지침 및 가이드

진단 질문: NeurIPS 2020 교육 챌린지

Zichao Wang[1] ,* 앵거스 램[4*] , 에브게니 사벨리예프[4] ,
파슈미나 카메론[4] , 요르단 자이코프[4] , 호세 미구엘 에르난데스-로바토[245] , 리차드 E. 터너[24] , 리차드 G. 바라니욱[1] , 크레이그 바톤[3] , 사이먼 페이튼 존스[4] ,
사이먼 우드헤드[3†] , 쳉 장[4†]

초록

교육 분야에서 디지털 기술이 점점 더 널리 보급되면서 전 세계 학생들이 개인화된 고품질 교육 리소스에 액세스할 수 있게 되었습니다. 이러한 리소스 중에는 *진단 질문이*포함되어 있는데, 학생들이 이러한 질문에 대한 답변을 통해 학생들이 가지고 있을 수 있는 오해의 구체적인 성격에 대한 핵심 정보를 알 수 있습니다. 이러한 진단 질문과 학생의 상호 작용에서 비롯된 방대한 양의 데이터를 분석하면 학생의 학습 상태를 더 정확하게 파악할 수 있으므로 학습 커리큘럼 추천을 자동화할 수 있습니다. 이 대회에서 참가자들은 객관식 진단 문제에 대한 학생들의 답변 기록에 집중하여 1) 학생들이 어떤 답변을 제공할지 정확하게 예측하고, 2) 어떤 질문의 품질이 높은지 정확하게 사전 예측하고, 3) 학생의 답변을 가장 잘 예측하는 각 학생에 대한 개인화된 질문 순서를 결정하는 것을 목표로 합니다. 이러한 작업은 실제 교육 플랫폼의 목표와 매우 유사하며 오늘날 직면한 교육적 과제를 잘 대변합니다. 유니티는 전 세계 수천 명의 학생들이 매일 상호작용하는 선도적인 교육 플랫폼인 Eedi에서 수학 문제에 대한 학생들의 답변 2천만 개 이상의 예시를 제공합니다. 이 대회 참가자는 전 세계 수백만 명의 학생을 위한 맞춤형 교육의 질에 지속적이고 실제적인 영향을 미칠 수 있는 기회를 얻게 됩니다.

키워드

개인 맞춤형 교육, 비지도 학습, 결측치 예측, 능동 학습

*동등한 기여, 공동 제1저자 † 공동 시니어 저자[1] 라이스 대학교,[2] 케임브리지 대학교,
[3]Eedi,[4] 마이크로소프트 리서치,[5] 앨런 튜링 연구소.

콘텐츠

1 소개

1.1 배경 및 영향

배경 무료 또는 저렴한 온라인 교육 시스템이 널리 보급되면서 더 많은 사람들이 양질의 교육을 받을 수 있게 되었습니다. 이러한 플랫폼에서 학생들은 교육용 비디오를 시청하고, 대화형 코스 자료를 읽고, 학습 포럼에서 다른 학생 및 멘토와 대화하면서 학습할 수 있습니다. 학생의 이해도를 측정하기 위해 이러한 플랫폼 중 상당수는 평가 요소를 포함하고 있습니다. 이러한 평가를 통해 수집된 데이터를 마이닝하면 이론적으로는 학생들의 학습 방식과 같은 유용한 교육 정보를 추출하고 학습 결과를 개선하기 위한 적절한 학습 개입을 추천할 수 있습니다. 그러나 도출된 인사이트의 품질은 평가에 포함된 문제의 품질에 따라 달라집니다.

형성 평가는 수업이 진행되는 동안 학생의 학습을 개선하는 데 사용할 수 있는 세부 정보를 이끌어내는 평가를 신중하게 설계하는 것과 관련이 있습니다. 형성 평가에 사용되는 놀랍도록 단순하지만 강력한 질문 유형은 *진단 질문입니다*.

