클린업 3주차

4팀 데이터마이닝

박정현
김지민
노정화
염예빈
전규리

1 Introduction

지도학습의 목적

예측성능 << X와 Y간의 관계를 추론

아이디어를 제공하는 GAM모델 소개

1 추론

2 예측

1 Basis Function

1. 비선형 모델
   - Basis Function
   - Cubic Spline
   - GAM
2. 군집화
3. 클린업을 마치며

기저함수

What is Basis function?

무한 공간에서 기저가 되는 단위함수

기저 : (2주차!) n차원을 span하는 n개의 선형적으로 독립인 벡터

무한대로 확장해보자!

Basis function : 무한차원 벡터공간 (Hilbert Space or 함수공간)의 기저가 되는 단위적(기본적) 함수

1 Basis Function

기저함수

About Basis function,,,

비선형 모델링의 관점에서…

함수의 정의역	Input matrix
기저함수	p차원 공간에서 스칼라 공간으로의 transformation

$$h_m(X) : \mathbb{R}^p \rightarrow \mathbb{R} \; is \; the \; m^{th} \; transformation \; of \; X.$$

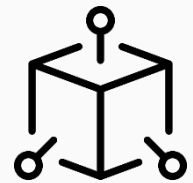
1 Basis Function

기저함수

$$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$$

- f: X 와 y의 관계
  - M개의 basis function들의 결합
  - → *Linear Basis Expansion*

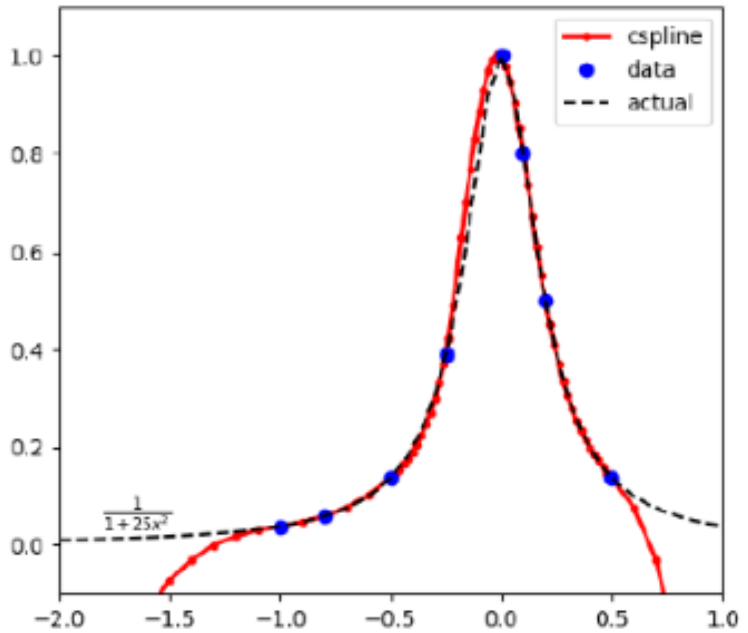
Cubic Spline



1 Cubic Spline

spline

What is spline?



전체 f를 smoothing 하는 Basis function

1 Cubic Spline

1. 비선형 모델
   - Basis Function
   - ▶ Cubic Spline
   - GAM
2. 군집화
3. 클린업을 마치며

Cubic spline

What is Cubic spline?

- Piecewise polynomial 중 하나
- X직선을 구간으로 분할하여 3차 다항식을 적합시키는 것

보통 4차이상은 잘 쓰지 않는다.

A. 현실에 존재하는 x와 y의 관계를 표현하기에 최대 삼차항까지 고려하는 것만으로 충분하기 때문!

1 Cubic Spline

1. 비선형 모델
   - Basis Function
   - ▶ Cubic Spline
   - GAM
2. 군집화
3. 클린업을 마치며

Cubic spline

What is Cubic spline?

Basis function으로
input matrix H을 만듦

베타계수 벡터(B)와 Y가
선형관계로 표현

LSE를 사용가능!

Generalized additive model

What is GAM?

