

TONGMYONG UNIVERSITY
동명대학교

파이썬 기초프로그래밍

SW융합대학 정보보호학과
박종규





TONGMYONG UNIVERSITY
동명대학교

목 차

1 KNN

2 SVM

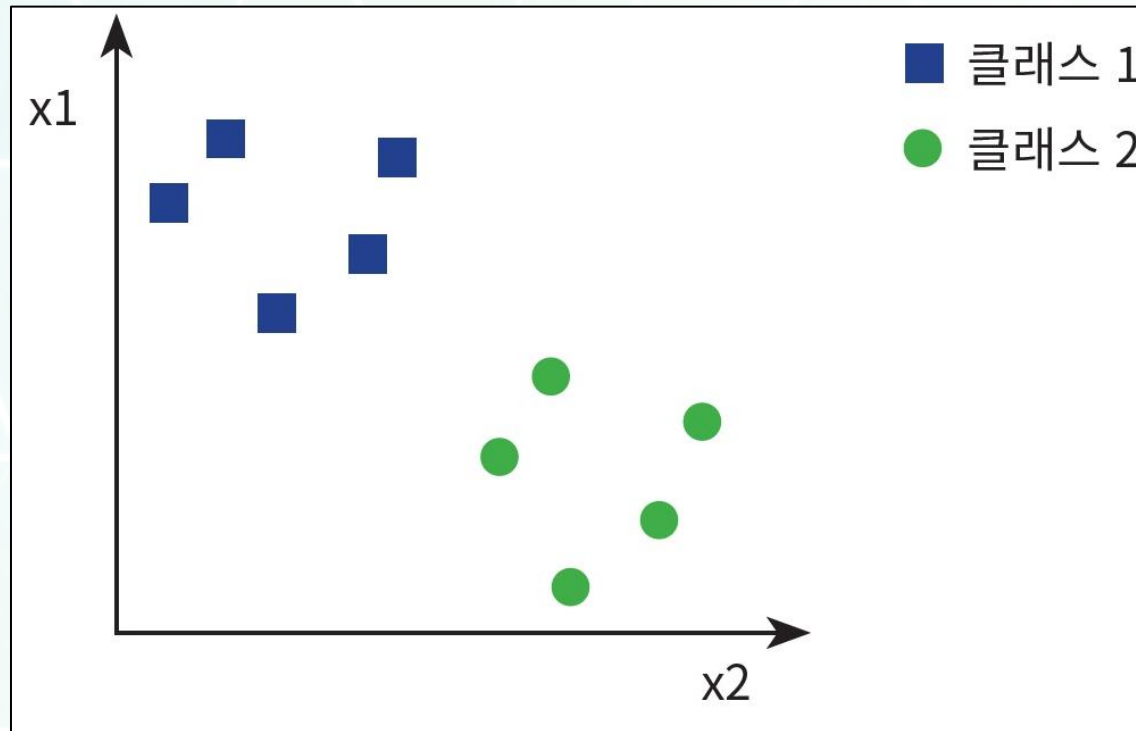
3 K-Means



ARTIFICIAL INTELLIGENCE

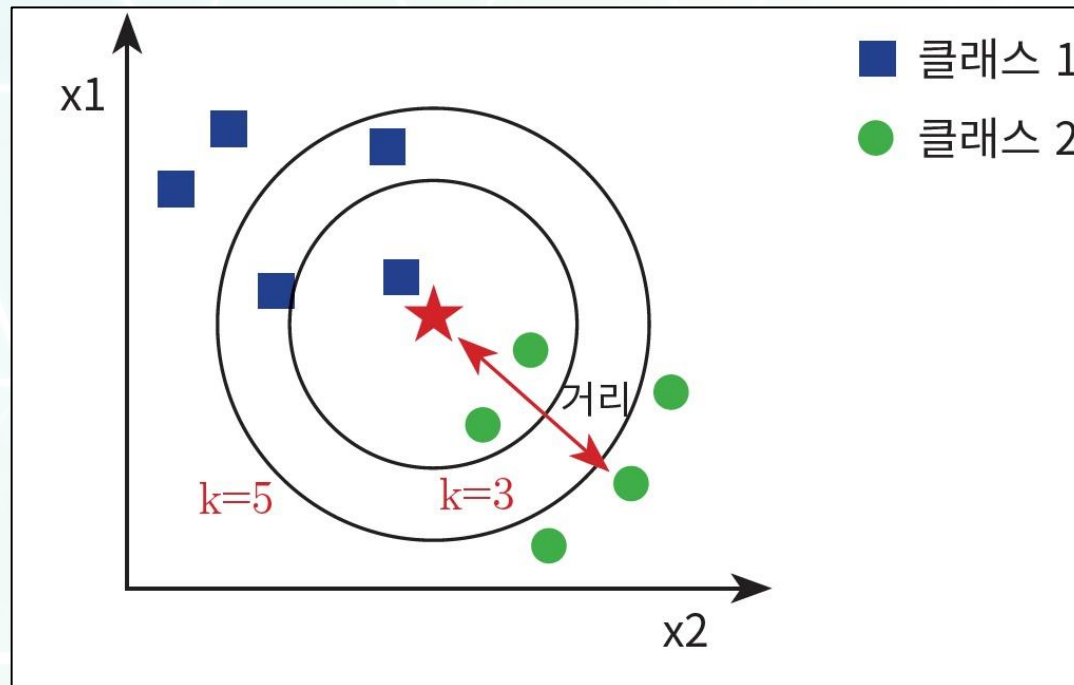
KNN

- K-Nearest Neighbor(KNN) 알고리즘은 기계학습 알고리즘 중에서 가장 간단한 분류 알고리즘 중 하나



KNN

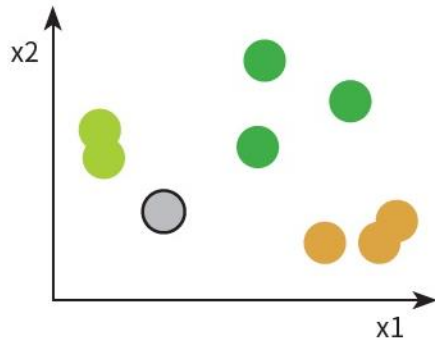
- 새로운 데이터가 입력되어서 그래프 상에 별표로 표시됨
- 별표는 파랑색 사각형 또는 빨강색 원 중에 속해야함
- 이것을 분류라고 함



KNN 알고리즘

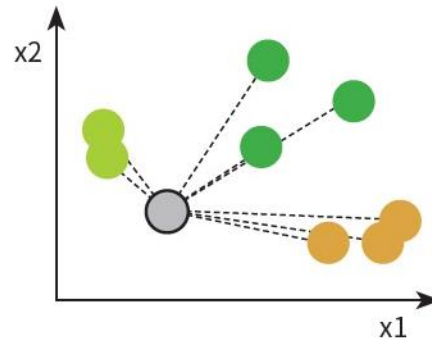
- KNN 알고리즘

1. 데이터를 관찰한다.



회색 원은 어디에 속해야 할까?

2. 거리를 계산한다.



회색 원과 다른 원들간의 거리를 계산한다.




3. 이웃을 찾는다.




점	거리	
 	2.1	→ 1등
 	2.4	→ 2등
 	3.1	→ 3등
 	4.5	→ 4등

거리에 따라서 이웃 원들을 정렬한다.

4. 새로운 데이터에 대하여 투표한다.

클래스 투표수

	2	
	1	
	1	


 색 원이 가장 많았으므로 새로운 원은  에 속한다.