진단 문제는 정답이 4개인 객관식 문제로, 정답은 정확히 하나이며 오답 3개는 각각 일반적인 오해를 강조하기 위해 선택됩니다. 학생이 진단 문제를 틀린 경우 교육자는 그 이유를 추측할 수 없습니다. 학생이 오답을 선택하면 오해의 본질에 대한 정보가 드러나며, 이는 오해를 해결하는 데 도움이 되는 귀중한 정보입니다[12]. 진단 질문은 오답과 관

련된 정보를 검색하도록 유도하기 위해 구성될 수 있습니다. 따라서 학생들은 다음과 같은 이유를 고려해야 합니다.

정답은 정답이지만 오답은 왜 오답인지에 대해 설명합니다[7].

오답 하나하나가 그럴듯한 방해 요소가 되는 좋은 진단 문제를 작성하는 것은 어려운 일입니다. 숙련된 교사조차도 시간이 많이 걸리는 작업이라고 생각합니다. 이러한 문제를 해결하기 위해 Eedi(https://eedi.com)의 개발팀은 교사들이 진단 문제를 크라우드소싱할 수 있는 플랫폼(https://diagnosticquestions.com/)을 만들었습니다. 필연적으로 생성된 질문의 품질에는 차이가 있을 수밖에 없습니다.

교사가 진단 문제를 만들 때는 오답을 하나씩 선택하여 일반적인 오해를 강조해야 하지만, 이러한 오해를 레이블로 표시하거나 문제 간에 연결하지 않습니다. 오답이 너무 잘못 선택되어 명백히 틀린 것이므로 학생이 절대 선택하지 않을 가능성이 전적으로 있습니다.

학생의 학습을 정확하게 진단하려면 좋은 질문을 제시하는 것이 필수적입니다. 좋은 진단 질문은 학생의 오해의 구체적인 성격을 파악합니다. 질문은 모호하지 않아야 하며, 중요한 것은 학생이 핵심 오해를 유지한 채로 정답을 맞출 수 없어야 한다는 것입니다. 또한 교수 학습 시간은 제한되어 있으므로 학생의 지식과 오해에 대한 정보를 가장 많이 파악할 수 있는 질문의 우선순위를 정해야 합니다.

대회 개요 및 교육에 미치는 영향 이 대회에서는 참가자들에게 학생들의 학습을 이해하고 개선하며 진단 문제의 질을 측정할 수 있는 새로운 방법론을 개발하도록 요청합니다. 섹션 3에 설명되어 있고 여기에 요약된 네 가지 과제가 있습니다:

1. 첫 번째 작업은 학생이 질문에 올바르게 답할지 여부를 예측하는 것입니다. **실제 영향력:** 특정 학생의 배경과 학습 상태에 가장 적합한 적절한 난이도의 문제를 추천할 수 있습니다.
2. 두 번째 작업은 이를 확장하여 학생들이 각 문제에 대해 어떤 답을 선택할지 예측하는 것입니다.
   실제 영향력: 동일하거나 관련된 오해를 나타낼 수 있는 질문-답변 쌍을 클러스터링하여 학생들이 가지고 있는 잠재적인 공통 오해를 발견할 수 있습니다.
3. 세 번째 과제는 질문의 품질을 측정할 수 있는 지표를 고안하는 것입니다. 이 지표는 도메인 전문가의 의견에 따라 평가됩니다.
   실제 영향력: 진단 문항 작성자에게 피드백을 제공하여 품질이 낮은 문항을 수정하고 교사가 학생을 위한 문항을 선택할 수 있도록 안내할 수 있습니다.
4. 네 번째 과제는 보이지 않는 문제에 대한 학생의 성적 예측을 위해 학생으로부터 제한된 답변 세트를 수집하는 것입니다. 이를 위해서는 개인화된 머신러닝이 필요합니다.

정보의 가치를 추정하는 학습 방법.
실제 영향력: 각 학생에게 맞춤화된 평가를 제공하여 학습 결과를 개선할 수 있습니다.

이 대회는 많은 교육 플랫폼에서 흔히 볼 수 있는 학습 분석 및 개인화 작업을 모방한 과제를 통해 교육 데이터 마이닝에 대한 심층적인 소개를 제공합니다. 이 대회에서는 이미 대규모로 배포되어 사용되고 있는 교육 플랫폼의 데이터를 사용합니다. 이 데이터는 실제 학생들이 실제 질문에 대한 실제 답변을 설명합니다. 실제 교육 데이터와 실제 문제를 매력적인 방식으로 해결할 수 있는 기회를 제공함으로써, 이 대회는 교육 분야에서 중요한 머신 러닝 분야로 인재를 유치할 것입니다.

