과제 #2, Decision Tree

데이터마이닝 01분반, 고윤용 교수님

소프트웨어학부 20223908학번 김주영

**ENV.**

* Ubuntu 22.04
* Python 3.12.3
* Make 4.3

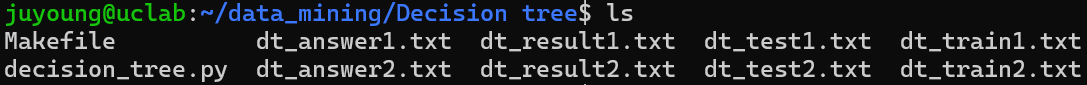
**Execute**

아래 사진과 같이 make 명령어를 실행하면 Makefile에 따라 파이썬 프로그램을 매개변수를 전달하며 실행한다. 총 2개의 데이터셋에 대해 각각 한 번씩 실행되어 총 2번 실행된다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

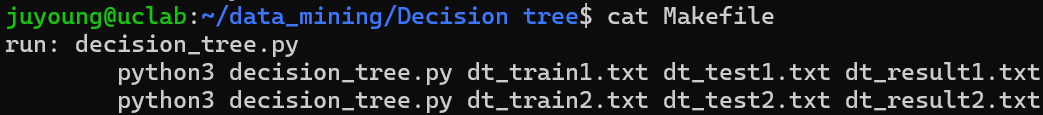
실행이 완료되면 동일한 디렉토리에 result 파일이 생성되고 이 파일 내에 결과가 저장된다.



텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

Makefile은 다음과 같이 작성하였다.



**Algorithm**

총 2개의 클래스, 4개의 함수를 통해 구현하였다.

아래는 명령 행 인자 개수를 확인하고 변수에 저장하는 코드이다.

MAX\_DEPTH 상수를 통해 결정 트리의 최대 깊이를 정할 수 있다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

먼저 트리를 정의하기 위해 아래와 같이 노드 클래스를 선언하였다.

Attribute와 label, depth를 필드로 가지며 노드가 리프인지 반환하는 메서드를 생성하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

아래 사진은 트리 클래스의 필드이다.

루트에 빌드 트리 메서드의 반환 값을 할당하며 루트 노드가 반환된다.

텍스트, 폰트, 스크린샷이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

아래 사진은 빌드 트리 메서드의 정의이며 학습하는 과정에서 사용된다.

모두 같은 레이블을 가지는 데이터셋, 이미 모든 attribute를 사용하여 더 이상 분할하지 못하는 경우, 트리의 최대 깊이를 넘어가는 경우(pre pruning)를 예외로 처리하였고 이러한 예외가 없다면 데이터를 분할한다.

Attribute 선정 기준은 gain ratio로 하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

아래는 학습된 트리를 통해 본 적이 없는 데이터를 분류하는 메서드이다.

현재 노드가 리프 노드가 아니라면 분류 기준에 해당하는 속성에 따라 다음 노드로 이동한다. 이를 리프 노드까지 반복하며 내려간다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

아래 사진은 디버깅 과정에서 트리의 구조를 확인하기 위해 구현한 함수이다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

1번 데이터셋에 대한 트리 구조는 다음과 같다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

2번 데이터셋에 대한 트리 구조는 다음과 같다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

아래 사진은 gain ratio를 구하기 위해 엔트로피를 구하는 함수이다.

엔트로피의 정의와 동일하게 작성하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

아래 사진은 gain ratio를 구하는 함수이다.

Gain ratio 또한 정의와 동일하게 작성하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

아래 사진은 정확도를 출력하는 함수이다. 과제 첨부파일로 주어진 dt\_answer 파일과 본 프로그램이 출력한 dt\_result 파일을 비교한다. 이 함수 또한 디버깅 과정에서 사용하고자 구현하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

1번 데이터셋에 대한 정확도는 다음과 같다



2번 데이터셋에 대한 정확도는 다음과 같다.



아래 사진은 전체적인 알고리즘 플로우이다.

먼저 주어진 파일에서 데이터를 리스트 형식으로 추출한다.

다음으로 각 속성이 가질 수 있는 값, 즉 도메인을 정의한다.

또한 학습 과정에서 이미 분류 기준으로 한 번 사용했던 속성을 또 사용하지 못하도록 사용한 속성인지 아닌지를 나타내는 attr\_flag를 정의하였다. 1이면 사용하지 않은 속성이고 0이면 사용한 속성을 나타낸다.

이후 훈련 데이터를 통해 결정 트리를 학습한다.

다음으로 테스트 데이터셋에 있는 데이터를 각 행마다 결정 트리에 입력하여 예측 레이블을 리스트로 추출한다.

마지막 줄에 정확도를 출력하는 과정도 있지만 과제 체점 환경에서는 answer 데이터가 주어지지 않으므로 주석처리 하였다.

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.