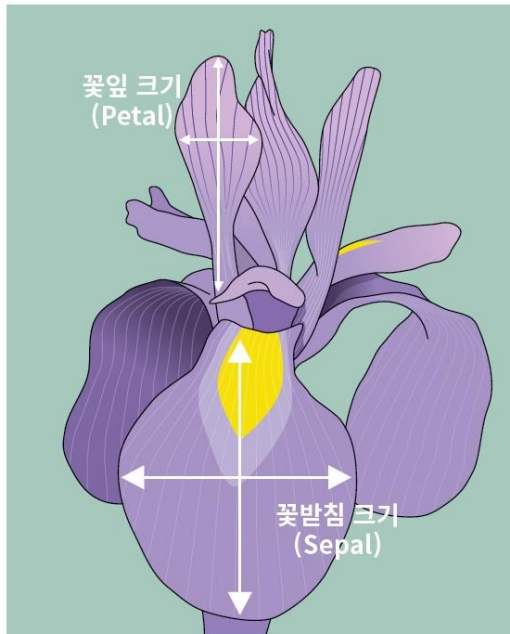
가장 가까운 k개의 이웃 중에서 가장 많은 표를 얻은 클래스로 분류한다.

KNN 알고리즘 장단점

- 특징 공간에 있는 모든 데이터에 대한 정보가 필요함
- 가까운 이웃을 찾기 위해 새로운 데이터에서 모든 기존 데이터까지의 거리를 확인해야함
- 데이터와 클래스가 많다면 많은 메모리 공간과 계산 시간이 필요함

실습 - KNN 알고리즘

- sklearn 라이브러리의 Iris Data Set을 활용한 분류
- 주피터 KNN_sklearn과 KNN_scratch 비교



```
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
```

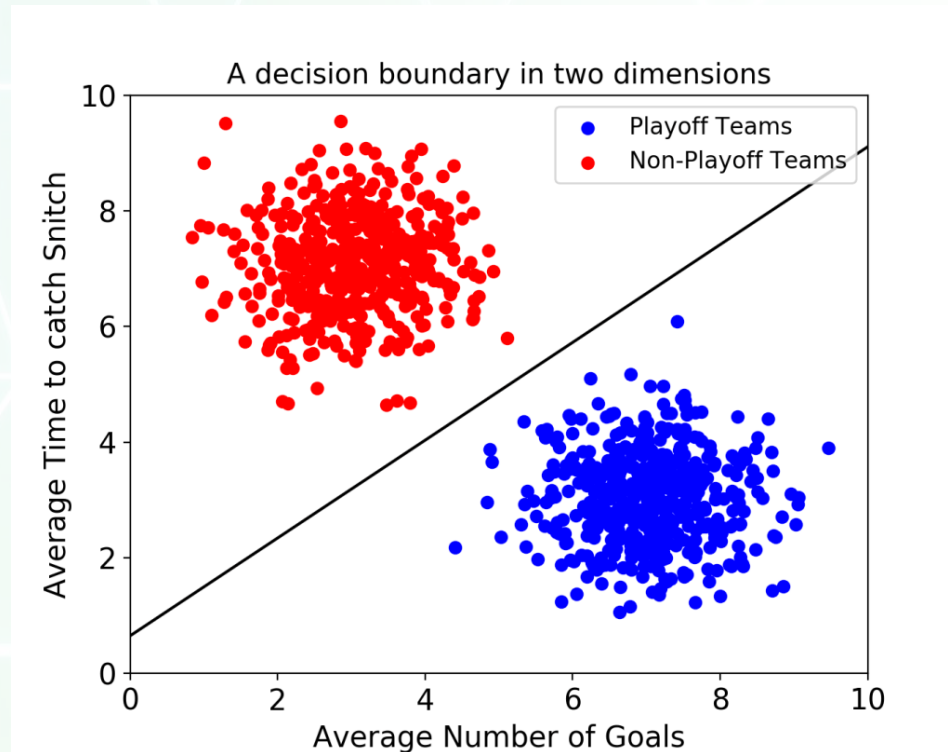
(아이리스 꽃: 꽃받침 길이와 너비, 꽃잎 길이와 너비)

SVM

- Support Vector Machine(SVM)은 주로 분류를 위해 사용함
- 두 카테고리 중 어느 하나에 속한 데이터 집합이 있을 때, SVM 알고리즘은 주어진 데이터 집합을 바탕으로 새로운 데이터가 어디에 속할지 판단하는 비확률적 이진 선형분류 모델을 만듦
- SVM은 선형 분류와 더불어 비선형 분류에서도 사용됨
- 비선형 분류는 고차원 특징 공간으로 사상하는 작업이 필요함

SVM

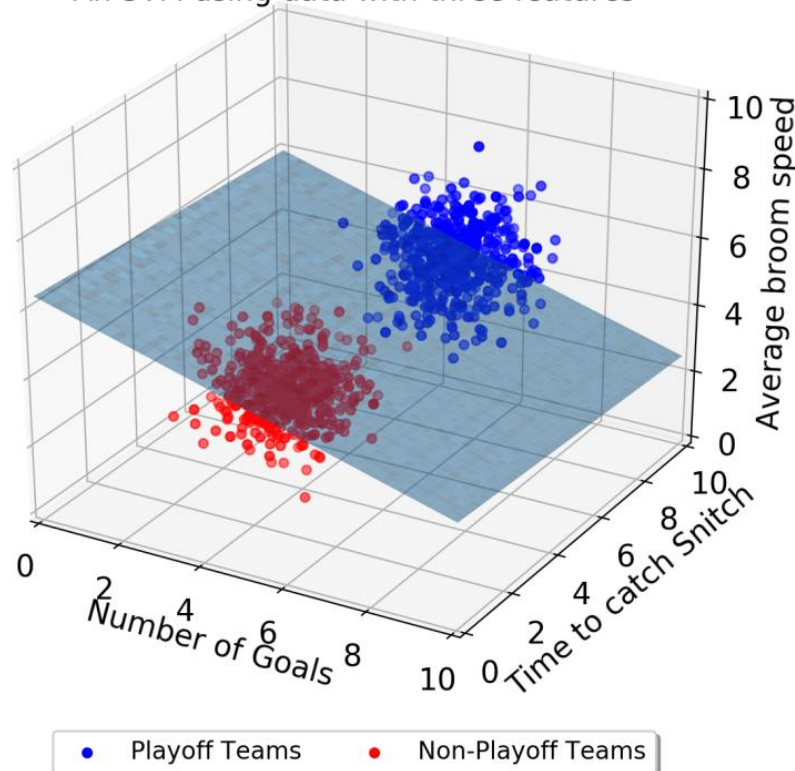
- SVM은 카테고리가 분류된 데이터 군집이 있을 때, 이를 나누는 초평면(hyper-plane)들의 집합으로 이루어져 있음
- 초평면과 데이터의 거리가 멀 수록 분류오차가 작기 때문에, 데이터와 거리가 먼 초평면을 찾는 것이 SVM 알고리즘의 목적



