Algorithm Practice: K-means Clustering

Dept. of Computer Science and Engineering

2019-06-13

STCIS
Seoul National University of Science and Technology
Cryptography and Information Security Laboratory

What is Clustering?
클러스터링(군집화) 알고리즘

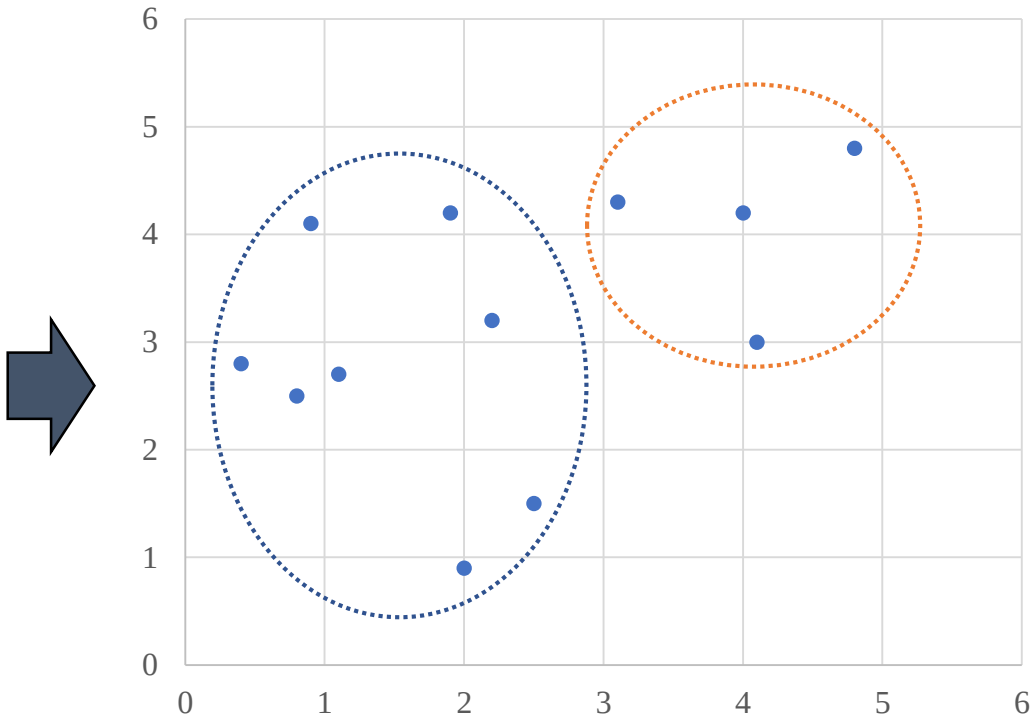
인공지능(Artificial Intelligence)

머신 러닝(Machine Learning)

클러스터링(Clustering)

예) 영화 장르별 사용자들의 선호도를 통해 사용사들의 성향을 분류

Username	Romance	Action
user1	1.1	2.7
user2	2.2	3.2
user3	4.1	3
user4	1.9	4.2
user5	0.4	2.8
user6	0.9	4.1
user7	0.8	2.5
user8	2.5	1.5
user9	4.8	4.8
user10	4	4.2
user11	2	0.9
user12	3.1	4.3