Genaralized Additive Model

1 X변수의 가산성(Additivity)는 유지

각 변수에 비선형 함수를 이용하여 적합

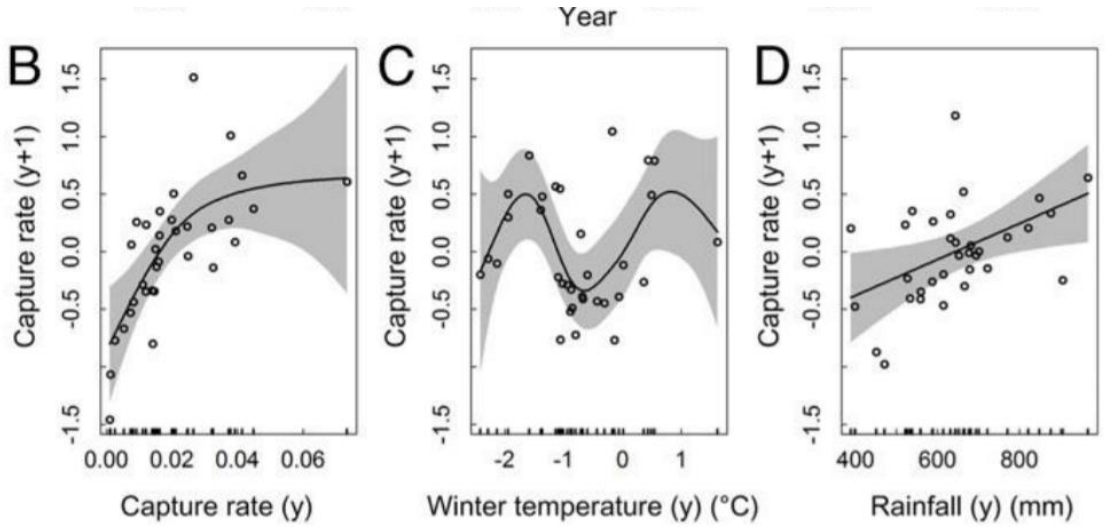
2 표준선형모델을 확장하는 일반적 체계를 제공!