SVM 그래프

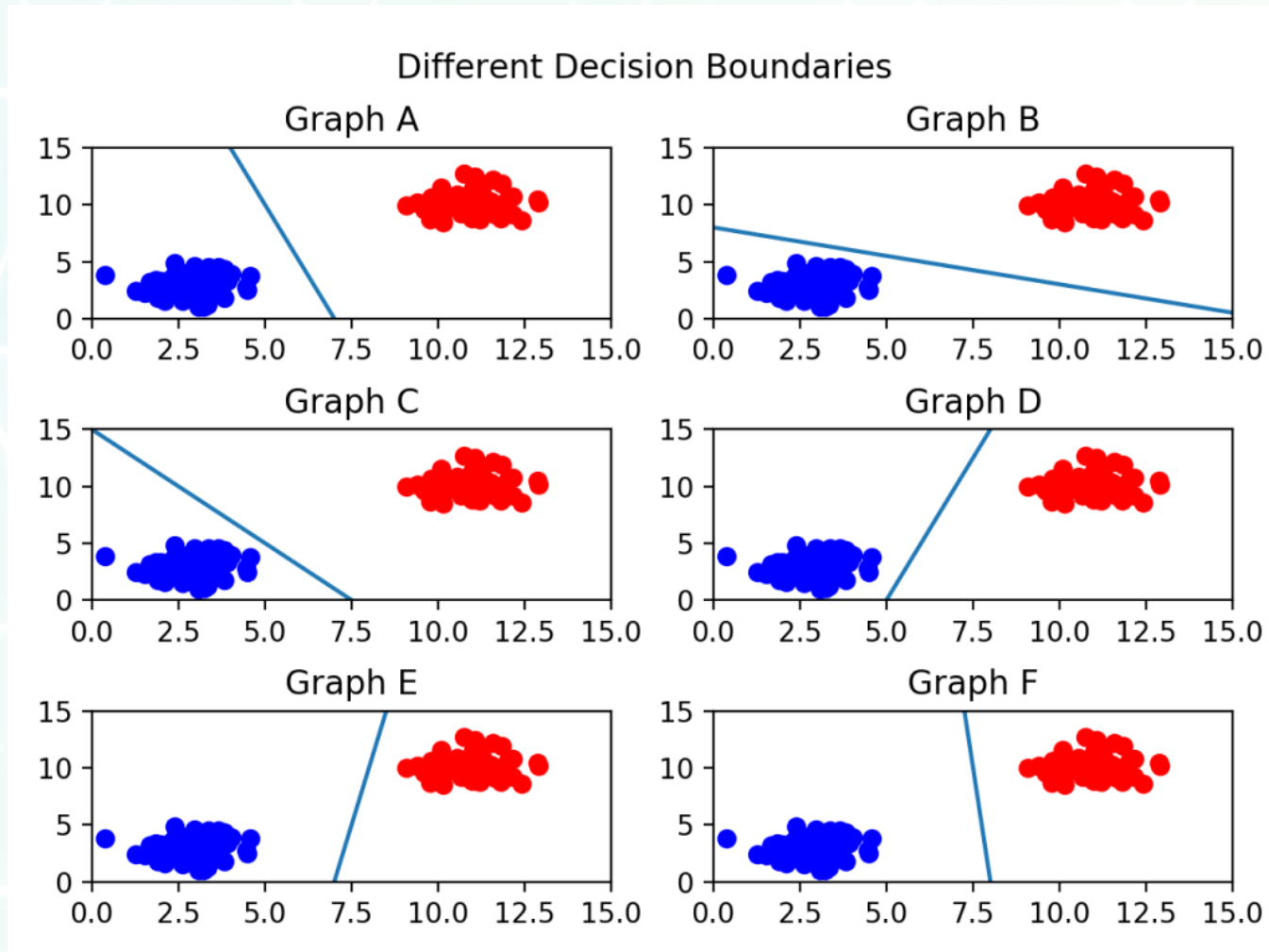
- 데이터의 속성이 3개면 그래프도 3차원으로 표현

An SVM using data with three features



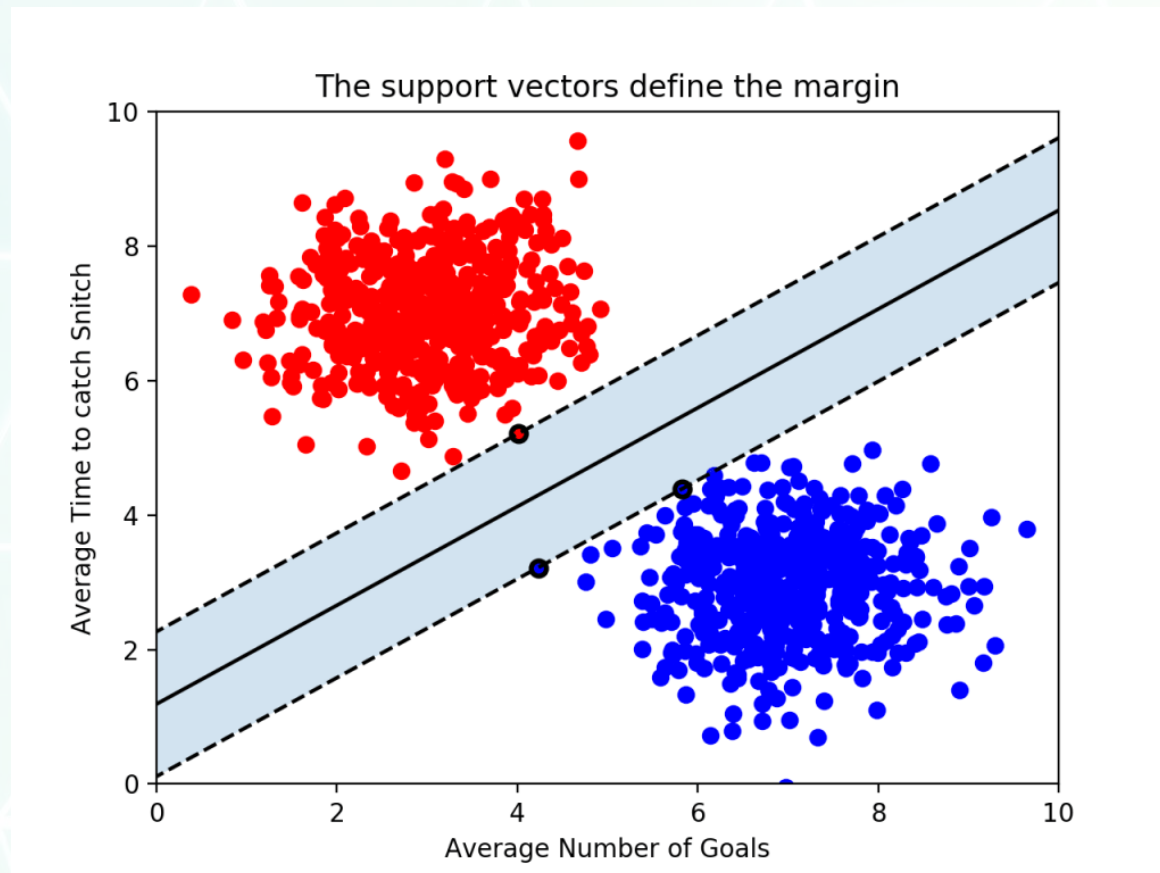
