Data Mining, Project (Part 1)

1. 데이터

- data.csv: 샘플 801개 × 유전자 20,531개의 발현 값 (행 = 샘플, 열 = 유전자)

2. 규정

- 생성형 AI 활용 허용 범위: 코드 작성 및 디버깅, 개념 설명 요청, 문장 다듬기
- 필수 조건
  - ✓ scikit-learn의 클러스터링 라이브러리 호출 금지
  - ✓ 알고리즘에 대한 이해 필요
  - ✓ 제출한 코드의 모든 라인에 대한 이해 필요

3. 프로젝트 (Part 1) 내용

K-means 클러스터링 알고리즘을 구현하고, 그 결과 분석 및 최종 K값 선택의 근거 제시

- **데이터 전처리**: 20,531차원을 그대로 쓸지, 줄일지 스스로 결정해 보기
  - o 최소 2가지 다른 전처리 설정(+전처리 하지 않은 설정)으로 실험하여, 동일한 K값 선택에 따른 클러스터링 결과가 유지되는지 비교하고, 선택한 전처리의 타당성 분석
- **K-means 구현** (강의 자료의 4단계 절차 참고)
  - o 초기 중심 선택(무작위로 k개 샘플 선택), 할당 단계와 갱신 단계, 수렴 조건(클러스터 구성이 더 이상 바뀌지 않는 조건)에 대한 처리 구현
- **결과 분석 및 K값 선택** (K = 2 ~ 10)
  - o WSS(응집도), BSS(분리도), 실루엣 점수를 직접 구현하고 k값 선택에 따른 결과 표(또는 그래프)를 함께 제시하여 K값 선정 기준을 설명
  - o 최종 K값에서 클러스터링 결과를 2차원 산점도(PCA)로 시각화

4. 제출

- 코드와 보고서를 모두 포함한 하나의 zip파일(학번_이름.zip)을 LearnUs를 통해 제출

제출물	내용	형식
코드	명령 한번으로(예, python problem1.py) 보고서의 모든 표와 그림이 재현될 수 있도록 전처리, K-means, 평가 지표, 실험 스크립트 코드를 구현할 것	.py + README.md(포함 가능)
보고서	과제 내용의 단계 순서대로 작성, 분량은 3p이내	PDF

- 제출 마감일: 11월 23일(월)

5. 평가 기준

항목	배점	만점 기준
구현 정확성	30	K-means가 4단계 절차대로 동작하는지 확인. WSS·BSS·실루엣 Score 계산 및 시각화 구현
전처리 근거	20	정규화·특징 선택·PCA 차원 수 등을 대안과 비교한 실험으로 뒷받침
결과 분석 및 K값 선택 근거	20	K값 도출의 근거가 논리적인지, 전처리를 바꿔도 결론이 유지되는지 확인
합계	70	

- **감점**
  - o 명령 한 번으로 보고서의 수치·그림이 재현되지 않으면 해당 결과는 인정하지 않음 (또한 해당 명령이 무엇인지 README.md 또는 보고서에 포함할 것)
  - o 제출 양식(.zip파일 또는 파이썬 코드, 보고서 pdf) 중 부족한 부분이 있으면 감점
  - o 그래프나 점수만 있고 해석이 없으면 해당 항목에 대해 최대 절반 감점 (시각 또는 표 자료가 최소 하나는 포함될 것)
  - o 차후 구두 확인에서 본인이 제출한 코드를 설명하지 못하면 감점이 이루어질 수 있음