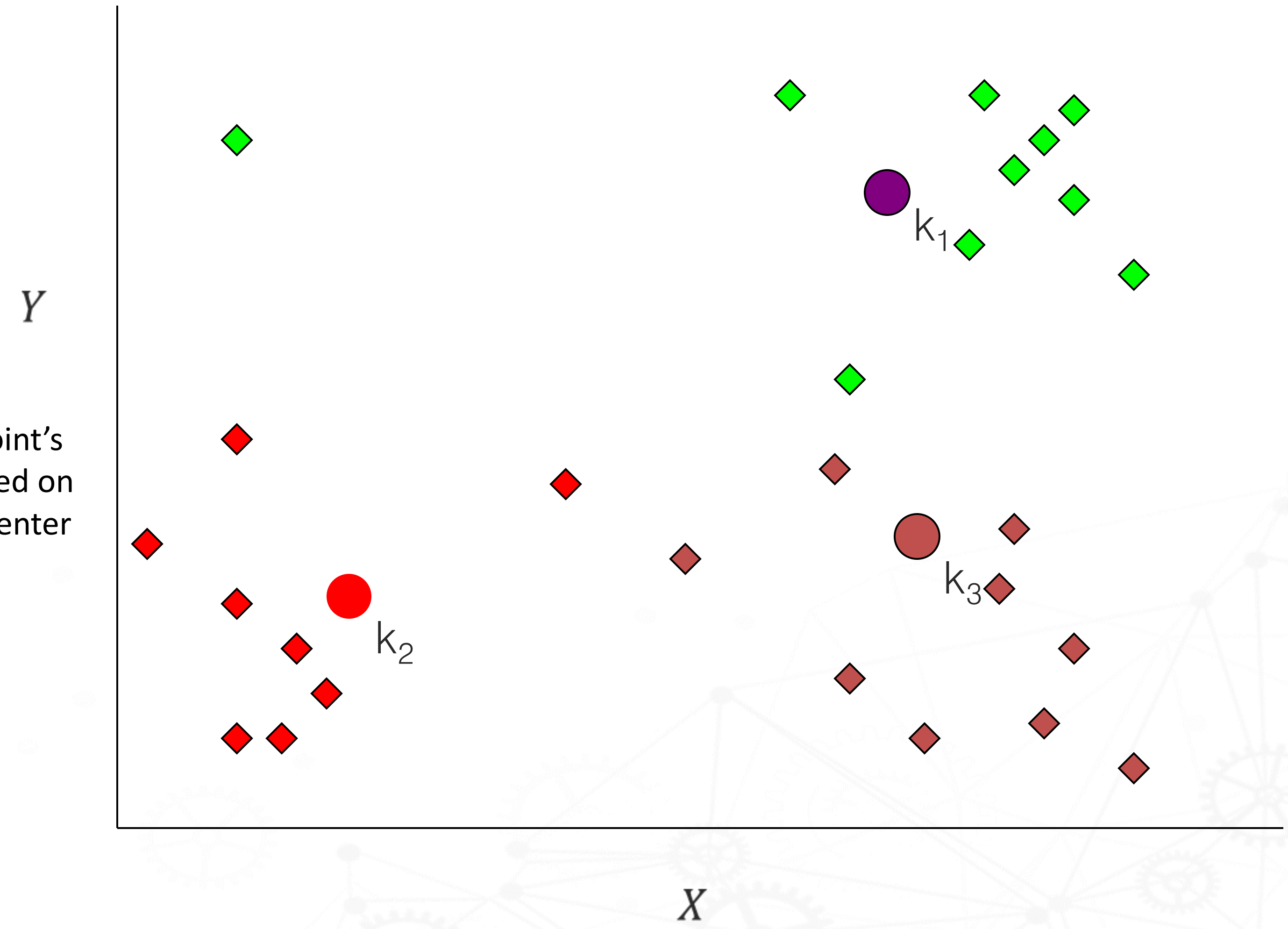
Assign each point to the nearest cluster center position

