COSE362 Term Project Proposal

2021320024 김민서
2021320303 정지원
2021320322 윤민서
2021320323 허준환

1. 풀고자 하는 문제

이번 Term Project는 고려대학교 학우들의 강의 평가 데이터를 통해 학부 강의들을 어떤 유형으로 구분지을 수 있을지에 대한 문제를 해결하는 것을 목표로 한다. 대학생들이 생각하기에 가장 보편적이고 대표적인 강의 유형에는 후술할 '명강'과 '꿀강'이 있는데, 이번 프로젝트의 우선적인 목표는 이와 같이 최소 2개 이상의 강의 유형을 찾아내는 것이다.

또한, 분석 과정 및 결과에서 2가지의 강의 유형 이외에도 새로운 유형이 발견된다면, 그 유형을 새롭게 정의하고 특징을 분석하고자 한다.

2. 왜 이 문제가 흥미로운지에 대한 의견

대학생들에게는 수강신청이 매우 중요하다. 수강하는 학부 강의에 따라, 대학생들의 학점, 공부량, 휴학 여부 등이 결정되기 때문이다. 그래서 대학생들은 수강 신청을 하기 전에 강의를 들었던 사람들의 후기나 강의평 등을 고려하여 자신이 어떤 강의를 수강해야 할지에 대해 상당히 많은 고민을 한다.

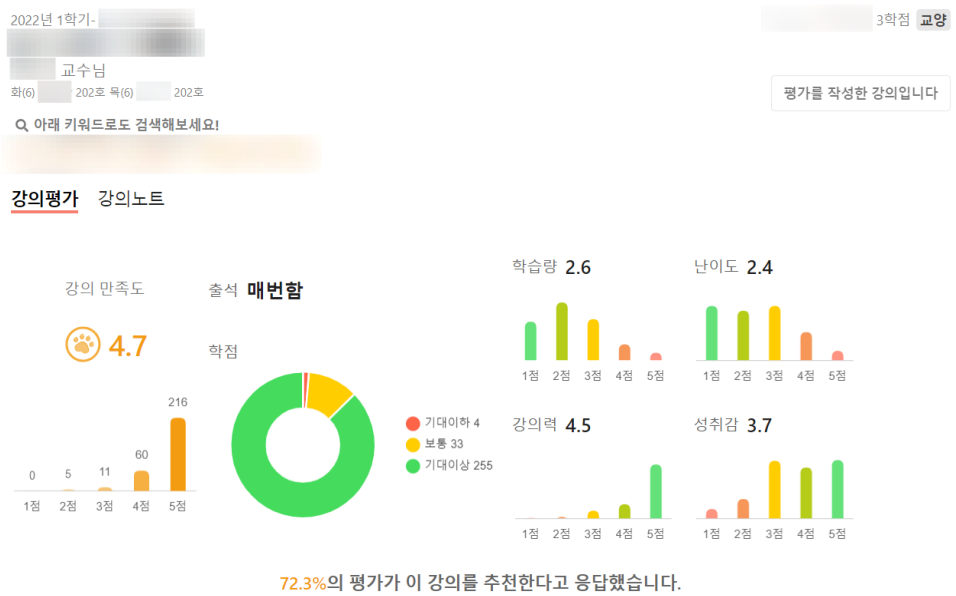
그러면서 대학생들은 학부 강의들을 몇 가지 유형으로 나누었는데, 앞서 말했던 명강과 꿀강이 그 예시이다. 여기서 '명강'은, 성적이 잘 안 나올지라도 교수님이 전달해 주시는 내용과 학문 자체에서 얻어갈 것이 많은 강의이며, '꿀강'은 강의 내용이 수강자에게 알차지 않거나 이미 배운 내용이 대부분 포함되어 있음에도 불구하고 단순히 최종 성적만을 바라보는 신청을 하는 강의이다. 이외에도, 따로 이름이 붙지는 않았으나 성적 산출이나 강의 진행, 강의력 등의 이유로 들으면 좋지 않은 강의 등이 있다.

대부분의 대학생들은 강의를 구분지을 때 강의평, 그리고 이에 주어지는 난이도, 학점, 학습량 등 여러가지 요소들을 참고하고 고려하면서 강의 유형을 파악한다. 하지만 이러한 유형이 있음에도 불구하고, 이 유형들에 대한 구분 기준은 명확하게 존재하지 않는다. 또한, 현재 4학기를 재학 중인 시점에서 앞으로의 전공 과목과 교양 과목을 이러한 구분 기준을 바탕으로 자신에게 적합한 강의를 더욱 신중하게 선택해야 할 필요가 있다. 그래서 이번 Term Project에서는 대학생들의 강의평 데이터를 바탕으로 강의 유형이 어떤 기준으로 나눠질 수 있는지 알아보고자 한다.

3. 주어진 문제 해결에 있어 ML의 적합성

- ML 방법론

Clustering에서 K-Means Algorithm을 중심으로 문제 해결을 진행할 예정이다. 데이터 수집은 고려대학교 전용 강의평가 사이트인 KLUE에서 진행한다.