최적의 경계

- 데이터를 분류하는 경계선은 아래와 같이 나올 수 있음



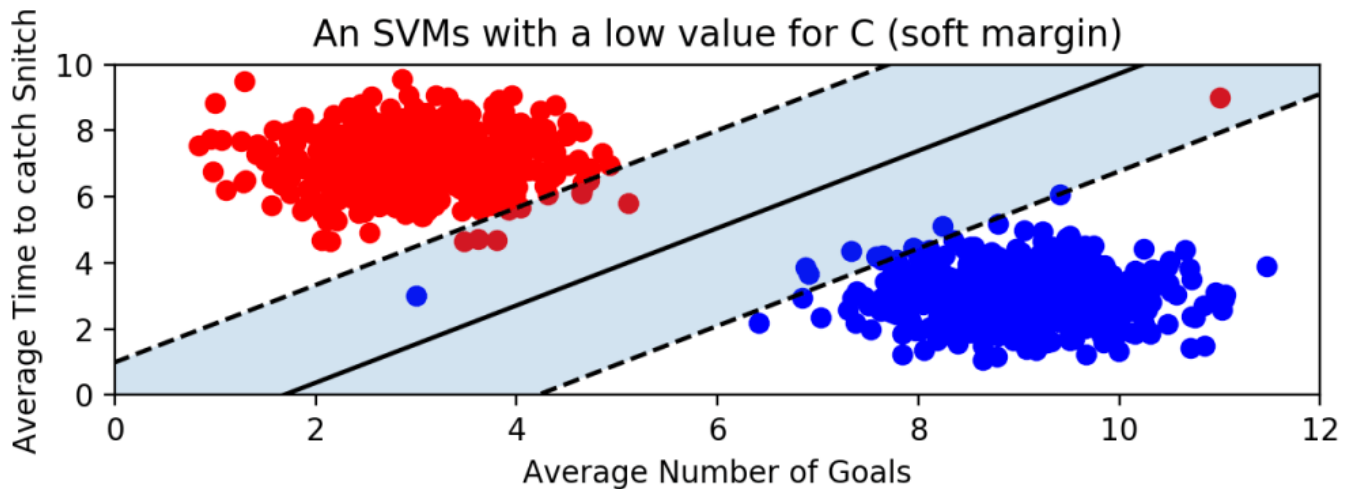
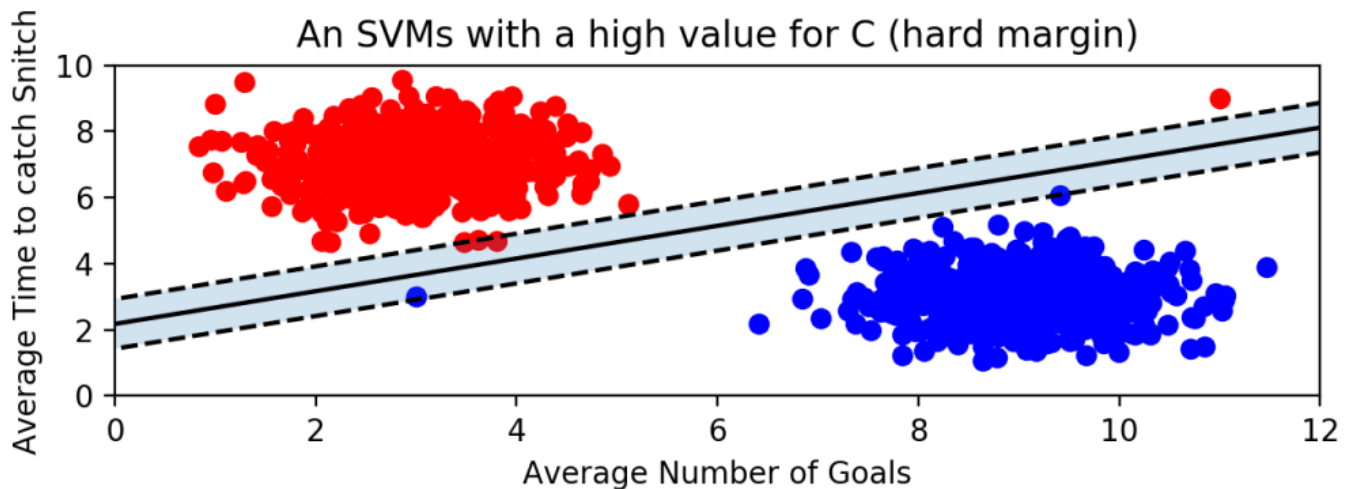
SVM 마진