K-means Clustering 3

Move cluster center position to cluster average position

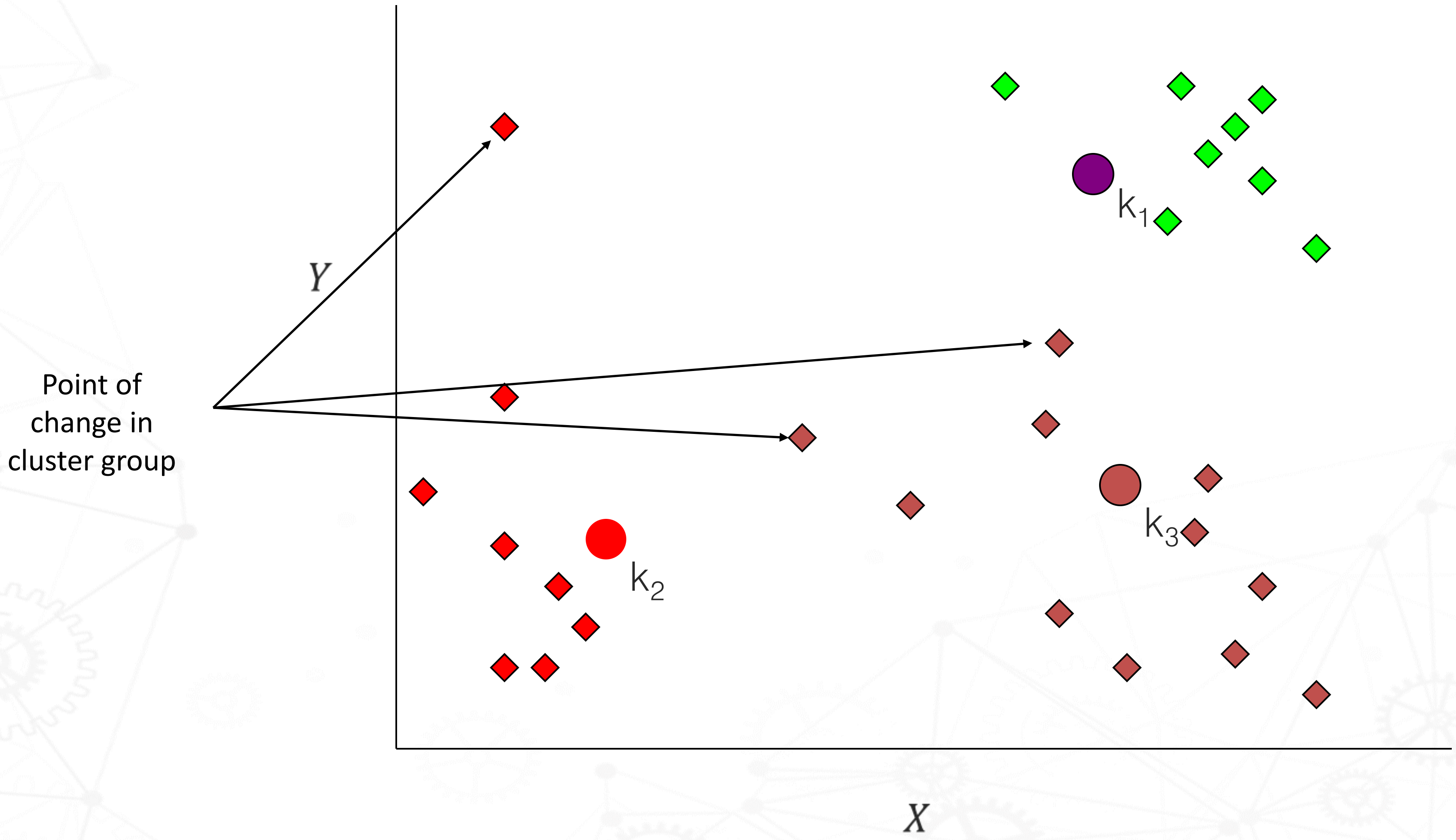


