

AI 특강

Prof. Jibum Kim

Department of Computer Science & Engineering

Incheon National University

Slide 1

■ Polynomial regression

-
- 지금까지는 입력 변수들과 출력 변수의 관계가 선형 (linear)관계, 2차원 직선, 라고 가정하였다
 - 하지만, 이러한 가정은 항상 성립하지 않는다
 - 이번에는 비선형 관계인 경우의 회귀분석인 다항 회귀 (polynomial regression)에 대해서 배워본다

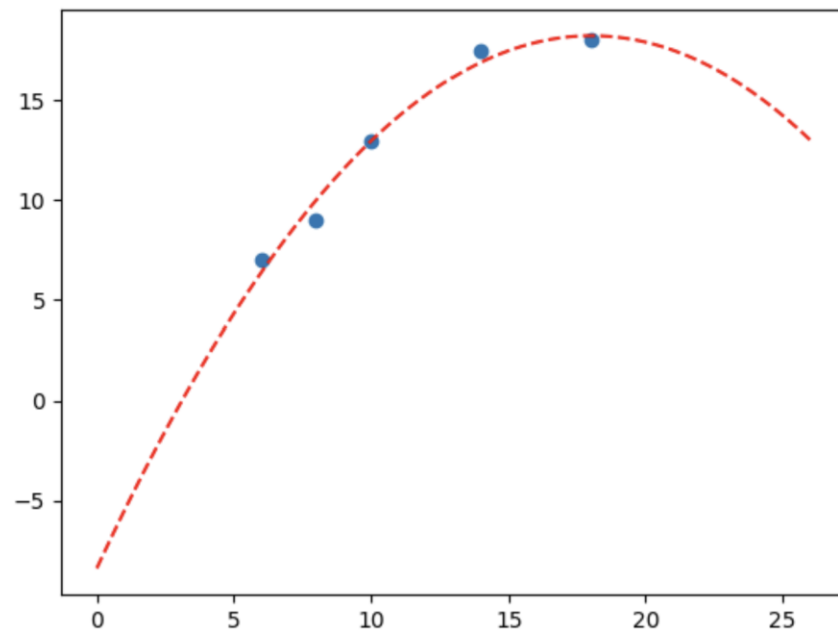
-
- 1. Quadratic regression (regression with a second order polynomial, 2차 다항식 회귀분석)
 - 입력 변수: x , 출력 변수: y
 - 수식: $y = \alpha + \beta_1 x + \beta_2 x^2$
 - 이번에는 best fit을 통해서 알아내야 하는 계수가 3개로 늘어남

-
- Scikit-learn 의 Polynomial Regression 모델은 비선형 데이터를 학습하기 위해, 선형 회귀 모델을 사용
 - $y = \alpha + \beta_1 x + \beta_2 x^2$
 - 즉, 원래 입력 data x 와 이를 제공한 x^2 을 먼저 만든 후에 다시 linear regression 수행

-
- 실습
 - 아래와 같은 데이터를 2차 다항식으로 fitting
 - $y = \alpha + \beta_1 x + \beta_2 x^2$, 목표: best fit하는 계수들을 polynomial regression으로 찾음

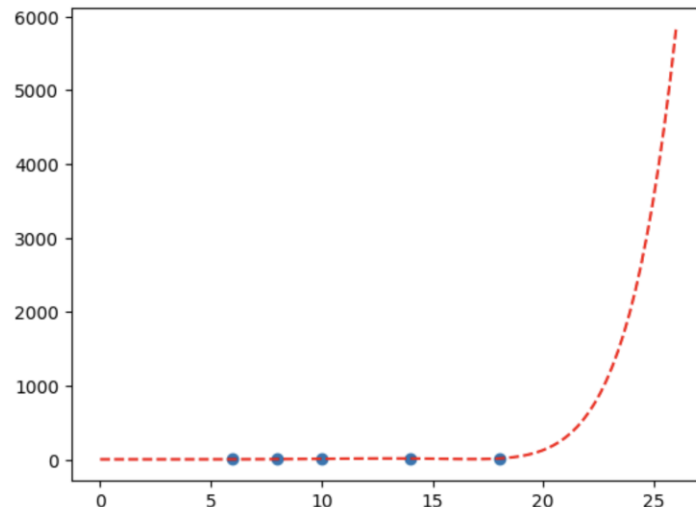
- $y = \alpha + \beta_1 x + \beta_2 x^2$
- Intercept: $\alpha = -8.39 \dots$
- $\beta_1 = 2.95 \dots$
- $\beta_2 = -0.08 \dots$

```
[ -8.39765458]  
[[ 2.95615672 -0.08202292]]
```