- 마진(Margin)은 경계와 서포트 벡터 사이의 거리
- 마진이 클 수록 좋은 분류



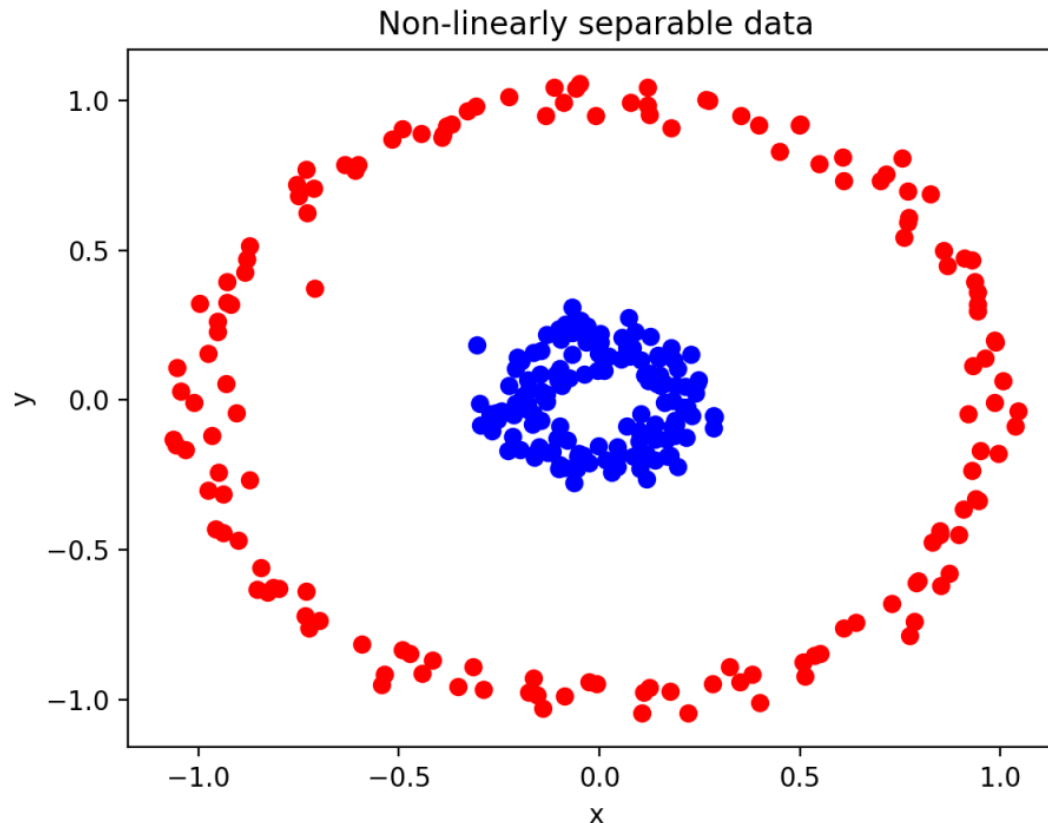
아웃라이어

- 아웃라이어(outlier)는 데이터 중 지나치게 높거나 낮은 값



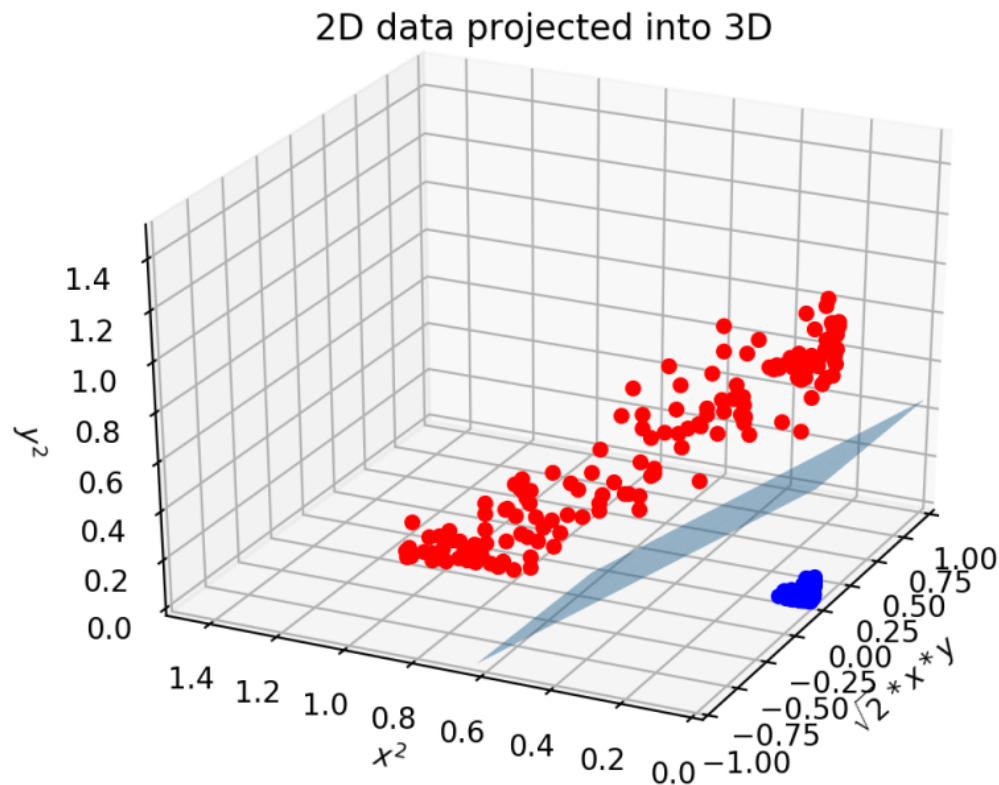
커널

- 초평면을 선형으로 정의할 수 없을 때에는 커널을 선형 (linear) 대신 다항식(poly) 같은 커널을 사용



커널

- 다항식 커널을 사용해 데이터를 다시 그리면 아래와 같음



실습 - SVM

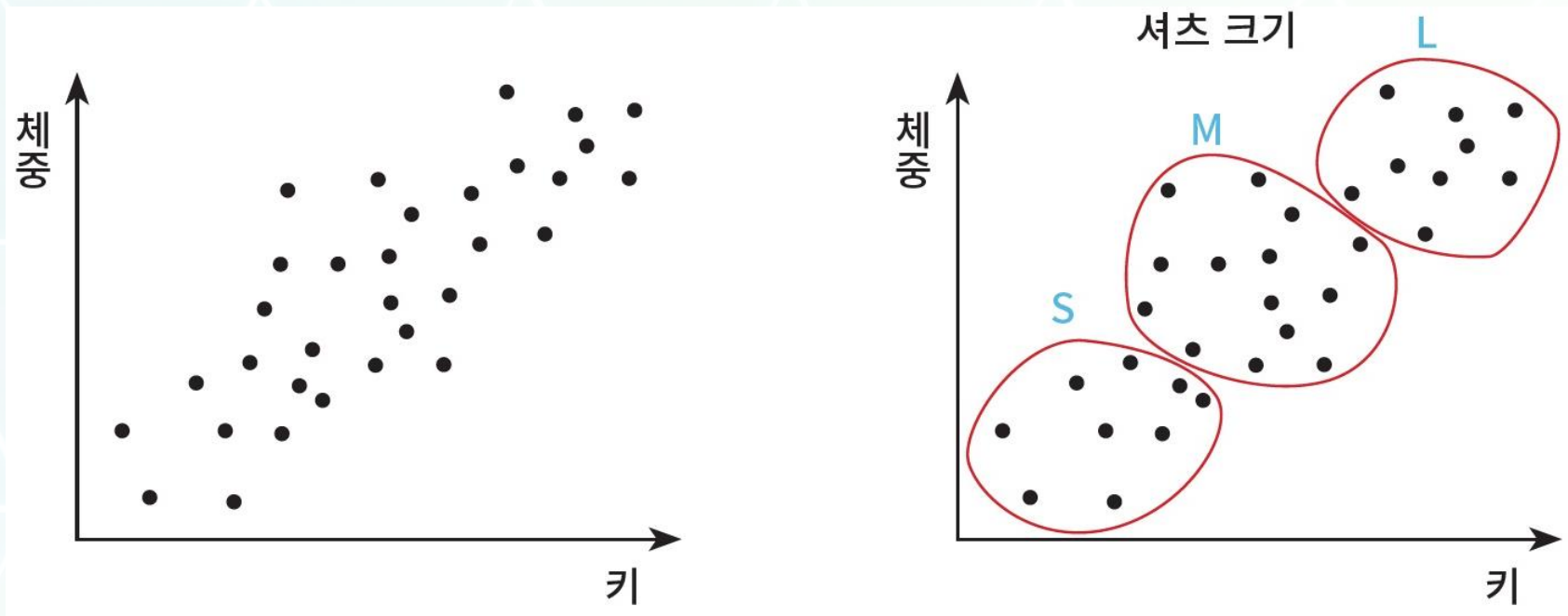
- 주피터 SVM 실습

K-means

- K-means 알고리즘은 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로 비지도 학습에 속함
- 각 클러스터와 거리 차이의 분산을 최소화 하는 방식으로 작동함
- 레이블이 달려있지 않은 데이터에 레이블을 달아주는 역할을 수행함