이번 대회를 통해 교육 데이터 마이닝 기술, 특히 학생의 학습 진도를 분석하고 개인화된 학습 커리큘럼을 추천하는 기술이 근본적으로 발전할 것으로 기대합니다. 이러한 방법은 실제 교육 플랫폼에 적용되어 수백만 명의 학생들의 학습 성과를 개선할 수 있을 것입니다.

머신러닝 커뮤니티에 미치는 영향 이 대회에는 해결해야 할 여러 가지 근본적인 머신러닝 과제가 있습니다. 추천 시스템에서 흔히 볼 수 있지만 교육 데이터 마이닝의 맥락에서 나타나는 몇 가지 과제는 다음과 같습니다. 각 학생이 전체 질문의 극히 일부만 답했기 때문에 데이터의 희소성을 어떻게 처리할 것인가? 학생 인구 통계와 같은 학생 및 문제 메타데이터를 효과적으로 사용하여 예측을 개선하는 방법은 무엇인가요? 예측 정확도를 극대화하기 위해 질문의 순서를 최적으로 선택하는 방법은 무엇일까요? 또 다른 과제는 정렬되지 않은 범주형 데이터에 대해 매트릭스 완성을 효과적으로 수행하는 방법입니다. 교육 데이터 마이닝의 고유한 맥락에서 이러한 과제를 해결하는 것은 NeurIPS 커뮤니티와 더 넓게는 머신 러닝 커뮤니티 전체에 중요한 기술적 관심사가 될 것입니다.

1.2 대회 탐색하기

이 문서는 대회에 대한 소개와 함께 참가자들이 대회를 탐색할 수 있도록 안내하는 역할을 합니다. **본 문서의 나머지 부분에는 대회의 다양한 측면에 대한 중요한 정보가 포함되어 있습니다**. 참가자는 대회 기간 동안 다음 정보를 숙지하고 정기적으로 다시 참조할 것을 권장합니다:

- 참가자가 받게 될 파일의 마스터 목록과 다운로드할 수 있는 위치(섹션 1.3);

- 데이터 파일에 대한 자세한 설명(섹션 2);
- 각 과제에 대한 평가 지표를 포함하여 대회의 각 과제에 대한 자세한 설명(섹션 3);
- 각 과제를 코다랩에 제출하기 위한 지침(섹션 4);
- 데이터를 로드하고, 로컬 평가를 수행하고, 샘플 제출을 준비하기 위한 스크립트를 제공하는 시작 가이드(섹션 5).

1.3 경쟁 파일 목록

아래는 참가자가 사용할 수 있는 파일 목록입니다. 아래 파일은 두 개의 zip 파일로 구성되어 있으며, 두 파일 모두 코다랩 경진대회 웹사이트에서 찾을 수 있습니다. 참가 탭을 클릭한 다음 데이터 가져오기를 클릭합니다.

공개 데이터 세 개의 하위 폴더가 있는 데이터 폴더가 포함되어 있습니다(각 파일에 대한 자세한 내용은 섹션 2에 나와 있습니다):

```
데이터
├── 이미지
│   └── 0.jpg, 1.jpg, ..., 947.jpg 메
├── 타데이터
│   ├── 답변 메타데이터 작업 1 2.CSV
│   ├── 답안 메타데이터 작업 3 4.CSV 질
│   ├── 문 메타데이터 작업 1 2.CSV 질문
│   ├── 메타데이터 작업 3 4.CSV
│   ├── 학생 메타데이터 작업 1 2.csv
│   ├── 학생 메타데이터 작업 3 4.csv
│   └── 과목 메타데이터.csv
└── 데이터 훈련
    ├── 훈련 작업 1 2.CSV
    └── 훈련 작업 3 4.CSV
```