분류, 회귀문제에 모두 적용가능

1 GAM의 활용

1. 비선형 모델
   - Basis Function
   - Cubic Spline
   - GAM
2. 군집화
3. 클린업을 마치며

Semiparametric model



$\mathbf{X_1}$: t시점의 capture rate

$\mathbf{X_2}$: t시점의 겨울의 기온

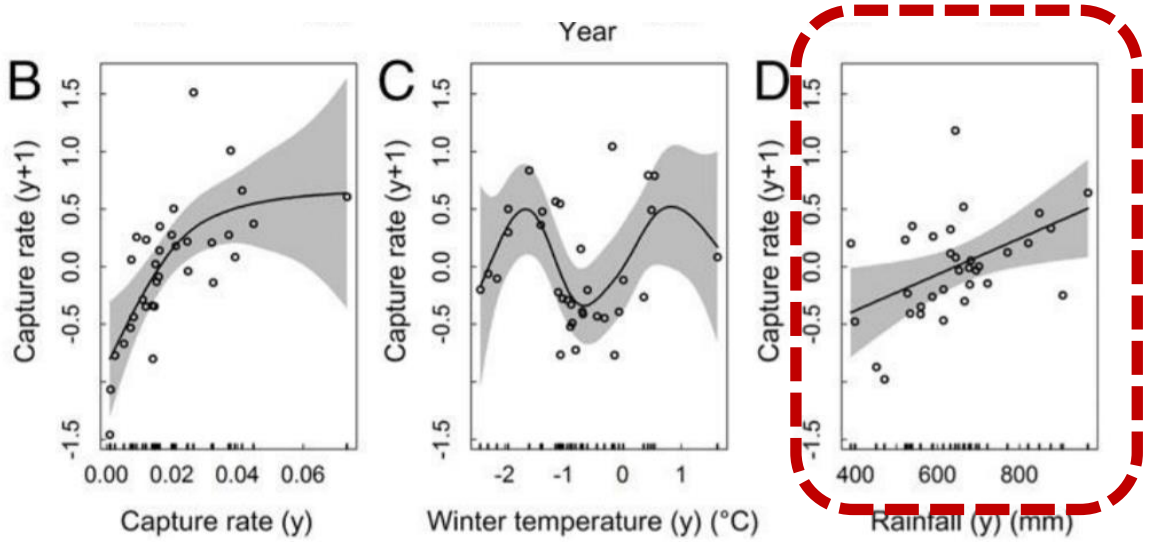
$\mathbf{X_3}$: t시점의 강수량

$\mathbf{Y}$: t+1시점에 전염병이 잡힐 capture rate

1 GAM의 활용

1. 비선형 모델
   - Basis Function
   - Cubic Spline
   - GAM
2. 군집화
3. 클린업을 마치며

Semiparametric model

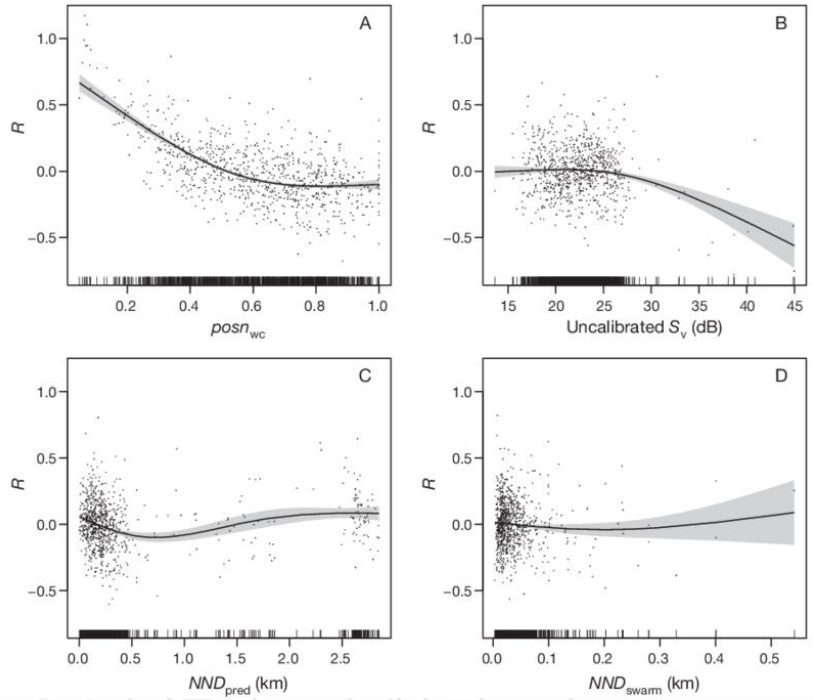


강수량(X_3)에 대해서만 일차선형회귀를 적합시킨다.

Capture rate at t+1

$$= \beta_0 + \beta_1(rainfall) + f_1(capture\ rate\ at\ t) + f_2(winter\ temperature\ at\ t)$$

1 GAM의 활용



Semiparametric regression을 구축하는 데
데이터 포인트의 추세를
"X 변수별로 smooth하게 보여준다"는 점에서
GAM은 매우 실용적!

2

군집화

Clustering

2 Before Clustering….

1. 비선형 모델
2. 군집화
   - Before Clustering
   - K-means & K-medoids
3. 클린업을 마치며

군집화(Clustering)란?

What is CLUSTERING?