K-means Clustering 4



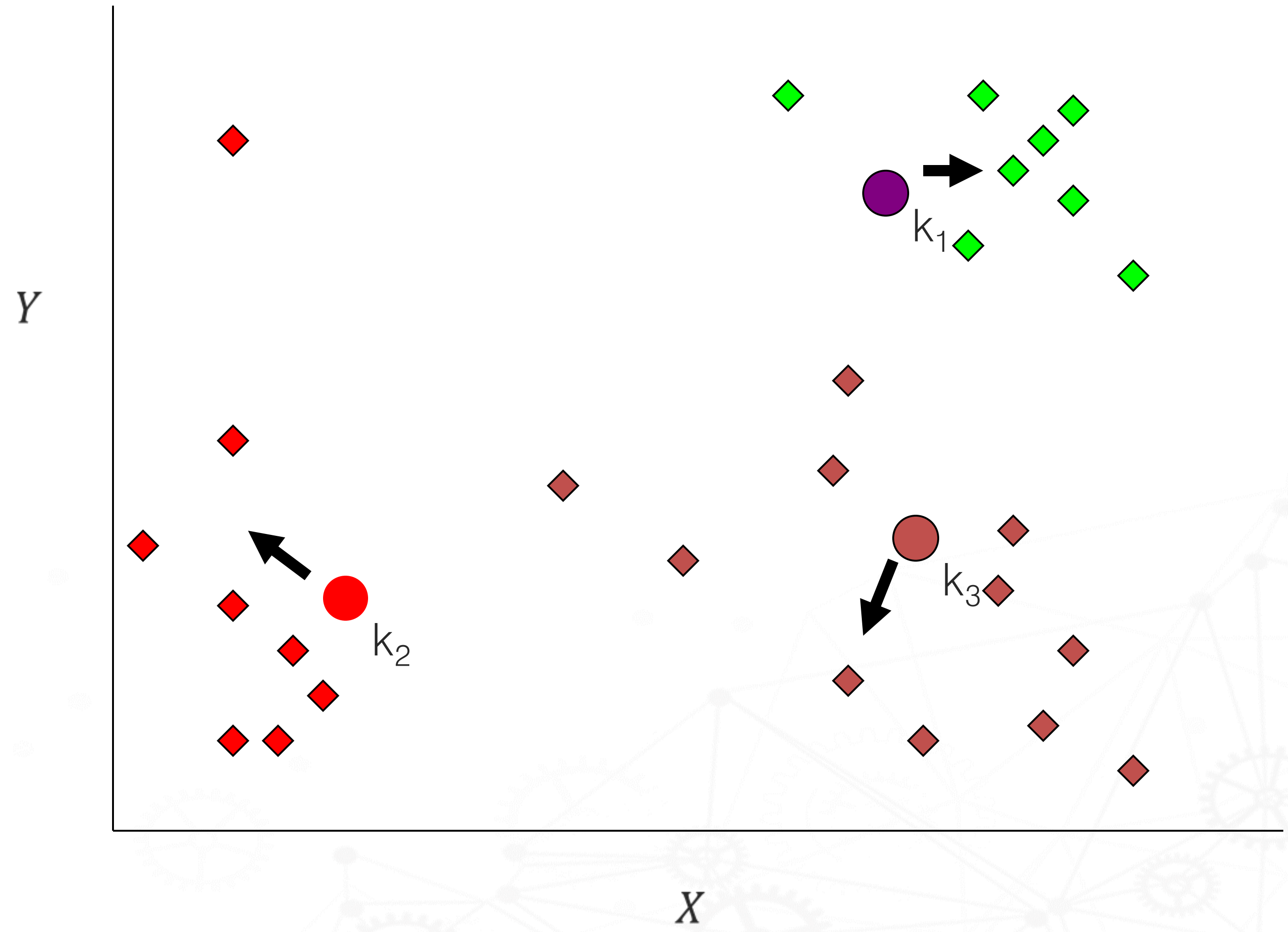
Reassign each point's cluster group based on the new cluster center