시작 키트 참가자가 데이터 로드, 로컬 평가 및 제출 준비를 시작하는 데 도움이 되는 폴더 스타터 키트가 포함되어 있습니다:

```
스타터 키트
├── 제출 템플릿
│   ├── 제출 작업 1 2.csv 제출 작업
│   ├── 3.csv 제출 모델 작업 4.py
│   └──
├── 작업-1
│   ├── 로컬 데이터 분할.py 로컬
│   ├── 평가.py 샘플 모델 대다수
│   └── .py
├── 작업-2
│   ├── 로컬 데이터 분할.py 로컬
│   ├── 평가.py 샘플 모델 대다수
│   └── .py
├── 작업-3
│   ├── 로컬 평가.py 샘플 모델 엔
│   └── 트로피.py
└── 작업-4
    ├── 로컬 데이터 split.py
    ├── numpy
    │   ├── local_evaluation.py
    │   ├── model.py
    │   └── train-model.py 파
    └── 이토치
        ├── local_evaluation.py
        ├──
        └──
```

model.py

train_model.py

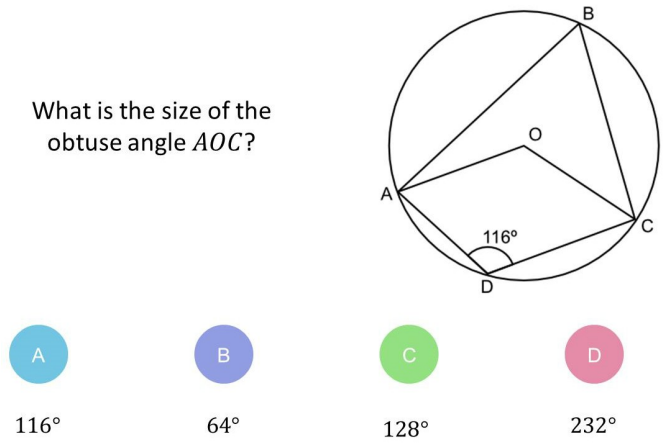

그림 1: 분석 데이터가 수집되는 교육 플랫폼의 질문 예시.

2 데이터

표 1: 데이터 예시.

QuestionId	UserId	AnswerId	AnswerValue	정답	IsCorrect
10322	452	8466	4	4	1
2955	11235	1592	3	2	0
3287	18545	1411	1	0	0
10322	13898	6950	2	1	0

현재 수만 개의 학교에서 사용되고 있는 온라인 교육 제공업체인 Eedi에서 제공하는 광범위한 데이터를 통해 2018년 9월부터 2020년 5월까지 제공된 객관식 주관식 문제에 대한 학생의 응답을 자세히 확인할 수 있습니다. 이 플랫폼은 초등학생부터 고등학생(대략 7세에서 18세 사이)을 대상으로 크라우드 소싱된 진단 문제를 제공합니다. 각 진단 문제는 4지 선다형 문제이며, 정답은 정확히 하나만 선택할 수 있습니다. 현재 이 플랫폼은 주로 수학 문제에 중점을 두고 있습니다. 그림 1은 플랫폼의 문제 예시를 보여줍니다. 모든 데이터는 코다랩의 대회 홈페이지에서 다운로드할 수 있습니다(섹션 1.3 참조).

대회는 4개의 과제로 나뉘는데, 과제 1과 2는 과제 3과 4와 마찬가지로 하나의 데이터 세트를 공유합니다. 이러한 데이터 세트는 형식은 거의 동일하지만 서로 다른 질문 세트를 사용합니다.

모든 질문아이디, 사용자아이디, 답변아이디는 익명으로 처리되어 제품에서 발견되는 것과 식별할 수 없는 관계가 없습니다. 작업 1과 작업 2의 모든 ID는 작업 3과 작업 4의 ID와 별도로 익명화됩니다. **중요: 질문, 사용자 및 답변 ID는 이러한 작업 쌍의 데이터 간에 연결되지 않아야** 합니다! 이는 두 데이터 세트가 모두 독립적으로 유지되도록 하기 위한 설계입니다.