Clustering Algorithm
클러스터링(군집화) 알고리즘

클러스터(군집)을 형성하는 기준 = 점(데이터)과 점 사이의 거리

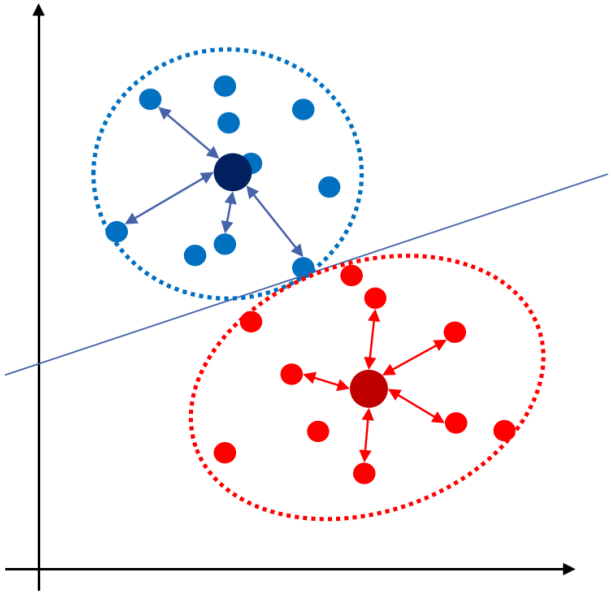
목적함수(Objective Function) = 거리의 총 합

n의 데이터(x), m개의 클러스터, d는 유클리드 거리 함수일 때,
목적함수 f는 동일 클러스터 내에 포함되는 데이터들과 클러스터 중심점(c)과의 거리의 총 합

$$f = \sum_{i=1}^{n} \sum_{j=1}^{m} d(x_i, c_j)$$

목적함수의 값이 최소화될 때, 클러스터링을 성공적으로 수행했다고 볼 수 있음.

목적함수의 값을 최소화하는 클러스터 중심(c_j)값들을 찾는 것이 알고리즘의 최종 목적

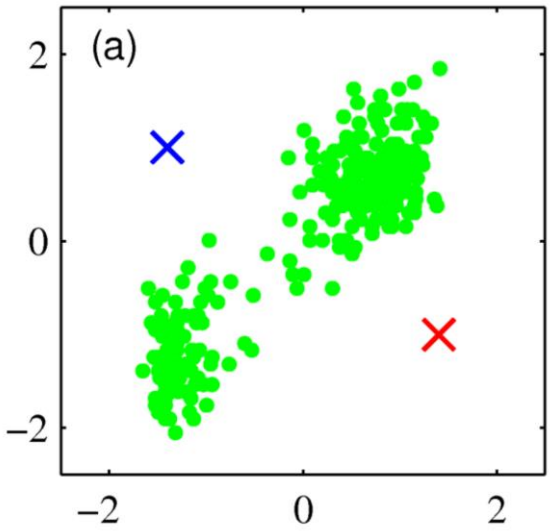


K-means Clustering Algorithm

K개의 클러스터를 이용하여, 데이터들의 평균점을 클러스터의 중심점으로 이용

Step 0: Initialization

클러스터의 중심점을 랜덤한 값으로 초기화 (혹은 각 데이터들을 랜덤한 클러스터에 배정)