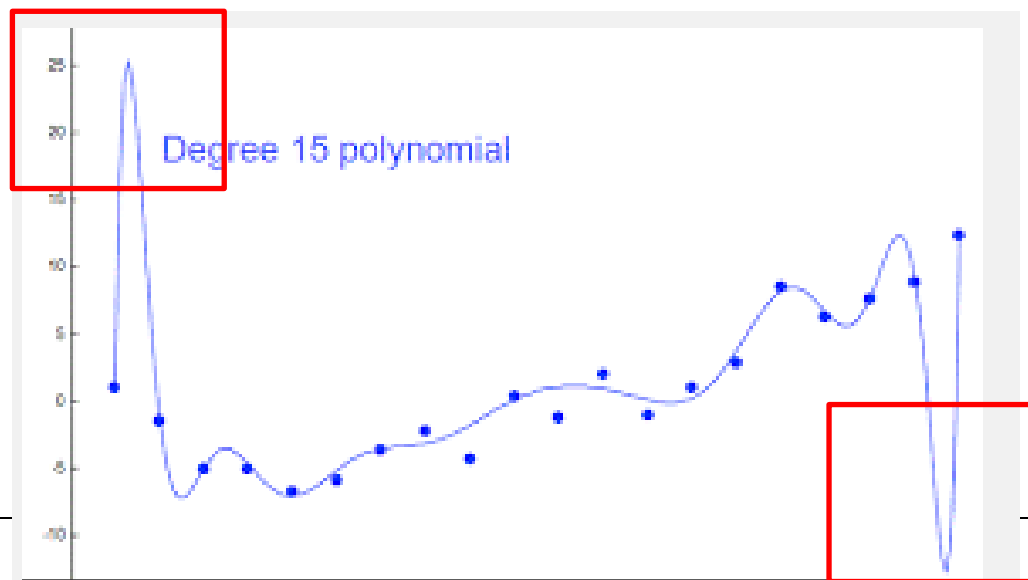


■ Overfitting-underfitting

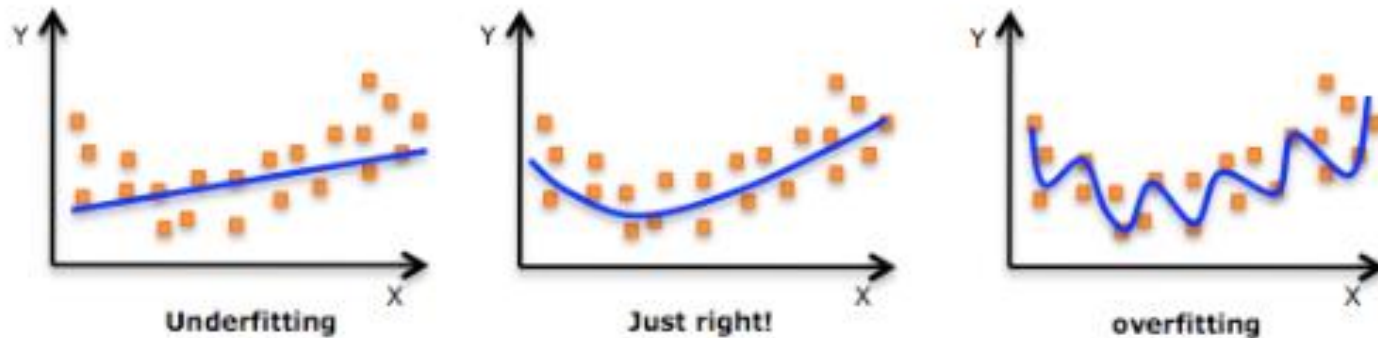
- 다항식의 차수를 올리다보면 거의 오차가 없게 만들 수 있다. 학습 데이터를 거의 완벽하게 지나게 (아래 그림)
- 아래의 경우에는 9차 다항식
- We created an extremely complex model that fits the training data exactly, but fails to approximate the real relationship
- This problem is called **over-fitting**



- Overfitting: 과도하게 training data를 regression함으로써 **미래의 데이터값을 예측하는데 실패할 확률이 높은 문제**
- (현재 데이터는 잘 근사화 하겠지만)