2.1 기본 데이터

이것은 학생들이 질문에 대한 답변 기록으로 구성된 기본 훈련 데이터입니다. 이 파일은 `train_task_1_2.csv` 및 `train_task_3_4.csv` 파일에서 찾을 수 있습니다. 열은 다음과 같습니다:

- **QuestionId:** 답변된 질문의 ID입니다.
- **UserId:** 질문에 답변한 학생의 ID입니다.
- **AnswerId:** 연결된 답변 메타데이터와 조인하는 데 사용되는 (QuestionId, UserId) 쌍의 고유 식별자입니다(아래 참조).
- **IsCorrect:** 학생의 정답 여부를 나타내는 이진 표시기입니다(1은 정답, 0은 오답).
- **정답:** 객관식 문제의 정답([1,2,3,4]의 값)입니다.
- **AnswerValue:** 객관식 문제에 대한 학생의 답입니다(값은 [1,2,3,4]입니다).

표 1은 이 형식의 데이터 레코드 4개를 나타낸 그림입니다. 각 학생은 일반적으로 가능한 모든 문제 중 극히 일부만 답했기 때문에 매트릭스는 매우 희박합니다. 과제 1과 2에서는 50개 미만의 답을 받은 문제와 50개 미만의 문제에 답한 학생을 제거했습니다. 고정된 문제 집합에 관심이 있는 과제 3과 4의 경우, 50개 미만의 문제에 답한 학생을 모두 제거했습니다. 또한 한 학생이 동일한 문제에 대해 여러 개의 답안 기록을 가지고 있는 경우 가장 최근의 답안 기록을 유지합니다. 데이터는 각 행이 학생을 나타내고 각 열이 문제를 나타내는 행렬 형식으로 변환할 수 있습니다.

과제 1과 2의 경우, 개별 답안 기록은 80%/10%/10% 훈련/공개 테스트/개인 테스트 세트로 무작위로 분할됩니다. 작업 3과 4의 경우, 사용자 아이디는 80%/10%/10% 훈련/공개 테스트/개인 테스트 세트로 무작위로 분할됩니다. 이러한 전처리 단계를 통해 다음과 같은 크기의 훈련 데이터 세트가 생성됩니다:

- 과제 1 및 2: 27613개의 문제, 118971명의 학생, 15867850개의 답안

- 과제 3 및 4: 948개 문제, 4918명의 학생, 1382727개의 답변

이러한 훈련 세트의 총 답변 레코드 수는 1,700만 개를 초과하므로 수동 분석은 비실용적이며 데이터 기반의 머신 러닝 접근 방식이 필요합니다. 데이터의 행렬 표현에 대한 그림은 섹션 3의 그림 2와 3을 참조하십시오.

2.2 질문 메타데이터

각 질문에 대해 다음과 같은 메타데이터를 제공합니다:

- **SubjectId** 각 과목은 수학의 한 영역을 다루며, 세부적인 수준은 다양합니다. 저희는 목록에서 문제와 관련된 각 주제에 대한 ID를 제공합니다. 예를 들어 "대수", "데이터 및 통계", "기하와 측정" 등이 있습니다. 이러한 주제는 트리 구조로 배열되므로 예를 들어 "인수분해"는 "단일 대괄호로 인수분해"의 상위 주제입니다. 이 트리에 대한 자세한 내용은 각 SubjectId와 관련된 주제 이름 및 트리 수준과 상위 주제의 SubjectId가 포함된 추가 파일 `subject metadata.csv`에 제공됩니다.
- **문제 내용:** 과제 3과 과제 4에서는 주제 외에도 각 문제에 대해 그림 1과 같이 각 문제에 대해 학생에게 제시된 이미지가 제공됩니다. 문제 이미지는 본 대회의 목적으로만 공유되며 다른 용도로 사용해서는 안 됩니다. 문제 이미지는 대회 외부의 누구와도 인쇄하거나 공유해서는 안 됩니다. 문제 문구는 이미지에 포함되어 있지만 텍스트로는 제공되지 않습니다.