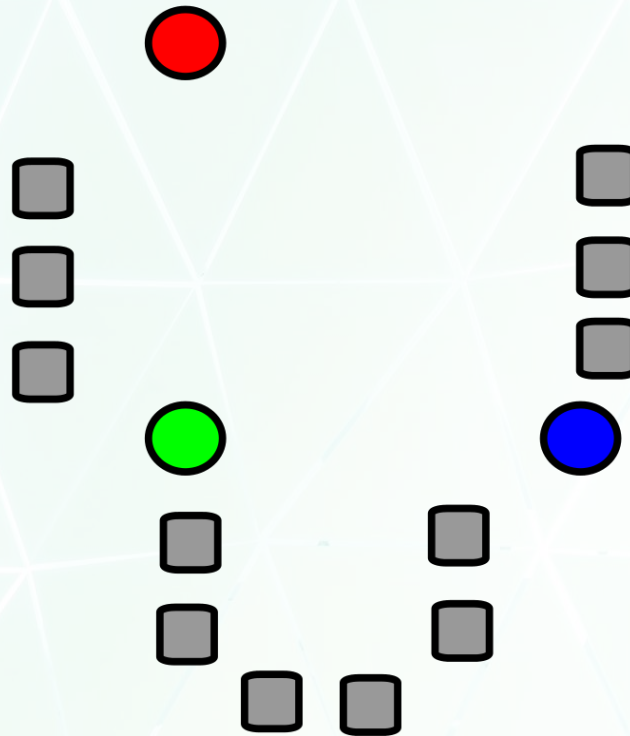
K-means 예시

- 키와 체중에 따라 셔츠 사이즈를 분류하는 모델



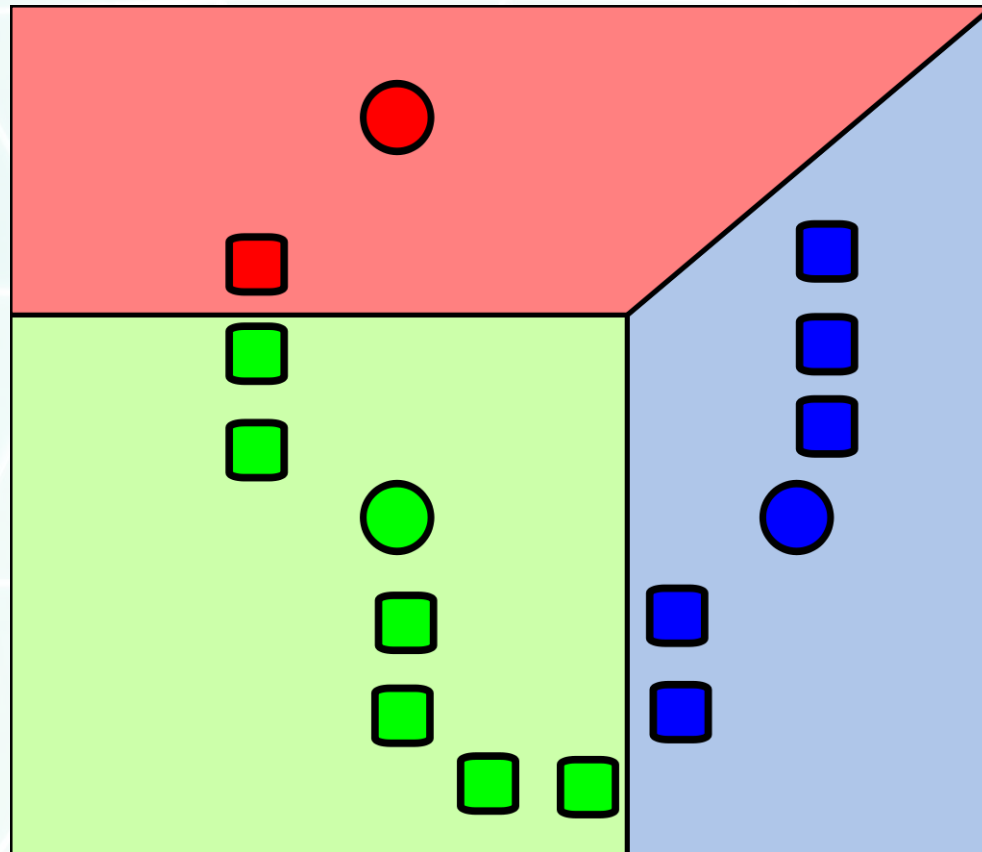
K-means 알고리즘 과정

- 초기 k 평균값은 데이터 오브젝트 중에서 무작위로 선정



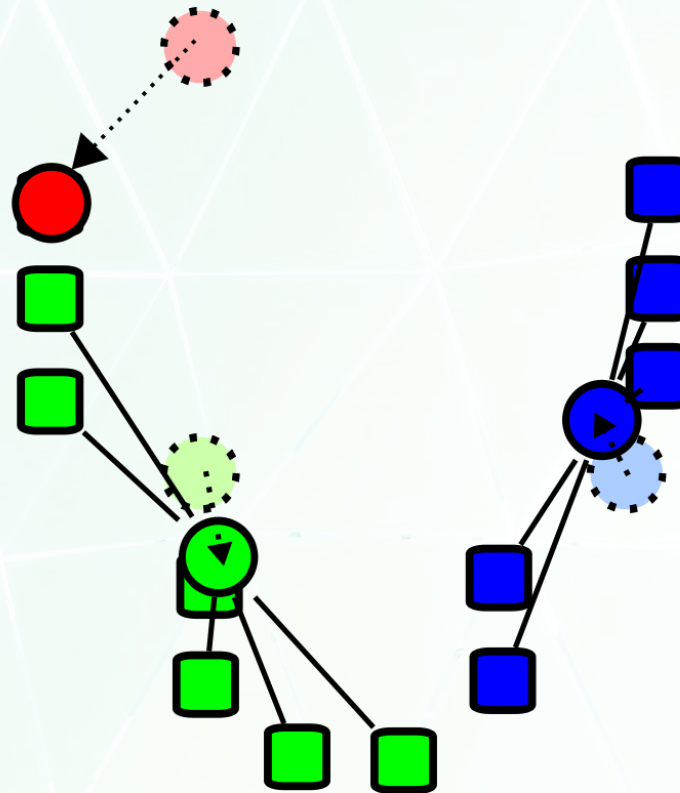
K-means 알고리즘 과정

- k 각 데이터 오브젝트들은 가장 가까이 있는 평균값을 기준으로 묶임



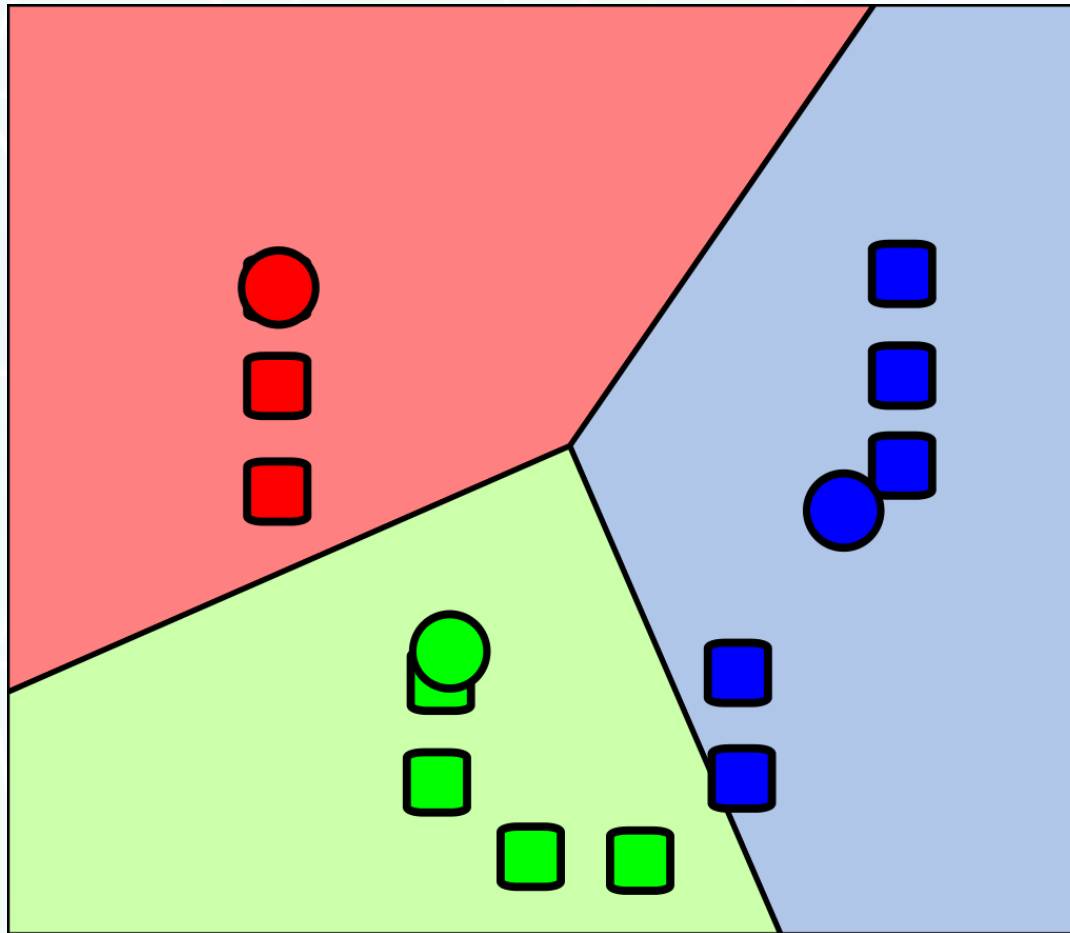
K-means 알고리즘 과정

- k 개의 클러스터의 중심점을 기준으로 평균값을 재조정함



K-means 알고리즘 과정

- 각 클러스터의 집합이 변하지 않을 때까지 반복



K-means 알고리즘 과정

- K-means 알고리즘 슈도 코드

입력값

k: 클러스터 수

D: n 개의 데이터 오브젝트를 포함하는 집합

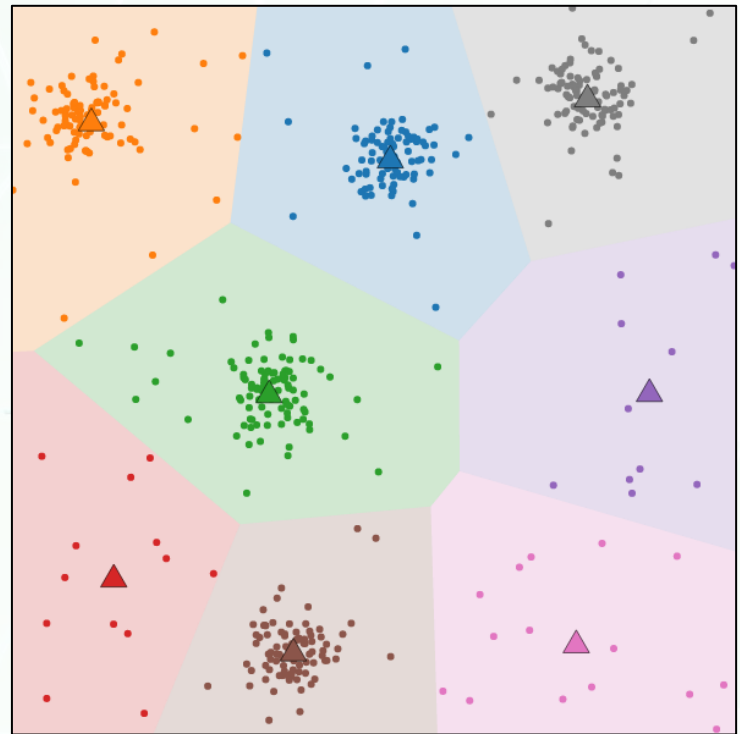