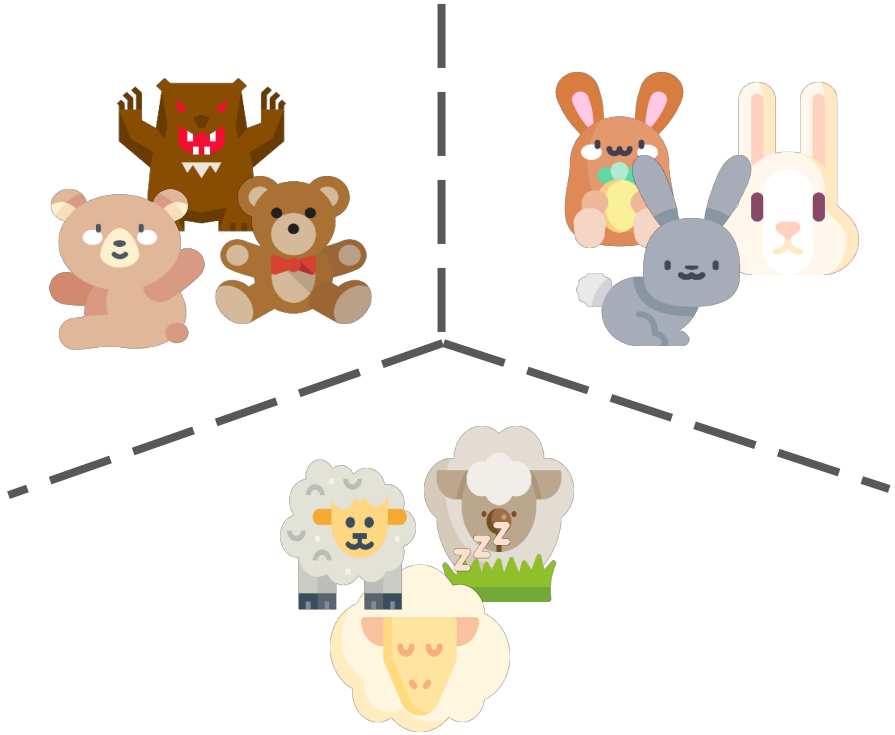
비지도학습(Unsupervised Learning) 기법

한 feature의 factor level이 너무 많을 때,
재범주화를 목적으로 쓰임

2 Before Clustering….

군집화(Clustering)란?

What is CLUSTERING?



다양한 오브젝트들 가운데

속성이 유사한 오브젝트끼리

지정된 개수의 집합을 묶어주는

방법론

비유사성 측정도로서의 거리

Proximity 의 측도 :

각 속성별로 오브젝트 간 **비유사성(dissimilarity)**을 계산하여 모두 합한 것

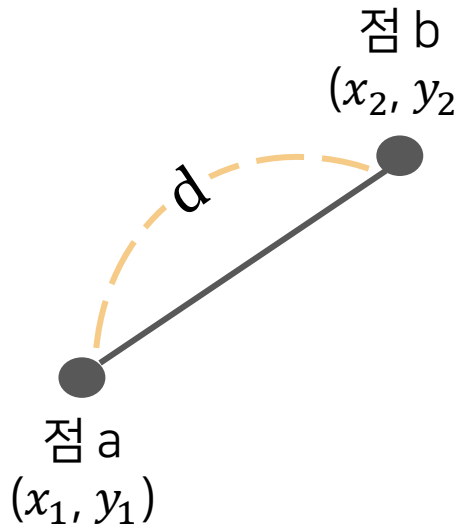
$$X = \begin{bmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \cdots & x_{np} \end{bmatrix}$$

- n개의 행 : 각 unique한 오브젝트
- p개의 열 : 오브젝트의 속성

비유사성 측정도로서의 거리

What is Euclidean Distance?

두 점 a와 b를 잇는 가장 짧은 거리



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

∵ *피타고라스 정리*

2 Before Clustering….

Silhouette & Elbow

[최적의 **군집 개수**를 정할 때에 고려해야 할 값]