K-means Clustering 5



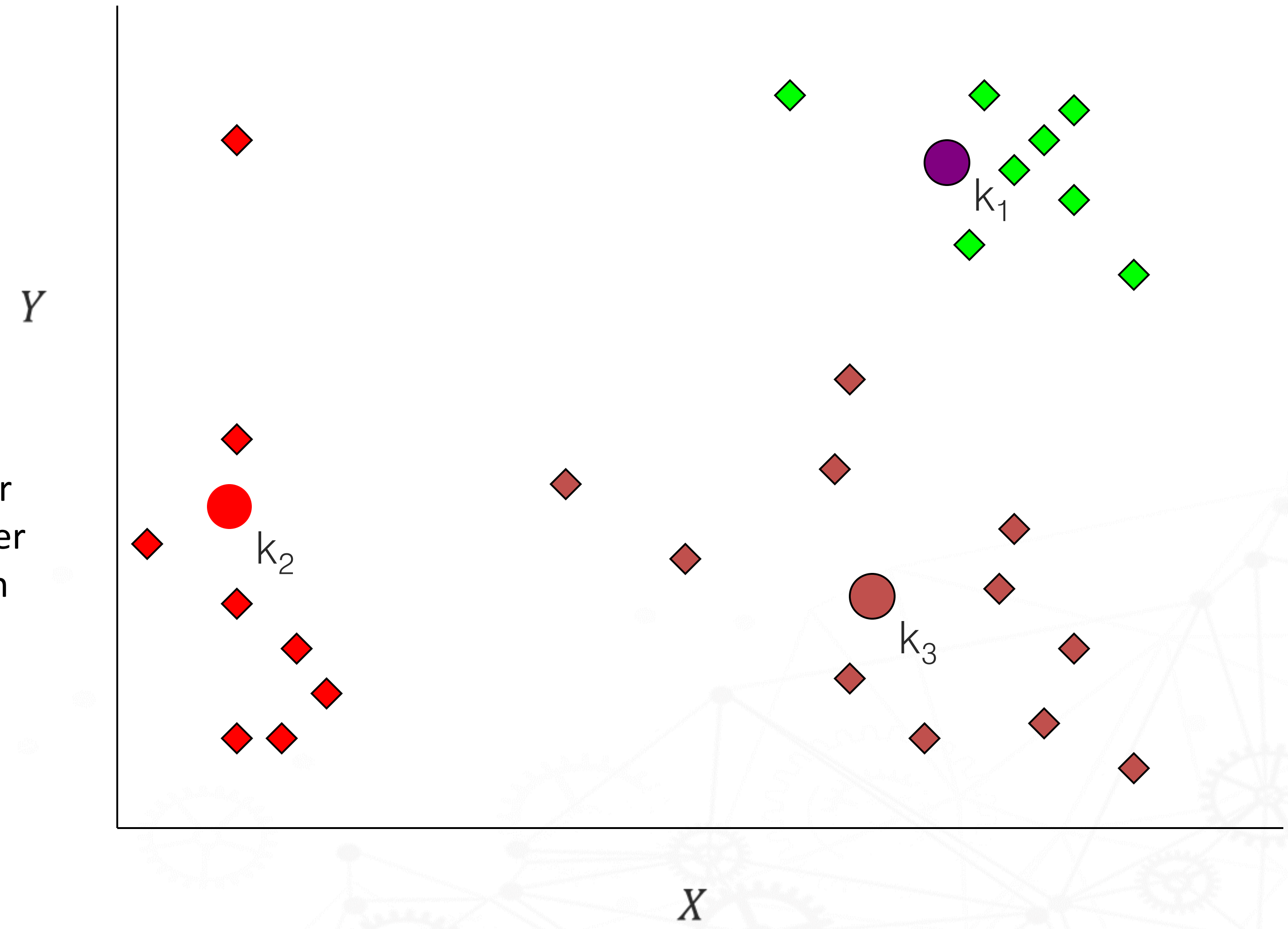
K-means Clustering 6

Recalculating cluster means



K-means Clustering 7

Change cluster center to cluster mean position



K-means Clustering

❖ k-means clustering

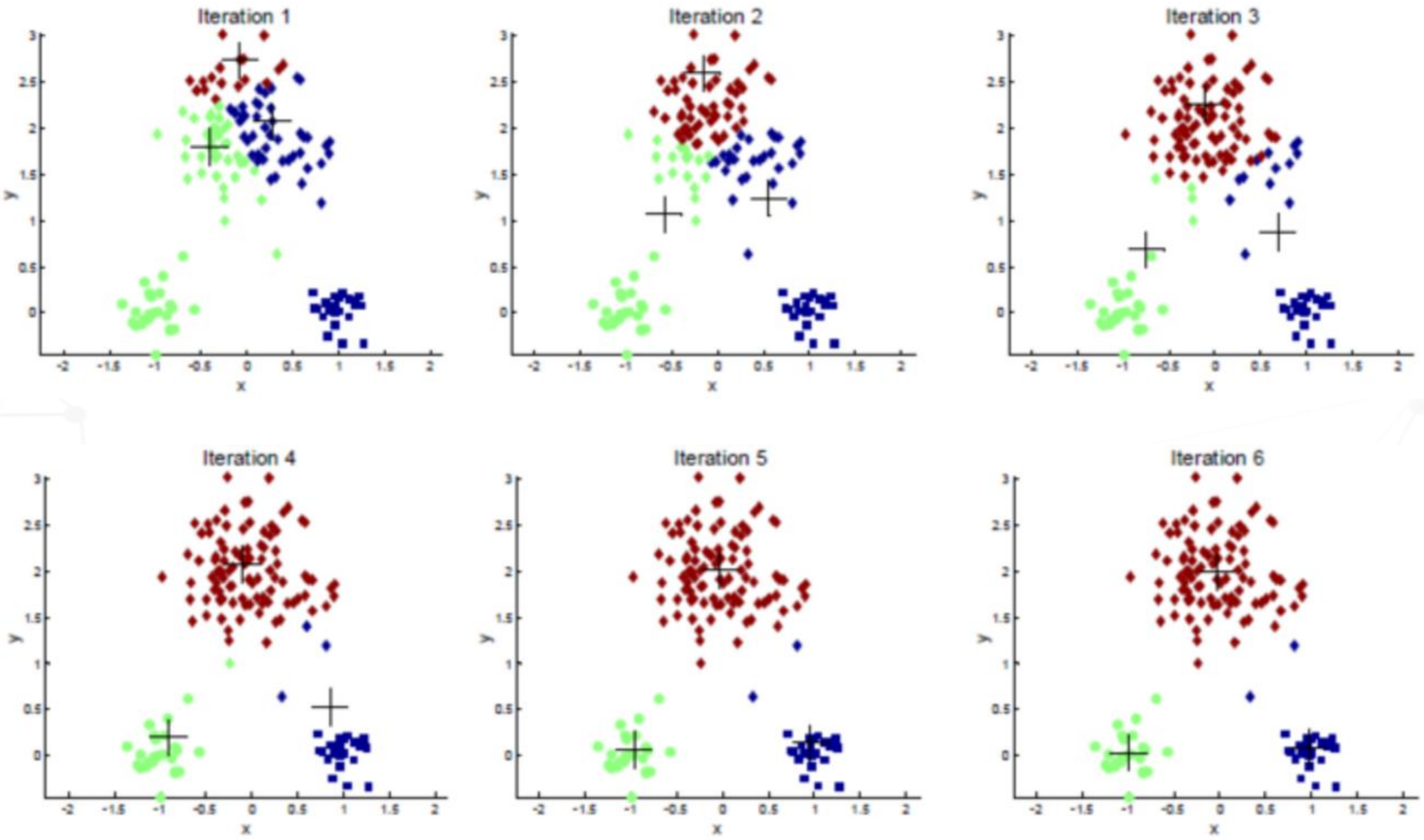
- i, when the center of the i^{th} cluster is μ_i, and the set of points S_i that belong to the cluster, the total variance
- Find a S_i that minimizes the variance V

$$V = \sum_{i=1}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2$$

- Process
  1. Set the initial i to random first
  2. Repeat the following two steps until the cluster remains unchanged
- Feature
  - The number of clusters k is specified in advance
  - Sensitive to initial cluster location