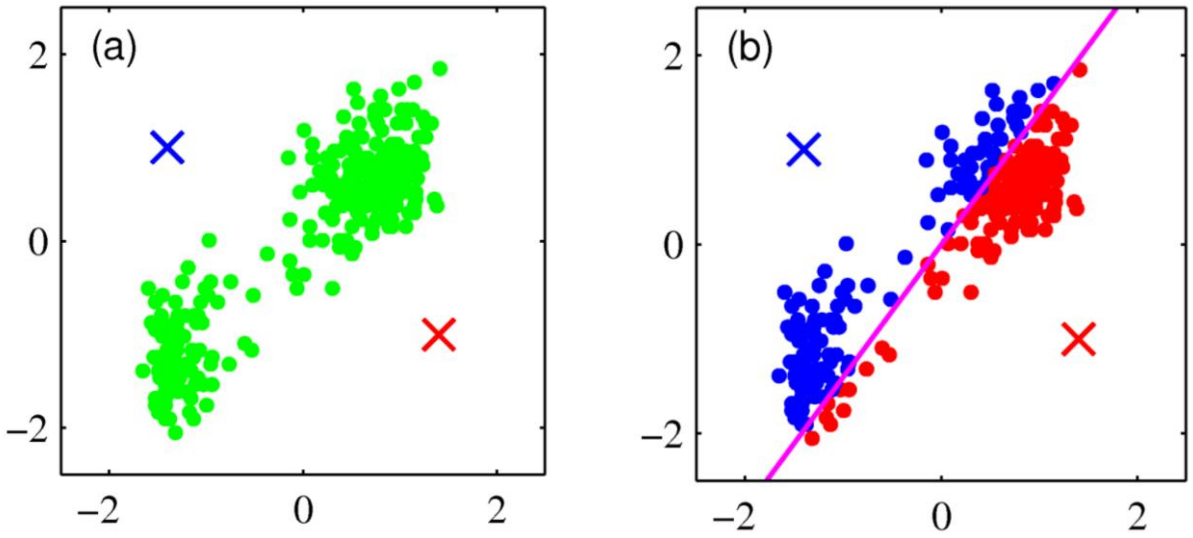


K-means Clustering Algorithm

K개의 클러스터를 이용하여, 데이터들의 평균점을 클러스터의 중심점으로 이용

Step 1: Expectation

모든 데이터들에 대해, 각 데이터와 각 클러스터의 중심점과의 거리들을 모두 구한 후, 해당 데이터와 가장 가까운 클러스터에 해당 데이터를 배정

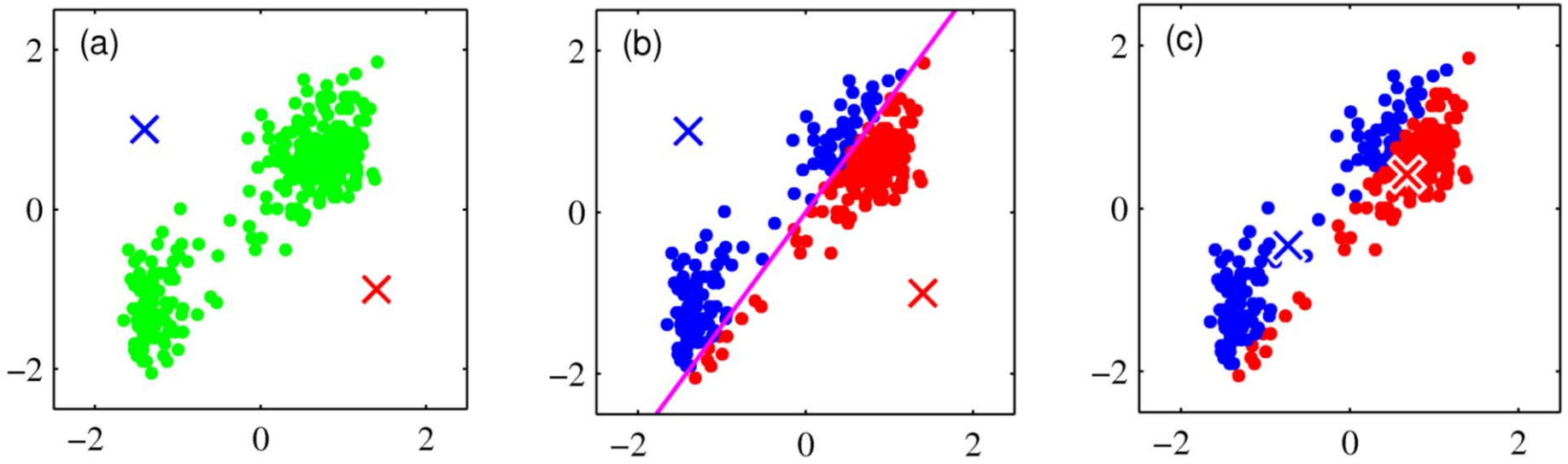


K-means Clustering Algorithm

K개의 클러스터를 이용하여, 데이터들의 평균점을 클러스터의 중심점으로 이용

Step 2: Maximization (or Minimization)

