Data Mining, Project (Part 2)

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 2) 내용

DBSCAN 클러스터링 알고리즘을 구현하고, 그 결과 분석 및 핵심 파라미터인 ε(eps, 이웃 반경)과 **MinPts**(core point가 되기 위한 최소 이웃 수) 선택의 근거 제시

- **데이터 전처리**: DBSCAN은 밀도 기반 알고리즘이므로, 20,531차원에서 밀도가 의미를 가지도록 어떻게 전처리할지 스스로 결정해 보기
  - o 최소 2가지 다른 전처리 설정(+전처리 하지 않은 설정)으로 실험하여, 클러스터링 결과 (클러스터 개수, outliers 비율 등)가 유지되는지 비교하고, 선택한 전처리의 타당성 분석
- **DBSCAN 구현** (강의 자료의 절차 참고)
  - o ε-neighborhood 탐색, core와 border 구분, core에서 시작하는 density-reachable (밀도-연결) 확장, outliers 처리(라벨 -1) 구현
  - o ε-neighborhood에 자기 자신이 포함되는지 여부, 사용한 MinPts 정의, 등 구현 조건을 보고서에 명시
- **결과 분석 및 파라미터 값 선택** (3개 이상의 MinPts 값 × 여러 개의 ε 값)
  - o MinPts 값 별로 k-거리 그래프(즉, 각 샘플에서 k번째 최근접 이웃까지의 거리를 오름차순 정렬한 그래프로 k 값은 MinPts 정의에 맞게 설정)를 그리고, outliers 후보(core가 되지 못하는 데이터 포인트)가 전체의 5%가 되는 지점의 거리를 ε으로 정할 것
  - o 각 MinPts 값에 따른 클러스터 개수, outliers 비율, 실루엣 점수(outliers 제외)를 직접 구현하고, 결과 표(또는 그래프)로 함께 제시하여 MinPts 선정 기준 설명
  - o 최종 파라미터 값을 이용한 클러스터링 결과를 2차원 산점도(PCA)로 시각화 (outliers는 구분해서 표시)

4. 제출

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

제출물	내용	형식
코드	명령 한번으로(예, python problem2.py) 보고서의 모든 표와 그림이 재현될 수 있도록 전처리, DBSCAN, 평가 지표, 실험 스크립트 코드를 구현할 것	.py + README.md(포함가능)
보고서	과제 내용의 단계 순서대로 작성, 분량은 3p이내	PDF

- 제출 마감일: 11월 30일(월)

5. 평가 기준

항목	배점	만점 기준
구현 정확성	30	DBSCAN이 절차대로(core와 border 구분, Density-Reachable 확장) 동작하는지 확인. 실루엣 등 평가 지표 계산 및 시각화 구현
전처리 근거	20	정규화·특징 선택·PCA 차원 수 등을 대안과 비교한 실험으로 뒷받침
결과 분석 및 파라미터 선택 근거	20	k-거리 그래프 등 ε 값과 MinPts 값의 도출 근거가 논리적인지, 전처리를 바꿔도 결론이 유지되는지 확인
합계	70	

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