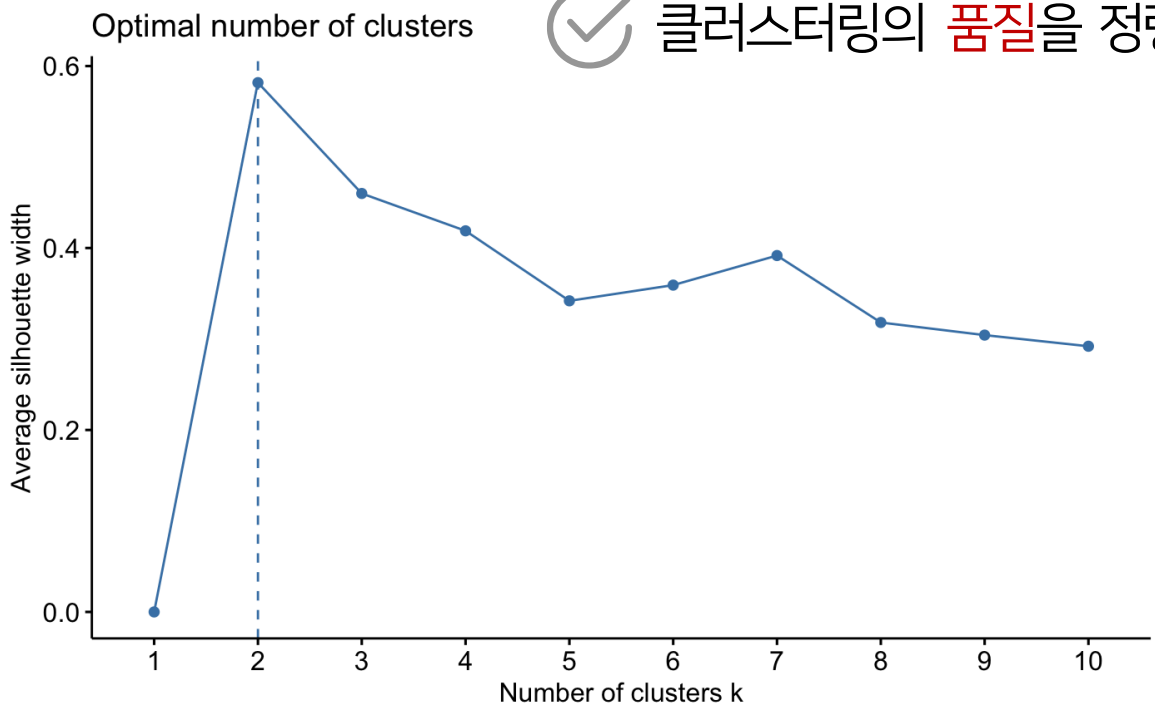
Silhouette

Elbow

Silhouette & Elbow

What is Silhouette?

클러스터링의 품질을 정량적으로 계산

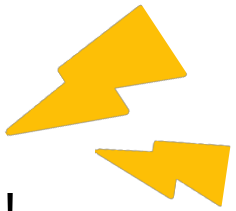


Silhouette & Elbow

What is Elbow?

클러스터 **내**의 분산이

최소가 되도록 하라! 실시!