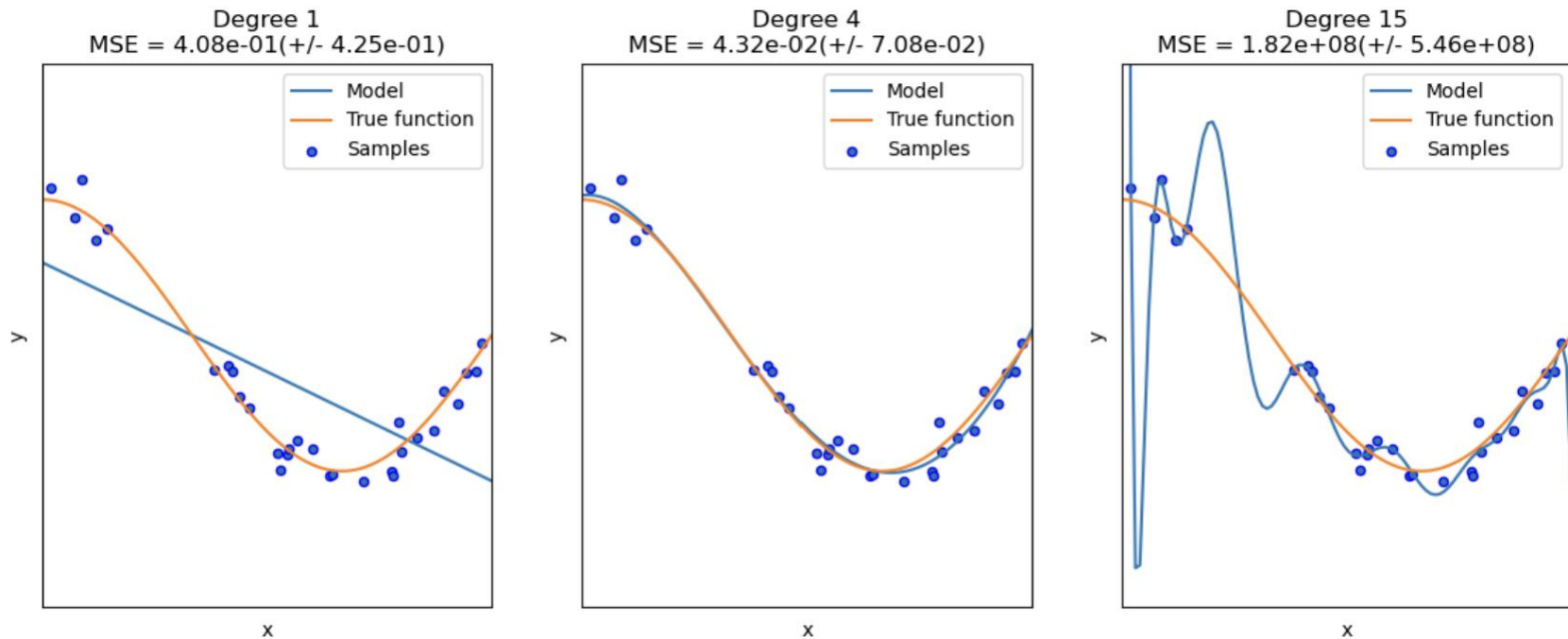
- 아래의 training 데이터들을 살펴보자



- Underfitting: 근사화한 것이 너무 간단해서 Fitting이 잘 안 된다
- Overfitting: training data에 대해서는 근사화가 잘 됨
근사화한 것의 차수가 너무 높아서 새로운 데이터가 조금의 변화 (variance)가 생기면 부정확한 결과 생김

-
- Scikit-learn의 over-fitting vs underfitting
 - http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

■ 왼쪽: under fitting, 오른쪽: over-fitting



■ 군집화 (clustering)

•Scikit-learn의 machine-learning 방법 분류

•군집화 (clustering)

•군집화는 매우 다양한 응용 분야를 가지고 있다

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.