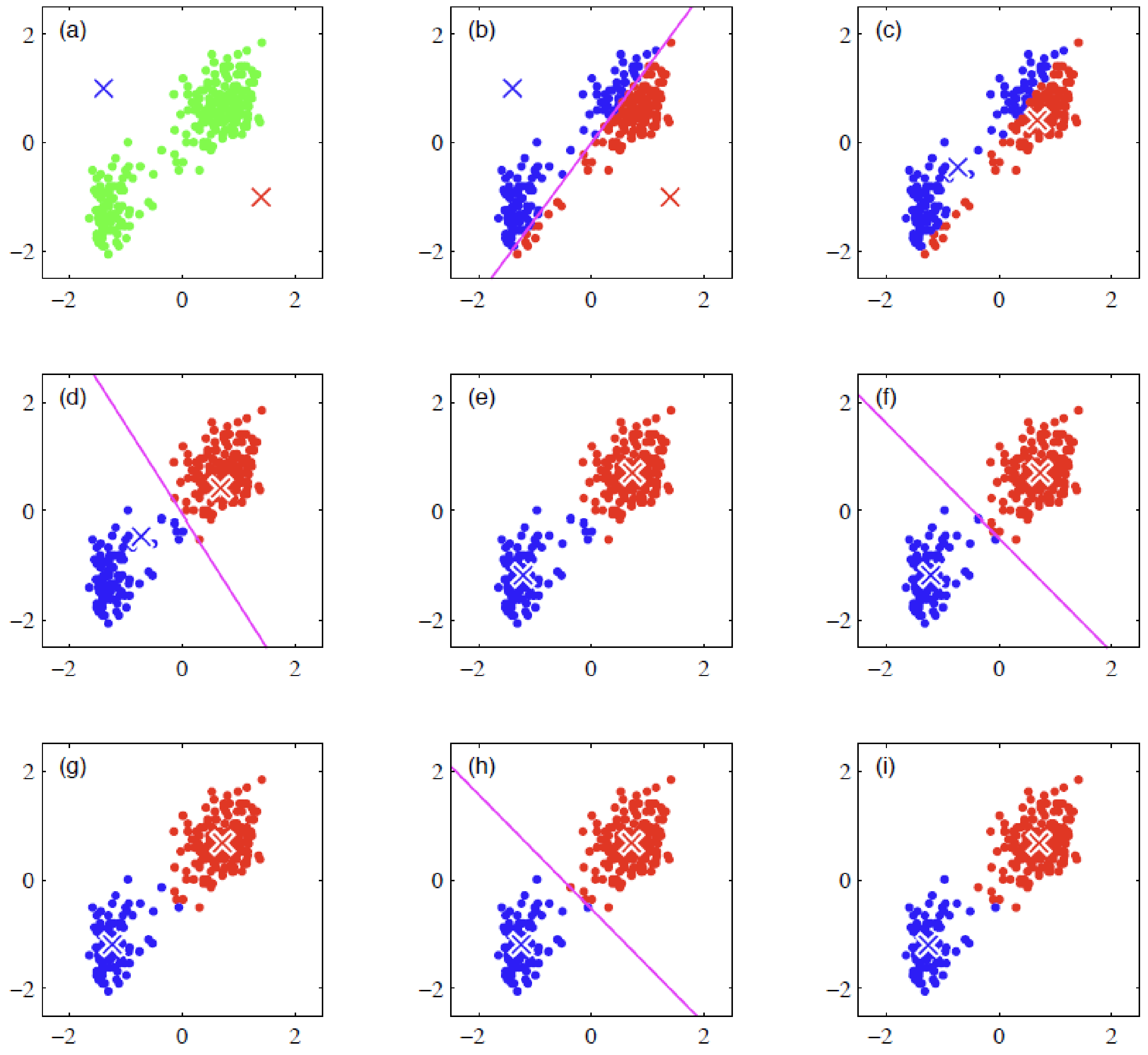
K-means Clustering

❖ Cluster results sensitive to initial center values



https://sites.google.com/site/myecodriving/k-means-ju-lei-fen-xi

K-means Clustering

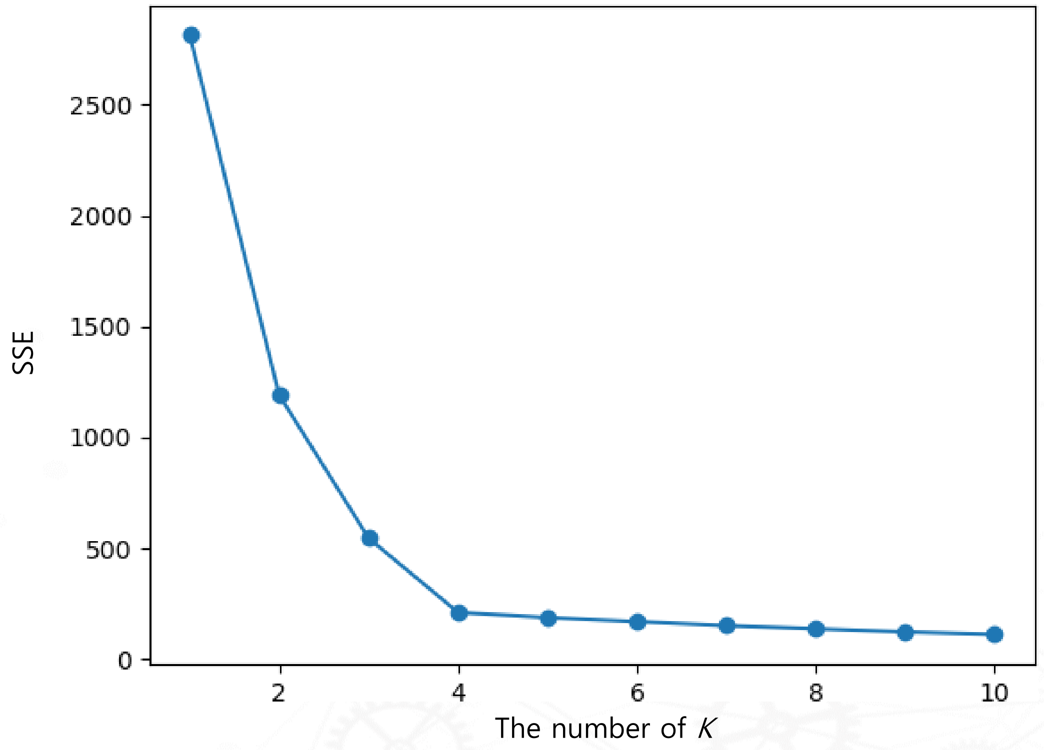


K-means Clustering

K-means Clustering

❖ Elbow method

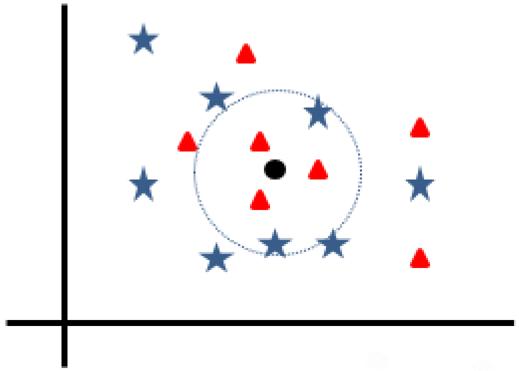
- To select *k* at the largest bend by graphing the Sum of Squared Errors (SSE) in a cluster for clustering results according to the number of clusters *k*



k-nearest Neighbor

❖ k-nearest neighbor, KNN

- In a situation where data with (input, output) is given, how to use the result information for the nearest *k* data that knows the result when estimating the result for a new input

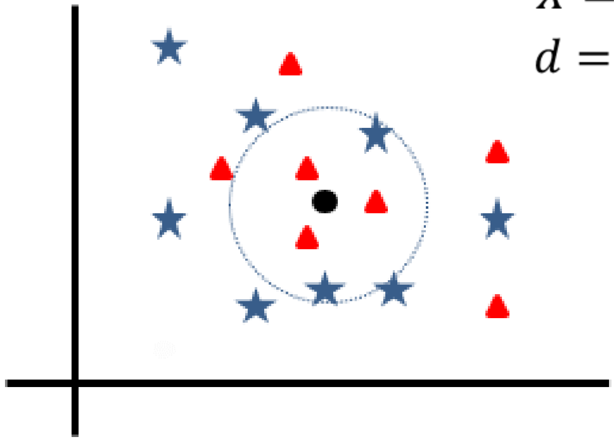


- Calculate the distance between query and data
- Efficiently search neighbor
- Estimate results from k close neighbors