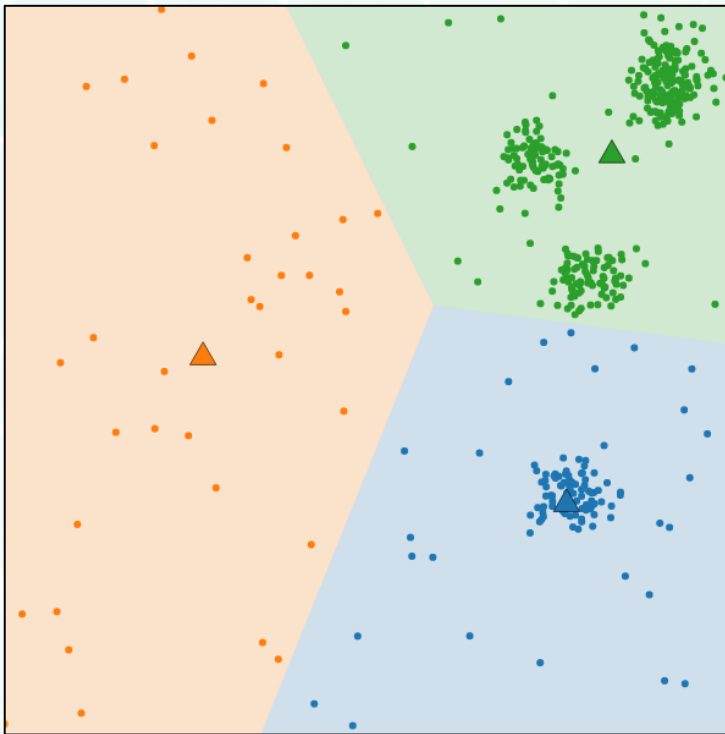
출력값 : k 개의 클러스터

알고리즘

1. 데이터 오브젝트 집합 D에서 k 개의 데이터 오브젝트를 임의로 추출하고, 이 데이터 오브젝트들을 각 클러스터의 중심 (centroid) 으로 설정한다. (초기값 설정)
2. 집합 D의 각 데이터 오브젝트들에 대해 k 개의 클러스터 중심 오브젝트와의 거리를 각각 구하고, 각 데이터 오브젝트가 어느 중심점 (centroid) 와 가장 유사도가 높은지 알아낸다. 그리고 그렇게 찾아낸 중심점으로 각 데이터 오브젝트들을 할당한다.
3. 클러스터의 중심점을 다시 계산한다. 즉, 2에서 재할당된 클러스터들을 기준으로 중심점을 다시 계산한다.
4. 각 데이터 오브젝트의 소속 클러스터가 바뀌지 않을 때까지 2, 3 과정을 반복한다.

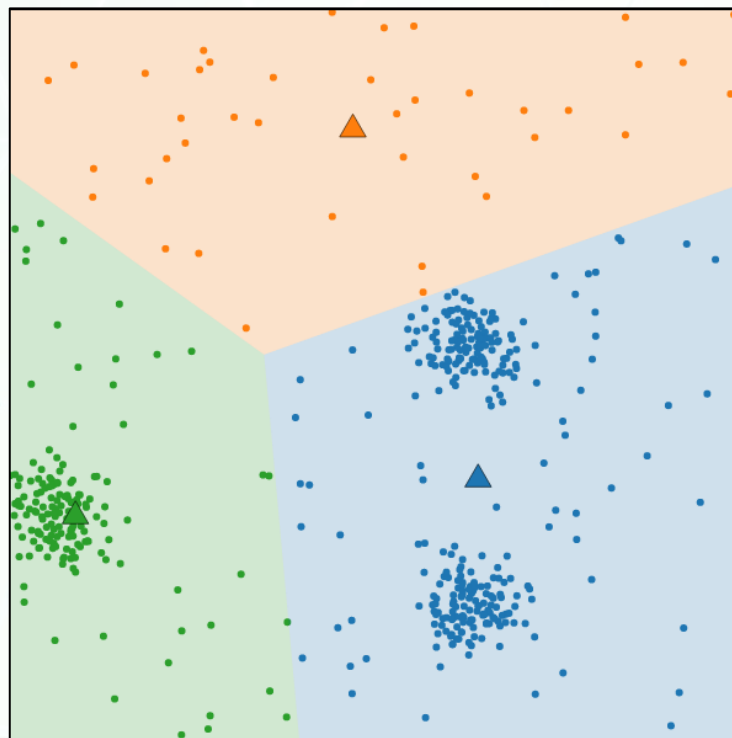
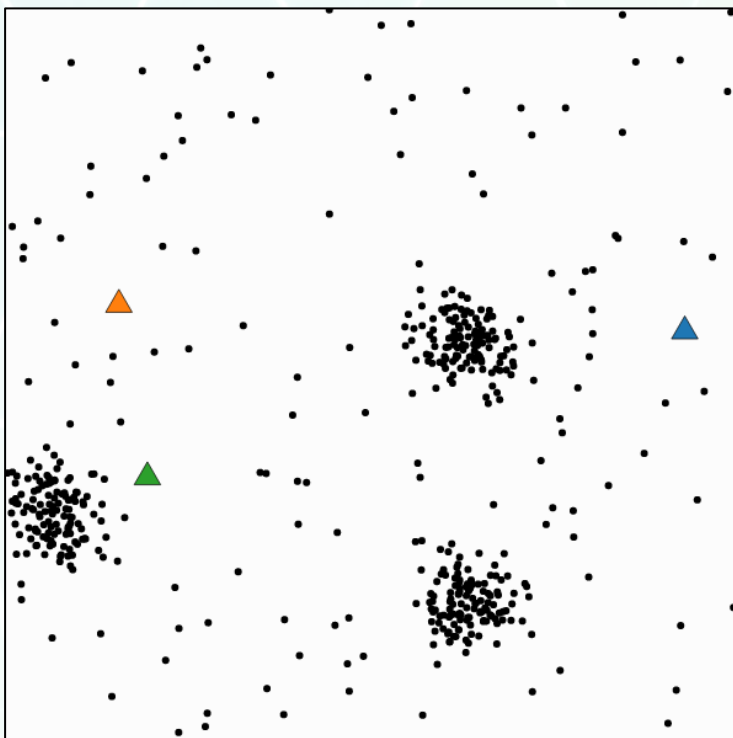
K-means 알고리즘의 한계