Algorithms: *SVM, nearest neighbors, random forest, ...* — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: *SVR, ridge regression, Lasso, ...*

— Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: *k-Means, spectral clustering, mean-shift, ...* — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: *PCA, feature selection, non-negative matrix factorization.* — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Modules: *grid search, cross validation, metrics.* — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: *preprocessing, feature extraction.* — Examples

-
- 군집화란?
 - Clustering is the task of grouping a set of objects in such a way that **objects in the same group (cluster) are more similar to each other than to those in other groups**
 - 대표적인 군집화 알고리즘: K-means 알고리즘

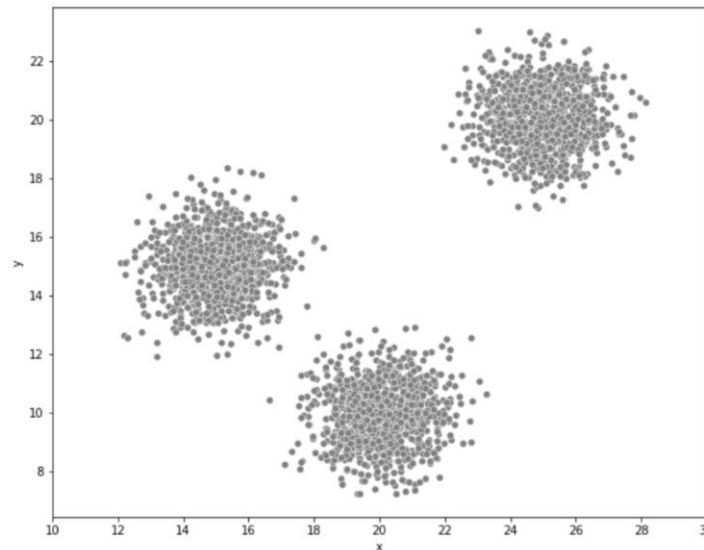
■ 군집화 사용 예시

- 어떤 사람이 몇년간 쇼핑몰을 성공적으로 운영해 왔다 (수백만명의 고객을 가지고 있다). 그 동안 한 종류의 홍보 판플렛을 만들어 발송했는데 이제부터는 고객의 취향을 분석하여 **4-6종류의 판플렛을** 만들어 맞춤 홍보를 하고자 한다. 일종의 개인화 (personalization) 홍보 전략이다. 고객에 대한 각종 정보는 DB에 저장되어 있어 이것을 기초 자료로 활용 가능하다
- 고객 정보는 월평균 구매액, 선호하는 물품의 종류와 수준, 결제 방법, 반품 성향, 직업, 성별, 나이, 거주 지역등 다양하다.
- **Q) 수백만명의 고객이 있을때 이를 어떻게 4-6개의 그룹으로 나눌수 있을까?**

-
- 해결 방법: 유사한 샘플(데이터)들끼리 모아서 4-5개의 그룹으로 만든다
 - 여기서 유사하다는 것은 데이터와의 거리가 가깝다는 것을 의미 한다
 - 이 각각의 그룹을 군집 (cluster)라고 하고 이 경우에는 4-5개의 군집이 만들어 진다
 - 여태까지 배운 Supervised learning과 다르게 오늘 배울 군집화 (clustering)은 unsupervised learning으로써 데이터 정보만 있을 뿐 이 데이터의 분류는 주어져 있지 않다

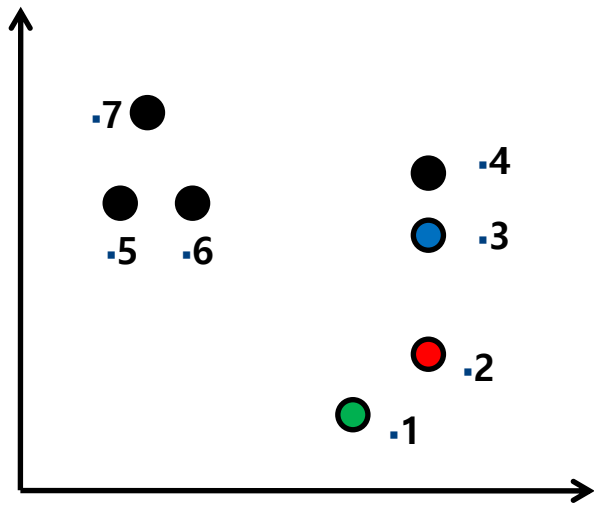
K-means 알고리즘

- K-means 군집화 알고리즘은 주어진 데이터를 K개의 클러스터 (군집)으로 묶는 알고리즘이다
- 아래와 같은 데이터의 경우 $K=3$ 이라고 생각할 수 있다
- K-means 군집화 알고리즘의 경우 기본적으로 입력으로 사용자가 K를 정하지만 뒤에서 배울 elbow 방법과 같은 방법으로 K를 정할수도 있다