k-nearest Neighbor

❖ k-nearest neighbor, KNN

- Euclidian distance



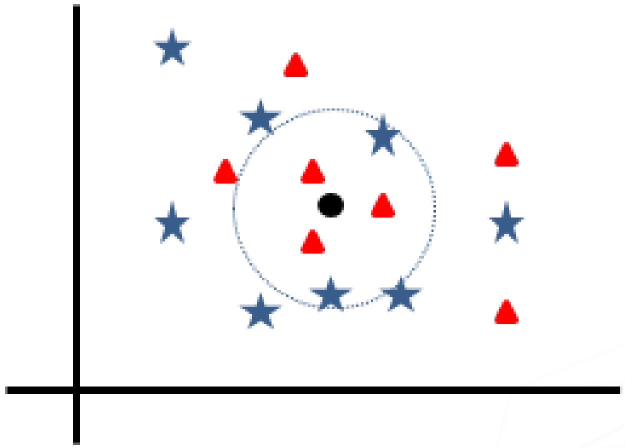
$$X = (x_1, x_2, \dots, x_n) \quad Y = (y_1, y_2, \dots, y_n)$$

$$d = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2$$

k-nearest Neighbor

❖ To estimate results from the last k

- Classification
  - Output is categorical value
  - Majority voting: selecting a large number of categories
- Regression analysis
  - Output is numeric value
  - Average: Last k mean
  - Weighted sum : Use weights that are inversely proportional to distance



k-nearest Neighbor

- Challenging issues
  - Regularization
    - To convert all values to values between 0 and 1 after fixing the minimum value to 0 and the maximum value to 1
    - To convert to z-score how far away from the mean using mean and standard deviation
  - How to select the number of K
    - K is too small : Overfitting problem
    - K is too large : Underfitting problem

Apply What You Learned

❖ Image segmentation with k-means clustering

1. Download Berkeley segmentation data set (Tiger) (Airplane)
2. Try your k-means with at least 3 different values of k (e.g., k=2,3,5)
3. Describe what you found

Fig.1 Orginal Airplane figure

Fig.2 Orginal Tiger figure

Apply What You Learned

❖ Image segmentation with k-means clustering

1. Download Berkeley segmentation data set (Tiger) (Airplane)
2. Try your k-means with at least 3 different values of k (e.g., k=2,5,10)
3. Describe what you found

Results

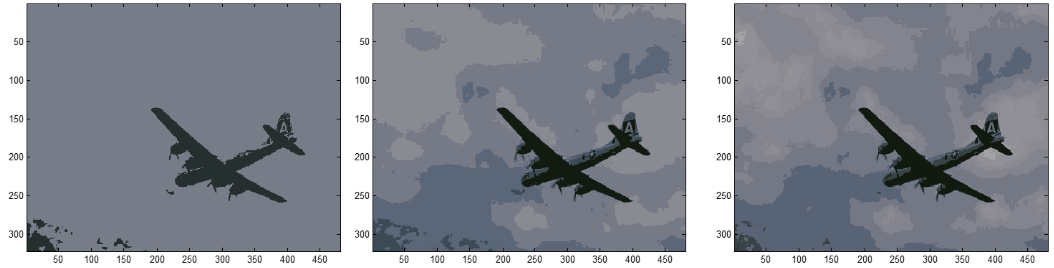


Figure 1. K=2, K=5, K=10, iteration= 10

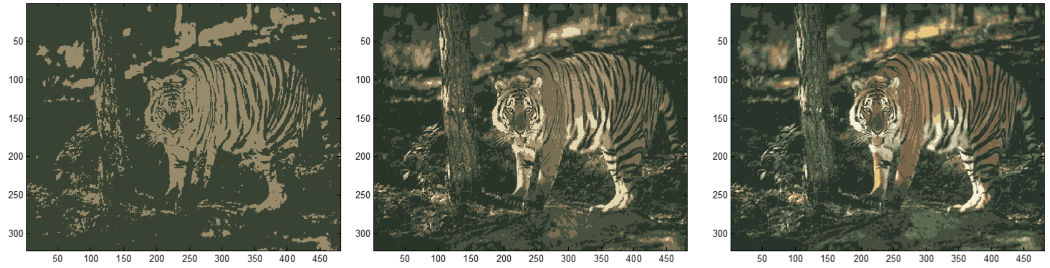


Figure 2. . K=2, K=5, K=10, iteration= 10