각 클러스터들의 중심점을 해당 클러스터들에 배정된 데이터들의 평균 값 지점으로 이동

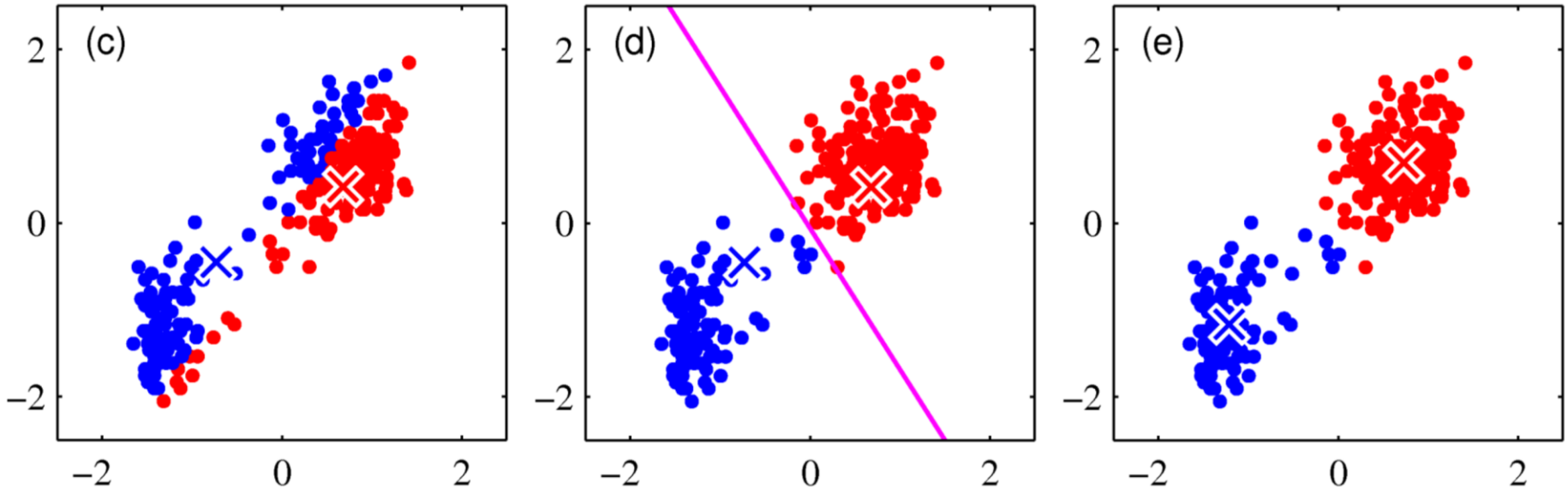


K-means Clustering Algorithm

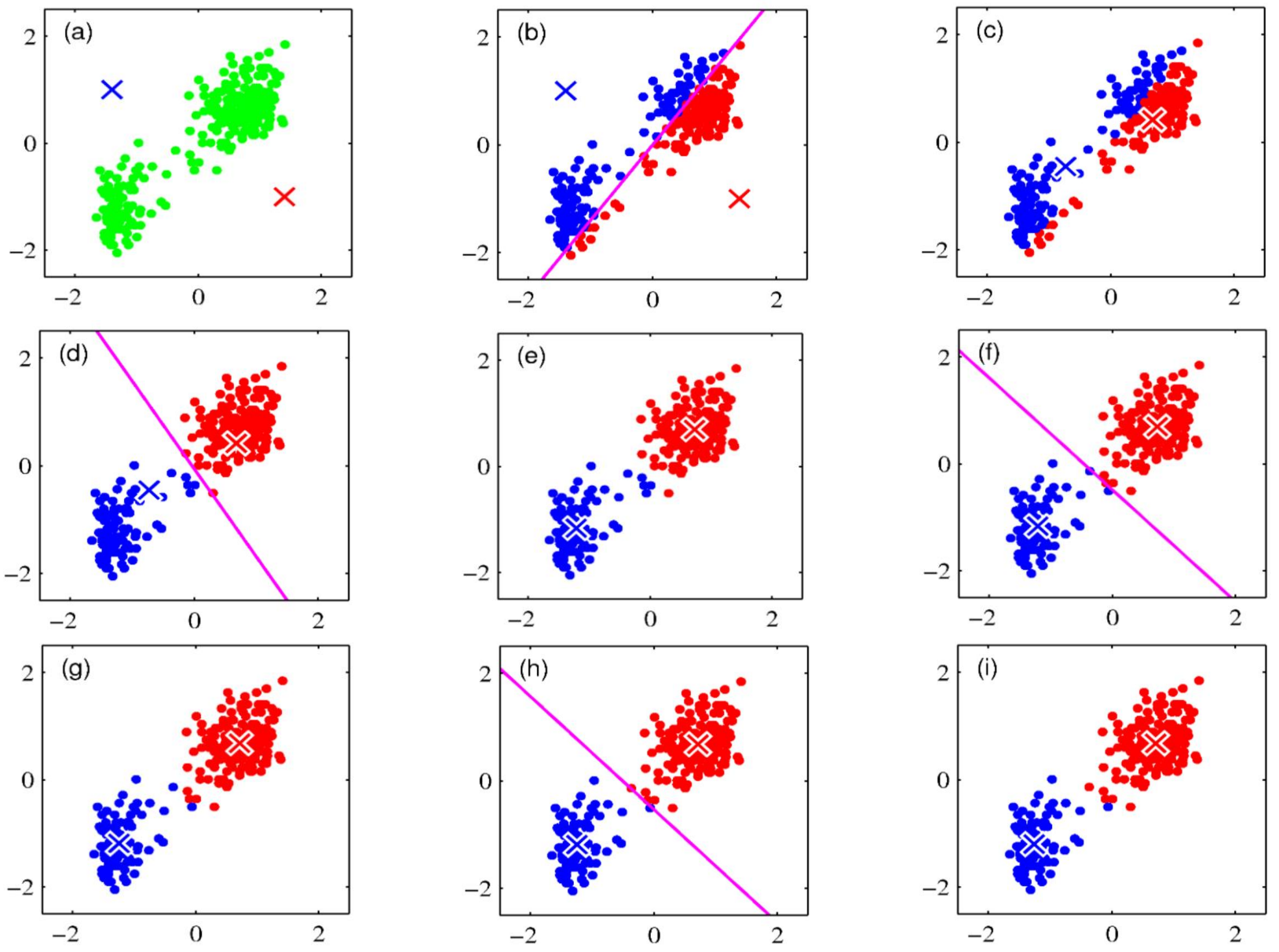
K개의 클러스터를 이용하여, 데이터들의 평균점을 클러스터의 중심점으로 이용

Step 3:

더이상 클러스터의 중심 값이 바뀌지 않을 때까지 Step 1~2를 반복
(각 클러스터에 배정된 데이터들의 변동이 없을 때까지 반복)



K-means Clustering Algorithm
클러스터링 과정



DataSet

데이터셋 설명 및 출력

UCI에서 제공한 ML 학습용 Iris(붓꽃) 종 분류 데이터 150건, 클러스터의 수는 3개

세 개의 Iris 품종을 구분하기 위한 데이터 셋

(품종 1: Iris-setosa, 2: Iris-versicolor, 3: Iris-virginica)

꽃받침 길이(cm)	꽃받침 너비(cm)	꽃잎 길이(cm)	꽃잎 너비(cm)
5.1	3.5	1.4	0.2
7.0	3.2	4.7	1.4
6.3	3.3	6.0	2.5
…	…	…	…

세 클러스터의 **중심점을 출력**하시오.

ex) X.X X.X X.X X.X (출력 순서는 상관 없음. 소숫점 첫째 자리까지만 출력 요망)
Y.Y Y.Y Y.Y Y.Y
Z.Z Z.Z Z.Z Z.Z

Thank you