-
- 다음 예제는 총 7개의 데이터에 대해서 K-means 군집화 알고리즘을 실행한 예제이다
 - 이 경우에 $K=3$ 으로 사용자가 정했다

• k (군집개수)=3이고 7개의 데이터가 있는 경우를 생각해 보자



1. x_1, x_2, x_3 를 초기 군집 중심으로 삼고

• 이를 z_1, z_2, z_3 라고 놓는다

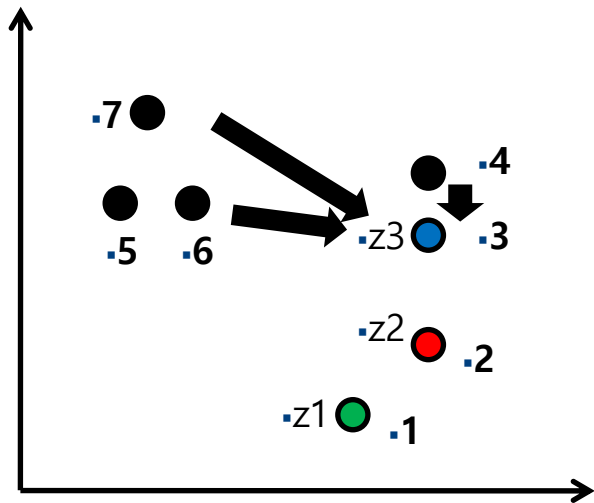
• $z_1 = (18, 5)$

• $z_2 = (20, 9)$

• $z_3 = (20, 14)$

• $x_1 = (18, 5), x_2 = (20, 9), x_3 = (20, 14), x_4 = (20, 17), x_5 = (5, 15), x_6 = (9, 15), x_7 = (6, 20)$

• k (군집개수)=3이고 7개의 데이터가 있는 경우를 생각해 보자



• 2. 각 데이터들은 z_1, z_2, z_3 중

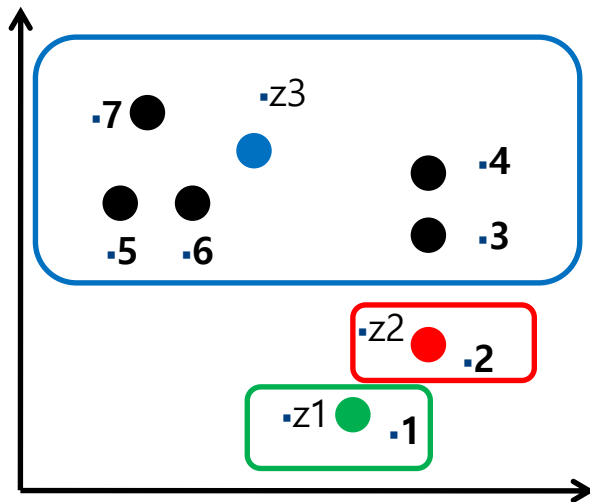
• 가장 가까운 군집 중심에 각각 배정된다

• $\{x_1\}$ 은 z_1

• $\{x_2\}$ 은 z_2

• $\{x_3, x_4, x_5, x_6, x_7\}$ 은 z_3 에 배정된다

• k (군집개수)=3이고 7개의 데이터가 있는 경우를 생각해 보자



• 2. $\{x_1\}$ 은 z_1

• $\{x_2\}$ 은 z_2

• $\{x_3, x_4, x_5, x_6, x_7\}$ 은 z_3 에 배정된다

• 3. 2에서 배정된 데이터들을 이용해 새로운

• 군집 중심을 찾고 (위치 평균)