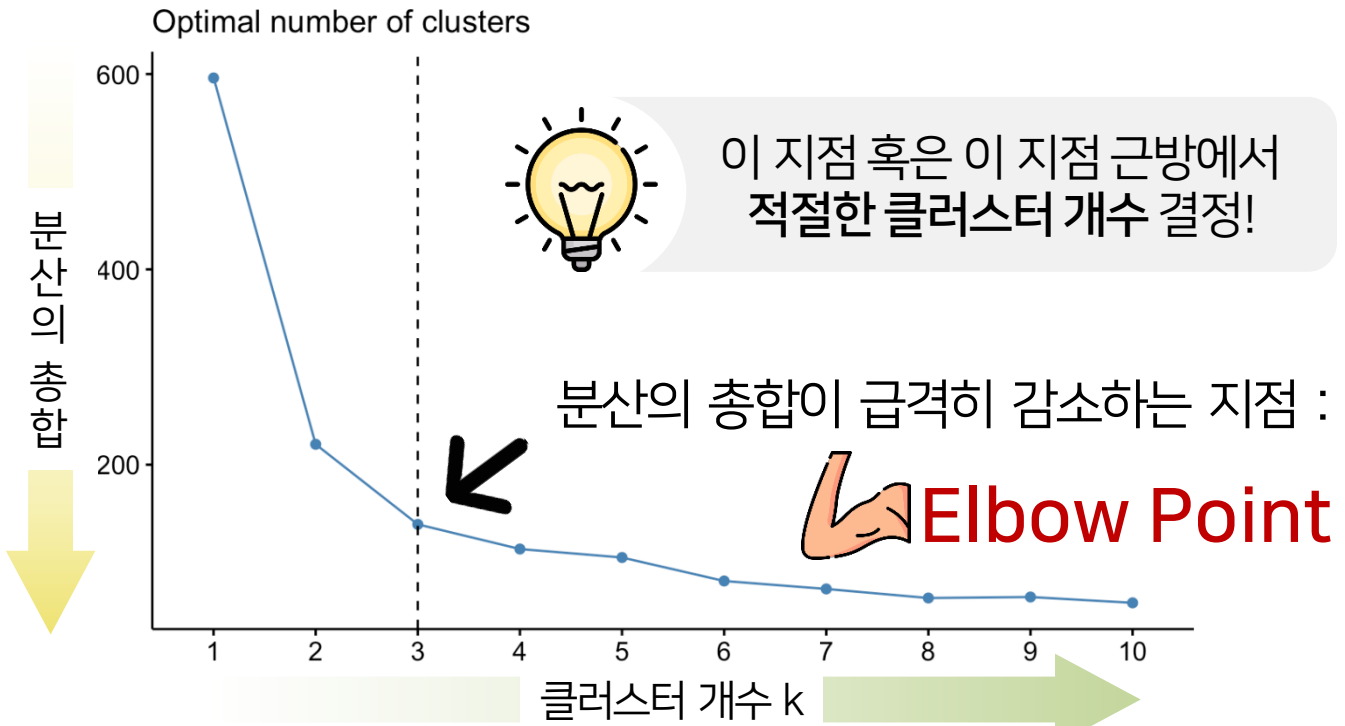


적합도 지표: 각 클러스터별로 클러스터 내의 분산을 계산하여 더한 것

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}\|x_i - x_{i'}\|^2 = \sum_{k=1}^{K} n_k \sum_{C(i)=k}\|x_i - \bar{x}_k\|^2$$

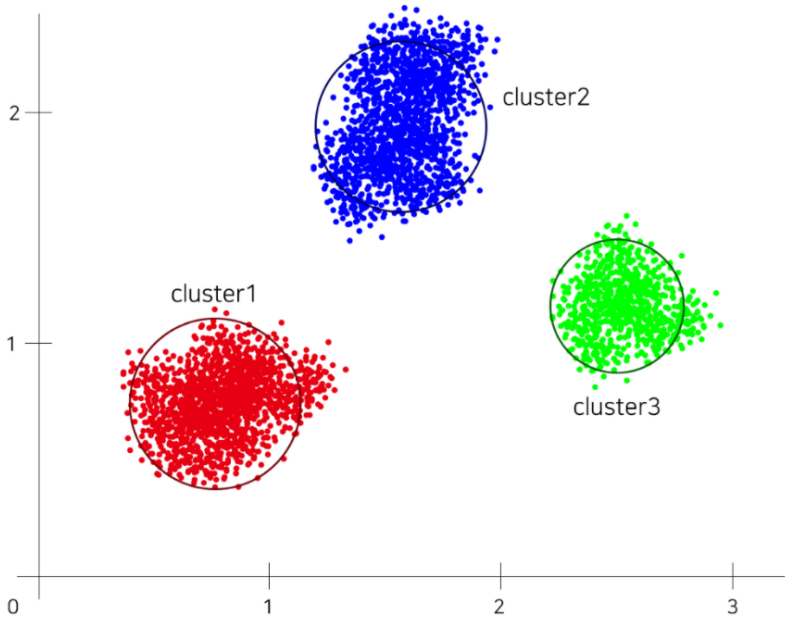
Silhouette & Elbow

What is Elbow? : WSS scree plot



Algorithm

What is K-means?



K-means

가장 흔하게 쓰이는 클러스터링 기법

K-Means & K-Medoids

Algorithm

1

군집의 수 k를 2로 설정하고,
랜덤으로 아무 포인트나 군집의 중심점으로 잡았다고 하자.

2 K-Means & K-Medoids

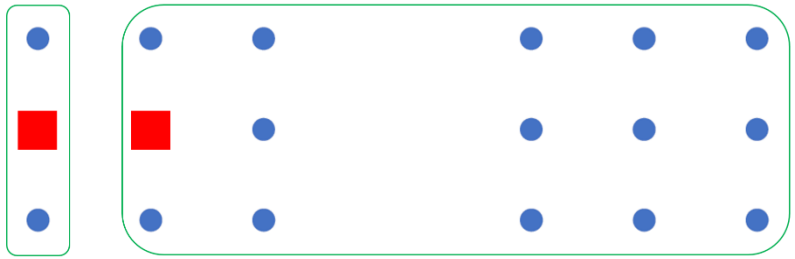
Algorithm

1

군집의 수 k를 2로 설정하고,
랜덤으로 아무 포인트나 군집의 중심점으로 잡았다고 하자.