<그림 1: 고려대학교 전용 강의평가 페이지 KLUE에 있는 강의>

KLUE의 강의평가에서 평가 요소로 작용하는 것은 그림 1에서 확인할 수 있는 것과 같이 수강 학점, 강의 구분, 교수, 요일 및 교시, 강의실 위치, 강의 만족도, 출석 빈도, 최종 성적의 만족도, 학습량, 난이도, 강의력, 성취감, 강의 추천 응답 비율 등이 있다. 이 feature들을 가진 데이터들을 clustering 한 후, 각 cluster에서 자연어로 쓰여진 강의 평가에 대해 word cloud를 비교해 보고 모델 학습이 잘 되었는지 확인 할 수 있다.

■ ML 방법론 선택 이유

우선, 강의 유형에 명확한 기준이 존재하지 않아 각 데이터에 대한 label이 존재하지 않는다. 그래서 데이터 준비 과정에 있어 label이 필요한 supervised learning이 아니라, label이 필요 없는 unsupervised learning이 적절하다. 또한 어떤 강의 유형이 존재하고, 어떠한 기준으로 나눠지는지 확인하기 위해서는 Clustering을 이용하는 것이 적절하다.

이뿐만 아니라, 앞서 말했듯 KLUE에 있는 강의평들을 통해 한 강의 데이터 당 얻을 수 있는 feature의 수는 14개 이상이 존재한다. 만일 이 feature들을 모두 사용한다면, dimension 수가 매우 커져 데이터를 처리하는 데 연산이 복잡해진다. K-Means Algorithm은 Clustering 중에서도 구현이 간단하고, 계산 속도가 빠르다는 장점이 있다. 고차원의 feature들을 다루면서 생기는 연산 복잡도의 문제로 인해, Clustering 방법론 중에서도 간단하고 빠른 K-Means Algorithm을 선택하였다.

4. 대략의 프로젝트 수행 일정

- 11월 첫째주 : 데이터 크롤링

모델 학습에 쓰일 강의들을 일정 기준에 따라 선별한 뒤 각 강의들의 강의 평가 데이터 수집

- 11월 둘째주 : 데이터 전처리 및 EDA

각각의 데이터들을 학습 가능하도록 전처리 e.g., 학점 feature를 0, 1, 2로 encoding
EDA를 통해 데이터의 분포를 시각적으로 살펴보고 어떤 feature들이 모델의 성능을 향상시키는데 좋은 영향을 줄 것인가에 대해 여러 가설을 세움
모델 학습에 쓸 feature들을 고르고 조건에 맞는 data set 준비

- 11월 셋째주 ~ 넷째주 : hyperparameter 여러 실험 및 모델 학습

모델의 성능을 향상시킬 수 있는 다양한 방법론 시도와 hyperparameter (feature의 종류, feature의 갯수 등) 를 수정해가며 model selection 작업을 진행

- 12월 : 모델 평가 및 최종 보고서 작성

여러 방법 및 test set으로 학습된 모델의 성능을 평가해보고 이를 바탕으로 보고서 작성

5. 예상되는 결과

강의평에 주어지는 여러 feature들 중 난이도와 학습량, 강의력과 성취감은 높은 relationship을 보일 것으로 예상된다. 따라서 이 두 relationship가 유의미할 것을 예상할 수 있으며, 이 외에도 다양한 feature들의 유의미한 relationship이 존재할 것이다.

또한, 앞서 말한 꿀강과 명강은 대체로 좋은 강의평, 즉 높은 별점을 받았다는 것을 확인할 수 있다. 이들의 차이점은 이 별점과 유의미한 relationship을 가지는 feature들에서 확인할 수 있는데, '꿀강'은 학점, '명강'은 성취감과 관련이 있다. 이를 통해 학점과 성취감이 높을수록 대체로 높은 별점을 받았다는 것을 알 수 있다.

그러나 단순히 학점과 성취감이라는 feature이 별점에 동등한 영향을 끼치는 것은 아니다. 대체로 성취감이 높은 강의는 학점과 무관하게 높은 별점을 받았다. 반면 학점을 잘주는 강의는 별점의 하한은 높지만 평균적인 관점에서 본다면 성취감이 높은 강의가 단순히 학점을 잘 주는 강의에 비해 더 높은 별점을 받은 것을 볼 수 있다. 이처럼 feature들 간의 relationship뿐만 아니라 그에 대한 association도 고려할 수 있을 것이다.

또한, 강의평의 별점과 강의 추천수의 관점에서 본다면, 대체로 별점이 낮은 강의는 강의 추천 수 또한 적고, 별점이 높은 강의는 강의 추천수가 높다는 것을 통해 별점과 강의 추천에 대한 relationship도 존재할 것이라고 예측할 수 있다.

위와 같은 relationship을 기반으로 clustering이 진행될 것이라고 생각한다.

강의평이 좋은 강의의 유형을 크게 두 가지로 나누어보면 앞서 말한 것과 같이 학습량이 적고 학점을 얻기 쉬워 강의평이 좋은 강의(꿀강)와 학점과는 별개로 수업을 통해 배워가는 것이 많아 강의평이 좋은 강의(명강)가 존재한다. 따라서 clustering을 통해 이러한 결과가 도출될 수 있을 것이라고 생각한다.