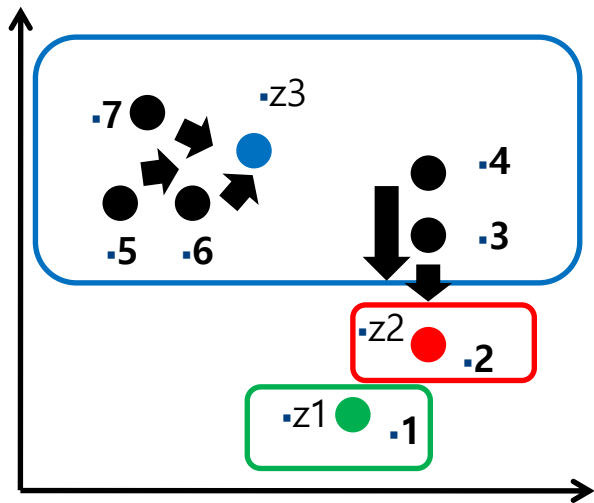
• 이를 새로운 z_1, z_2, z_3 라 놓는다

• $z_1 = (18, 5)$

• $z_2 = (20, 9)$

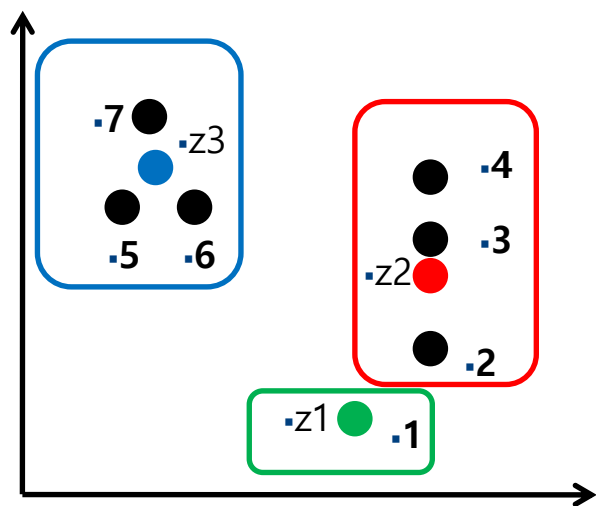
• $z_3 = (x_3 + x_4 + x_5 + x_6 + x_7) / 5 = (12, 16.2)$

• k (군집개수)=3이고 7개의 데이터가 있는 경우를 생각해 보자



- 앞의 step 2를 반복한다. 즉
- 각 데이터들은 가장 가까운 군집 중심에
- 각각 배정된다
- $\{x_1\}$ 은 z_1
- $\{x_2, x_3, x_4\}$ 는 z_2
- $\{x_5, x_6, x_7\}$ 은 z_3 에 배정된다