2.3 학생 메타데이터

데이터 집합의 학생에 대해 다음과 같은 메타데이터가 제공됩니다:

- **UserId:** 학생을 고유하게 식별하는 ID로, 기본 데이터 집합에 조인할 수 있습니다.
- **성별: 성별:** 학생의 성별(사용 가능한 경우)입니다. 0은 지정되지 않음, 1은 여성, 2는 남성, 3은 기타입니다.
- **DateOfBirth:** 학생의 생년월일(월 1일로 반올림)입니다.
- **프리미엄 학생:** 학생이 재정적으로 취약하여 무료 학교 급식 또는 학생 프리미엄을 받을 자격이 있는지 여부입니다.

2.4 답변 메타데이터

데이터 집합의 각 개별 답안 레코드에 대해 다음과 같은 메타데이터가 제공됩니다:

- **UserId:** 기본 데이터 집합에 조인할 수 있는 답변을 고유하게 식별하는 ID입니다.
- **DateAnswered:** 질문에 답변한 시간 및 날짜(가장 가까운 분 단위)입니다.
- **자신감: 신뢰도:** 정답에 대한 백분율 신뢰도 점수입니다. 0은 무작위 추측, 100은 총 신뢰도를 의미합니다.
- **GroupId:** 학생이 문제를 할당받은 클래스(학생 그룹)입니다.
- **QuizId:** 학생이 답한 질문이 포함된 할당된 퀴즈입니다.
- **작업 계획 ID:** 학생에게 문제가 할당된 작업 계획입니다.

3 작업 세부 정보

이 섹션에서는 대회 과제와 평가 지표를 소개합니다.

대회는 다양한 스타일의 네 가지 과제로 구성됩니다. 처음 두 과제는 데이터 세트의 모든 질문에 대한 학생의 응답을 예측하는 것을 목표로 합니다. 이 두 과제는 추천 시스템 과제[5, 2, 3] 또는 결측치 대입 과제[8, 13, 10, 4]와 같이 여러 가지 방식으로 공식화할 수 있습니다. 여기서 각 학생은 질문의 일부에만 답하고 다른 질문에 대한 학생의 답변은 개인화된 교육에 큰 관심을 갖습니다. 세 번째 과제는 교육 영역에서 필수적이며 미해결 과제로 남아 있는 질문의 품질을 평가하는 데 중점을 둡니다[11]. 마지막 과제는 개인화된 동적 의사 결정[1, 6, 9]이 필요한 개인화 교육의 과제를 직접적으로 다룹니다. 처음 두 과제는 후자의 두 과제에 대한 솔루션을 구축하는 데 유용한 기반이 될 수 있지만, 반드시 이러한 접근 방식을 취할 필요는 없습니다. 참가자는 원하는 과제에 대한 솔루션을 원하는 순서대로 자유롭게 제출할 수 있습니다.

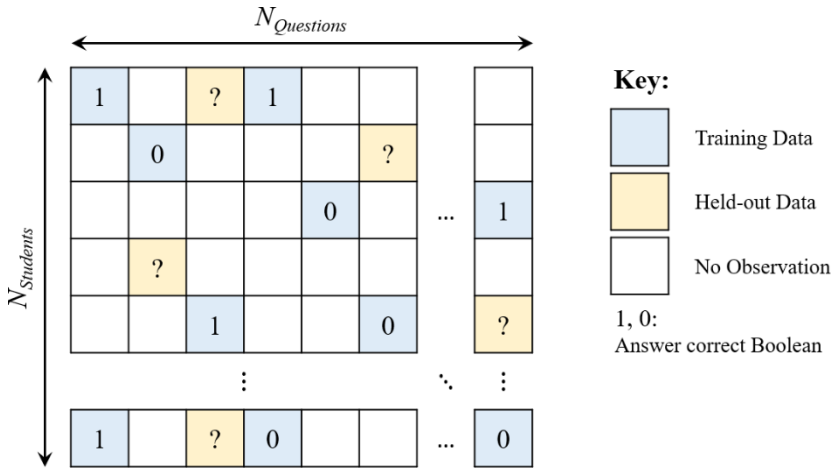

그림 2: 작업 1에 대한 데이터의 스파스 행렬 표현 그림.

