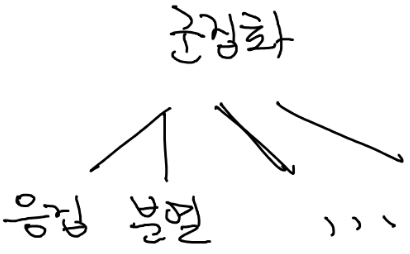
중간고사 요점정리.

군집화의 일원으로 여러특성을 가진 샘플들을 하나의 군집으로 묶는 비지도학습의 방법
⇒ Clustering

군집화를 할 때 샘플들을 군집하는 기준.



거리와 유사도

1. 민코프스키 거리 $\left(\sum (x_1 - x_2)^p\right)^{\frac{1}{p}}$

   (p=2) 유클 $\sqrt{(x_1 - x_2)^2}$

   (p=1) 맨하튼 $|x_1 - x_2|$

k-means

Clustering의 방법에서 응집, 분열 등 중의 분열알고리즘에 속하는 방식

군집의 개수가 주어지며 Dmax, Dmin, Davg 를 이용하여 군집의 중심을 수정하며 마지막이 이전과 군집의 중심이 같게되면 끝내는 알고리즘

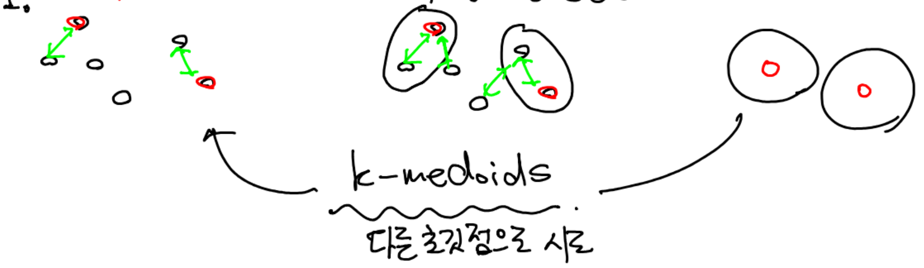
초깃값과 outliver 에 민감하다 (거리측정에 대한 오류가 발생).

해결책 : 표준편차로 정확도를 판단함에 따라 초깃값을 달리하며 반복 (k-medoids)

장점 : 타 알고리즘에 비해 빠르다.

알고리즘순서

1. 아무 초깃점과 개수 2. 거리값측정 군집갱신 3. 반복 끝내면 중심고정



단점 : 초깃값과 outliner 에 약하다.

SVM : 샘플을 어떻게 분류시킬 것인가에 대한 방법론

분류하여 모고과하는 답

Category Data : 특성이 있는 범주형 데이터

Data Embedding : 위의 데이터를 벡터형으로 변환

Decision Function 인 Hyperplane 을 통해 나눈다.

$W^T X + b > 1$ Positive

$\;\;\;\;\;\;\;\;'' \;\;\; < -1$ Negative

장점 : 분류와 예측이 둘다 사용이 가능하다.

좋은 Hyperplane 은? 일거인듯은 Data가 잘맞는 Hyperplane

↳ Margin을 최대화 할시 좋은 Hyperplane.
(여백)

단점 : kernel 및 파라미터 조정을 여러 번해야한다.

SVM은 위 Margin의 끝에 걸려있는 샘플의 벡터이다.

kernel Gamma : 분류가 어려운 데이터를 Gamma 의 파라미터 조정을 통해 일부 샘플의 위치를 조정하는 기술

C와 Gamma의 Parameter를 찾는 기술 : Grid Search

) Rbf kernel

$W^T W + C$: Cost ↑ margin ↓ / Cost ↓ margin ↑

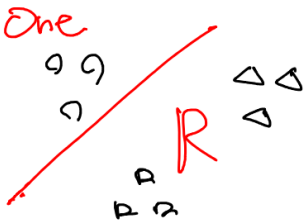
위의 과정을 거친 SVM을 Soft SVM이라 일컫음.

장점 : 실생활의 적용이 가능. 예측, 분류모델에 따라 사용이 가능하다.

단점 : Gamma. C 파라미터를 어떠한 Test를 통해 알아보아야한다.

Multi의 SVM

One vs Rest



클래스 한쪽의 이진분류기 효율적임.

One vs One

$\frac{K(K-1)}{2}$ 개 만큼의 클래스들끼리

kernel 알고리즘이 사용.

OvR 보다 빠르게 연산.