• k (군집개수)=3이고 7개의 데이터가 있는 경우를 생각해 보자



• $\{x_1\}$ 은 z_1

• $\{x_2, x_3, x_4\}$ 는 z_2

• $\{x_5, x_6, x_7\}$ 은 z_3 에 배정된다

• 앞의 step 3를 반복한다. 즉

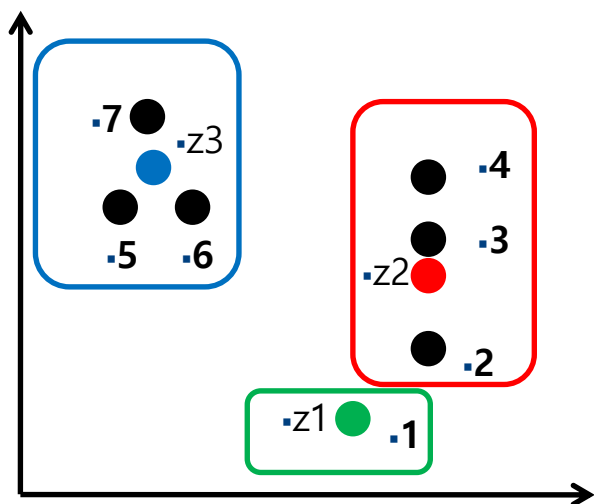
• 배정된 데이터들의 군집 중심을 찾고

• 이를 새로운 z_1, z_2, z_3 라 놓는다

• $z_1 = (18, 5)$, $z_2 = (x_2 + x_3 + x_4)/3 = (20, 13.333)$

• $z_3 = (x_5 + x_6 + x_7)/3 = (6.667, 16.667)$

• k (군집개수)=3이고 7개의 데이터가 있는 경우를 생각해 보자



• 그 다음에 step 2와 3를 반복한다

• Step 2: 각 데이터들은 가장 가까운

• 군집 중심에 각각 배정된다

• 변화가 있는가? 없음

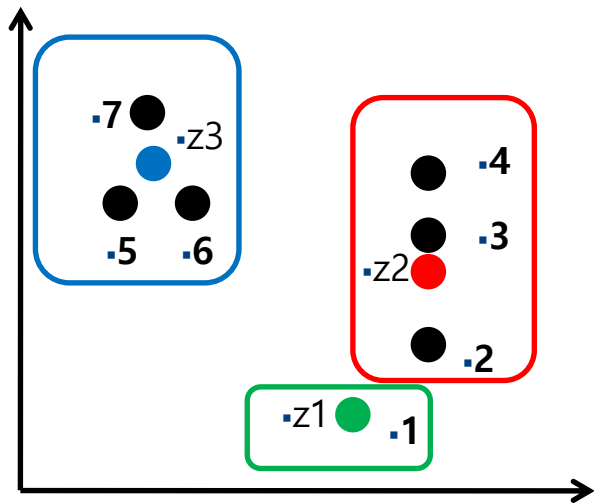
• Step 3: 배정된 데이터들의 각각의

• 군집 중심을 찾고 (위치 평균)

• 이를 새로운 $z1, z2, z3$ 라 놓는다

• 따라서 step3의 변화도 없음

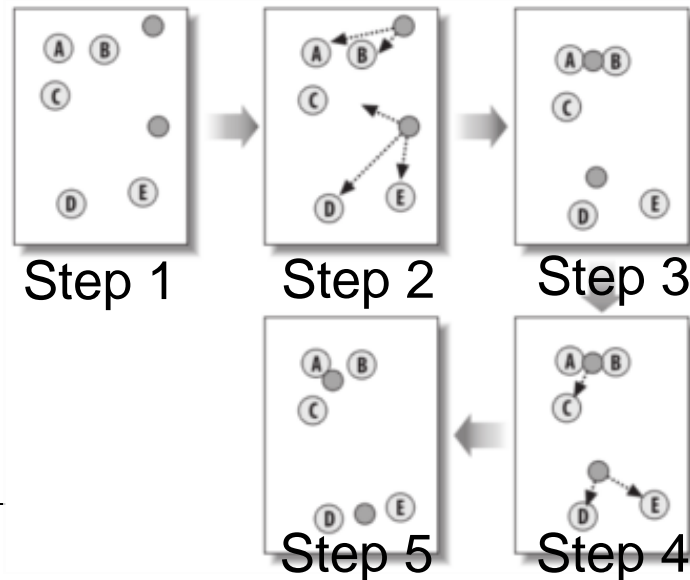
• k (군집개수)=3이고 7개의 데이터가 있는 경우를 생각해 보자



- 즉 최종적으로 왼쪽의 데이터들을
- $K=3$ (군집개수=3)으로 군집화 한
- 경우
- $\{x_1\}$, $\{x_2, x_3, x_4\}$, $\{x_5, x_6, x_7\}$ 과 같이
- 세 개의 군집으로 군집화 할 수 있다

K-means 알고리즘

- **K=2이고 5개의 데이터 (A, B, C, D, E)가 있는 경우**
- 1. 임의의 위치에 군집 중심 2개가 위치
- 2. 각각의 데이터가 가장 가까운 군집 중심을 찾음. 이 경우에는 A, B는 위쪽 군집 중심과 가깝고, C, D, E는 아래쪽 군집 중심과 가까움
- 3. A, B의 평균 위치가 새로운 군집 중심의 위치, C, D, E의 평균 위치가 새로운 군집 중심의 위치
- 4. 다시 2 반복. 이 경우 A, B, C는 위쪽 군집 중심과 가까움. D, E는 아래쪽 군집 중심과 가까움
- 5. 다시 3 반복.
- **언제까지 2,3 반복?**

