0415_규민

-Clustering(군집) vs Classification(분류)

- `clustering`
  - 비지도학습
  - 데이터간의 유사도를 정의하고 그 유사도에 가까운 것부터 순서대로 합쳐가는것
- `classification`
  - 지도학습
  - 기존에 존재하는 데이터의 category 관계 파악하고, 새롭게 관측된 데이터 category를 스스로 판별

-군집분석

> 개체를 분류하기 위한 명확한 분류기준이 존재하지 않거나 기준이 밝혀지지 않은 상태에서 주어진 데이터들의 특성을 고려하여 같은 그룹(클러스터)를 정의하고, 다른 클러스터의 개체보다 서로 더 유사한 개체가 되도록 그룹화하여 그룹의 대표성을 찾아내는 방법.

중심 기반(Center-based clustering)

> 동일한 군집에 속하는 데이터는 `어떠한 중심을 기분으로` 분포할 것이다.

밀도 기반(Density-based clustering)

> 동일한 군집에 속하는 데이터는 `서로 근접하게` 분포 할 것이다.

종류

- **K-Means**
  - 장점
    - 아주빠름
    - 계산량 O(n)
  - 단점
    - 그룹수 (k) 직접정해야함
    - 처음에 point를 랜덤하게 고르기 때문에 돌릴떄마다 다른결과나옴
- **Mean-Shift Clustering**

  > data points의 가장 밀도 높은 지역을 찾으려함

  - 장점

        - K-means와 대조적으로 클래스 혹은 그룹의 개수 정할 필요X
    - 단점
        - kernel = 반지름r 사이즈 선택해야함
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**

    > 임의로 고른(가본적없는)start point에서 시작
    >
    > 모든 방문점 방문할때까지 계속됨

    - 장점
        - 클래스나 그룹수 지정필요 없다
        - outlier를 noise로 정하고 클러스터 찾는다
    - 단점
        - 잘 작동되지 않는다
        -
- **GMM (Expectation-Maximization (EM) Clustering using Gaussian Mixture Models)**
    - 장점
        - k-meas보다 flexible
        - datapoint가 겹쳐진 두개의 클러스터 사이에있는것도 확률로 나눠서 판단
- **Agglomerative Hierarchical Clustering**
    - 장점:
        - 클래스나 그룹개수 정하지않는다
        - 거리에 민감하지않다
        - 데이터가 계층적 구조를 갖고있을때 매우좋다
    - 단점:
        - 높은계산량 O(n^3)

활용데이터

> 기술통계조사꺼!

- 기술개발 애로요인`H1` -> 원하는게 다를듯
- 기술개발 지원제도필요 시기 `I3태그`-
- 수출여부
- 대기업납품여부
- 향후 1년간 중점 투자계획분야
- 단계별 기술개발 자금지원 필요성
- 개발기술 사업화를 위해 가장 필요한 지원책

의견

`각자 분석기법맡아서 한두개씩 해보고 좋은거 찾는것도 좋을듯`

군집수를 정해놔야 mbti가 적어진다

하지만 군집수 정해놓는건 안좋을수도