2

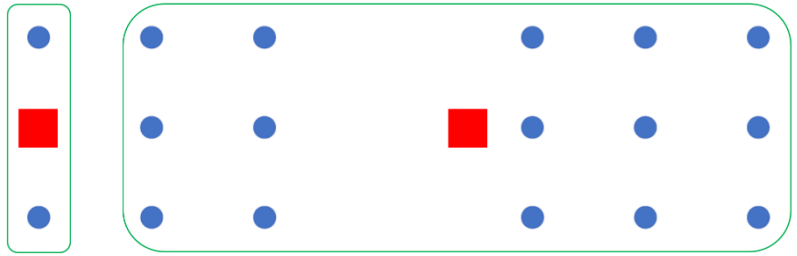
모든 개체들을 **가장 가까운 중심의 군집**으로 할당한다.



Algorithm

3

이번엔 **중심**을 군집 경계에 맞게 업데이트해준다.
이때 중심지점은 각 데이터 포인트의 평균을 사용한다.

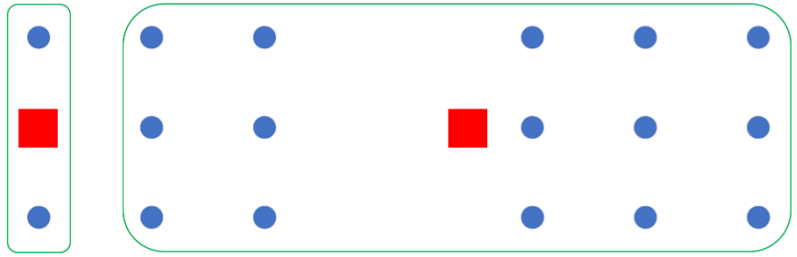


K-Means & K-Medoids

Algorithm

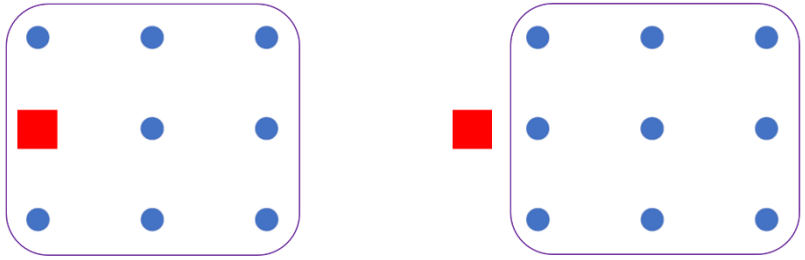
3

이번엔 **중심**을 군집 경계에 맞게 업데이트해준다.
이때 중심지점은 각 데이터 포인트의 평균을 사용한다.



4

다시 모든 개체들을 **가장 가까운 중심점**의 군집으로 할당해준다.

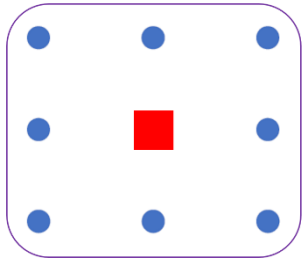


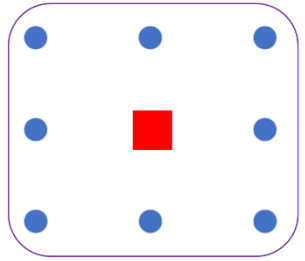
K-Means & K-Medoids

Algorithm

5

중심을 다시 업데이트해준다.





중심점이 **고정된 점**으로 수렴할 때까지 위 과정을 반복해주자.

혹 수렴하지 않더라도 사용자가 정한 **반복수**를 채우게 되면 학습은 끝!

Algorithm

What is K-medoids?

[중심점을 설정하는 방식]

K-means	각 클러스터에 속하는 오브젝트의 속성(X)을 **평균** 낸 벡터
K-medoids	각 속성값의 **중앙값** 벡터