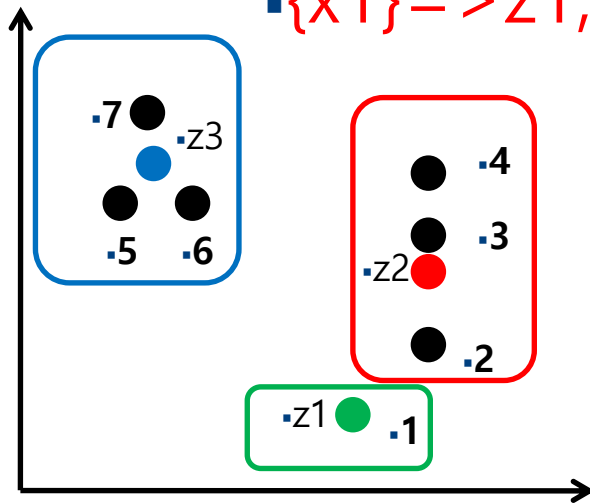


- SSE와 Elbow method

-
- 군집화가 잘 되었다라는것은 어떻게 판단하나?
 - K-means 에서 군집화가 잘 되어 있다 라는건 데이터들이 군집 중심 근처에 모여 있는 것이다. 데이터 들이 군집내에서 퍼져 있는건 군집화가 잘되지 않았다고 생각할 수 있다
 - 군집화가 잘되었나 판단하는 척도 (metric) 중 하나는
 - 각각의 데이터가 그것에 가장 가까운 군집 중심 사이의 거리 (혹은 거리의 제곱)을 모두 더해 보는 것이다
 - 이를 **Sum of squared error (SSE)**라고도 한다

예를 들어 마지막 최종 군집화 결과의 경우

$\{x_1\} \Rightarrow z_1$, $\{x_2, x_3, x_4\} \Rightarrow z_2$, $\{x_5, x_6, x_7\} \Rightarrow z_3$



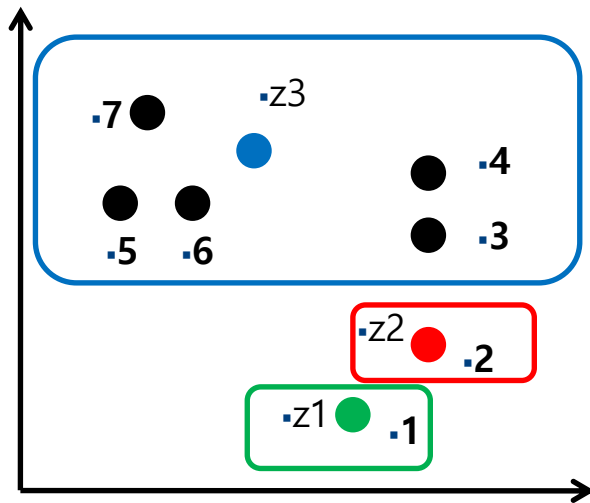
•예를 들어 마지막 최종 군집화 결과의 경우

• $\{x_1\} \Rightarrow z_1$, $\{x_2, x_3, x_4\} \Rightarrow z_2$, $\{x_5, x_6, x_7\} \Rightarrow z_3$

•SSE: 제곱 오류

•SSE=58

• 최종 군집화 바로 이전의 경우



• $\{x_1\} \Rightarrow z_1, \{x_2\} \Rightarrow z_2, \{x_3, x_4, x_5, x_6, x_7\} \Rightarrow z_3$

• $SSE=244.8 \Rightarrow$ 즉 최종 군집화된 결과가 조금 오류가 더 적다!

-
- Iris dataset을 이용한 군집화 예제

■ Iris dataset

총 150개의 데이터

데이터설명 : 아이리스(붓꽃) 데이터에 대한 데이터이다. 꽃잎의 각 부분의 너비와 길이등을 측정한 데이터이며 150개의 레코드로 구성되어 있다. 아이리스 꽃은 아래의 그림과 같다. 프랑스의 국화라고 한다.



필드의 이해 :

데이터의 이해를 돕기 위해 포함된 6개의 변수에 대하여 간략하게 설명한다.

총 6개의 필드로 구성되어있다. caseno는 단지 순서를 표시하므로 분석에서 당연히 제외한다.

2번째부터 5번째의 4개의 필드는 입력 변수로 사용되고, 맨 아래의 Species 속성이 목표(종속) 변수로 사용된다.

caseno	일련번호이다. (1부터 150까지 입력된다.)
Sepal Length	꽃받침의 길이 정보이다.
Sepal Width	꽃받침의 너비 정보이다.
Petal Length	꽃잎의 길이 정보이다.
Petal Width	꽃잎의 너비 정보이다.
Species	꽃의 종류 정보이다. setosa / versicolor / virginica 의 3종류로 구분된다.

```
from sklearn.datasets import load_iris
```

```
iris = load_iris()  
iris
```

This code gives:

```
{'data': array([[5.1, 3.5, 1.4, 0.2],  
                [4.9, 3. , 1.4, 0.2],  
                [4.7, 3.2, 1.3, 0.2],  
                [4.6, 3.1, 1.5, 0.2],...  
'target': array([0, 0, 0, ... 1, 1, 1, ... 2, 2, 2, ...  
'target_names': array(['setosa', 'versicolor', 'virginica'], dtype='<U10'),  
...}]
```

실습