- 실제로 분류되어야 할 k 값을 입력 파라미터로 지정해주어야 함
- 실제 클러스터가 4개인데 $k=3$ 을 입력(좌)와 실제 클러스터는 5개인데 $k=8$ 을 입력(우)한 경우



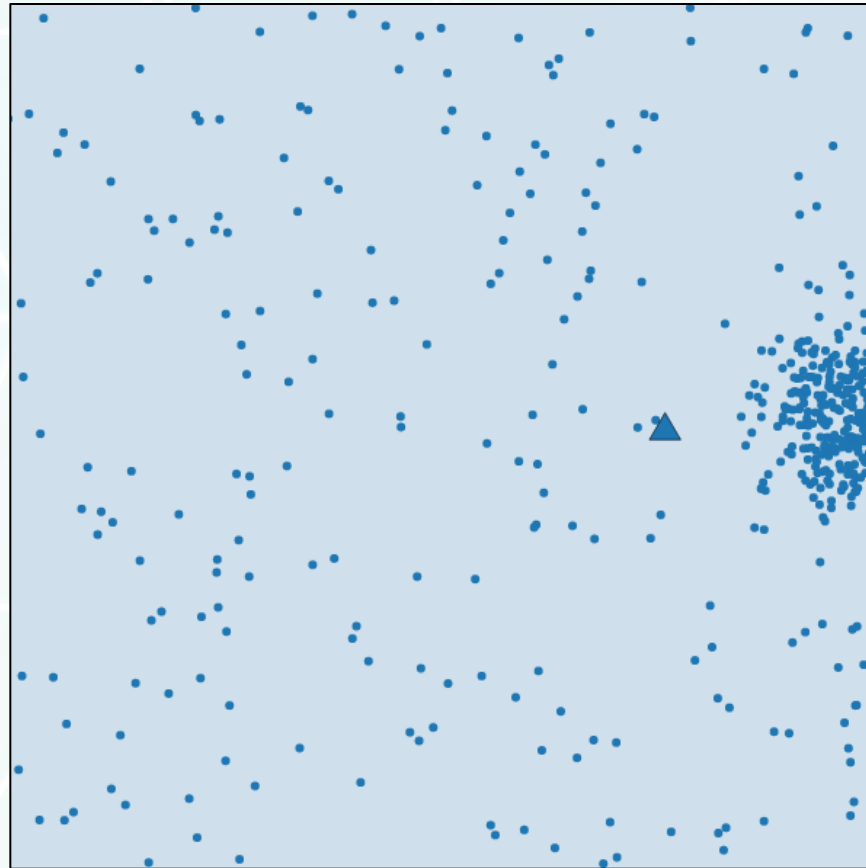
K-means 알고리즘의 한계

- 알고리즘의 에러 수렴이 전역 최솟값이 아닌 지역 최솟값으로 수렴할 가능성
- 클러스터와 k값 둘 다 3이지만, 클러스터의 평균을 내는 과정에서 지역 최솟값에 빠져 그대로 수렴해버린 경우



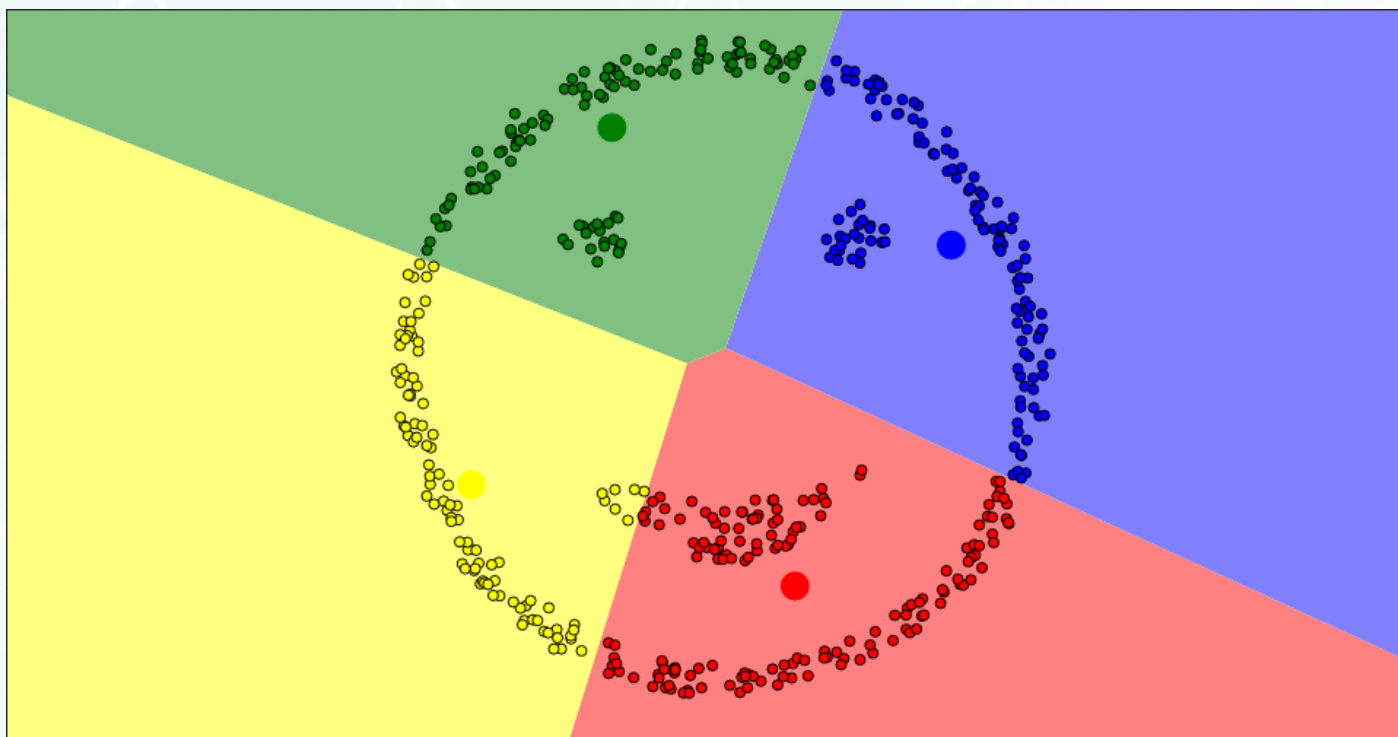
K-means 알고리즘의 한계

- 아웃라이어에 민감함



K-means 알고리즘의 한계

- 데이터 분포가 구형이 아닌 경우 클러스터가 잘 만들어지지 않음



K 값을 찾는 방법

- Rule of Thumb
 - 가장 간단한 방법으로 데이터의 수가 n 이라고 할 때, 필요한 클러스터의 수는 $k \approx \sqrt{n/2}$
- Elbow Method
 - 클러스터의 수를 1부터 순차적으로 늘려가며 결과를 모니터링 함
 - 평균에서의 거리 오차를 더한 값이 줄어들기 시작하는 k 값을 선택
- 정보 기준 접근법
 - 클러스터링 모델에 대해 가능도를 계산하는 것이 가능할 때 사용하는 방법으로 베이지안 정보 기준 등이 있음

실습 - K-means

- 주피터 K-means 실습
- 주피터 ElbowMethod 실습



TONGMYONG UNIVERSITY

• 동명대학교

Thank You!

ABSTRACT BACKGROUND