3.1 과제 1: 학생의 반응 예측 - 맞거나 틀림

첫 번째 작업은 학생이 문제에 정답을 맞혔는지 예측하는 것입니다. 이 작업에 사용되는 기본 데이터는 마지막 열이 이진 값인 기록 테이블(학생 ID, 문제 ID, 정답 표시자)입니다. 이 데이터의 스파스 행렬 표현은 그림 2에 나와 있습니다. 참가자는 학생 답변의 보류된 하위 집합에 대한 정답률 지표를 예측하도록 요청받습니다. 보다 구체적으로, 각 학생에 대해 사용 가능한 기록의 일부가 평가가 수행될 숨겨진 테스트 집합으로 보류됩니다.

아직 정답이 없거나 새로 도입된 문제에 대한 학생의 정답을 예측하는 것은 실제 개인화 교육 플랫폼에서 학생의 능력 수준을 추정하는 데 매우 중요하며, 고급 작업을 위한 토대가 됩니다. 이 작업은 행렬 완성 클래스에 속하며, 이진 데이터의 경우 추천 시스템 영역에서 흔히 볼 수 있는 문제를 연상시킵니다. 행렬 인수분해 또는 최인접 기반 방법과 같이 이 분야에서 널리 사용되는 접근 방식이 이 작업에 효과적일 수 있습니다.

3.1.1 평가 지표

예측 정확도, 즉 실제 정확도 지표와 일치하는 예측 수를 총 예측 수로 나눈 값을 측정 지표로 사용합니다.

테스트 세트):

$$\text{정확도} = \frac{\#\text{정확한 예측}}{\#\text{총 예측 수}}$$

3.2 과제 2: 학생 반응 예측 - 답안 예측:

두 번째 작업은 학생이 특정 문제에 대해 어떤 답을 했는지 예측하는 것입니다. 이 작업에 사용되는 기본 데이터는 레코드 테이블(`StudentId`, `QuestionId`, `AnswerValue`, `CorrectAnswer`)이며, 마지막 두 열은 [1, 2, 3, 4]의 값을 취하는 범주형입니다(각각 객관식 답안 옵션 A, B, C 및 D에 해당). 따라서 과제 1에 표시된 스파스 행렬 표현은 이제 그림 3과 같이 보입니다. 데이터 세트의 질문은 모두 객관식이며, 각각 4개의 선다형과 1개의 정답이 있으므로 이 과제는 다중 클래스 예측 문제로, 참가자는 (StudentId, QuestionId) 쌍의 숨겨진 보류된 하위 집합에 대한 학생들의 응답을 예측해야 합니다.

학생의 답에 대한 실제 객관식 옵션을 예측하면 학생이 주제에 대해 가질 수 있는 일반적인 오해를 분석할 수 있으므로 실제 교육 플랫폼에서 개인화된 조언 및 안내의 기초를 형성할 수 있습니다. 상관관계가 높은 질문-답변 쌍의 클러스터는 동일하거나 관련된 오해에 해당할 수 있습니다. 오해 사이의 관계를 이해하는 것은 커리큘럼 개발을 위해 해결해야 할 중요한 문제이며, 주제를 가르치는 방식과 주제 순서에 영향을 줄 수 있습니다.

과제 1과 마찬가지로 행렬 완성 과제이지만 이번에는 정렬되지 않은 범주형 데이터를 사용합니다. 일반적으로 응답이 이진 또는 서수(예: 별 1~5개)인 추천 시스템 영역에서는 이러한 유형의 데이터가 드물기 때문에 학생의 답변을 정확하게 예측하고 오개념을 정확하게 모델링하려면 보다 새로운 접근 방식이 필요할 수 있습니다. 이러한 분석을 목표로 하는 대부분의 모델은 학생을 정확하게 모델링하여 답을 모델링하는 데 의존하기 때문에 처음 두 과제는 학생의 학습 분석의 기초를 형성합니다. 따라서 이 두 가지 과제에서 경쟁하는 참가자는 교육 데이터 마이닝의 기본 과제에 노출됩니다.

3.2.1 평가 지표

정답이 이진형이 아닌 범주형이라는 점을 제외하면 위와 동일한 메트릭 **예측 정확도를** 사용합니다.

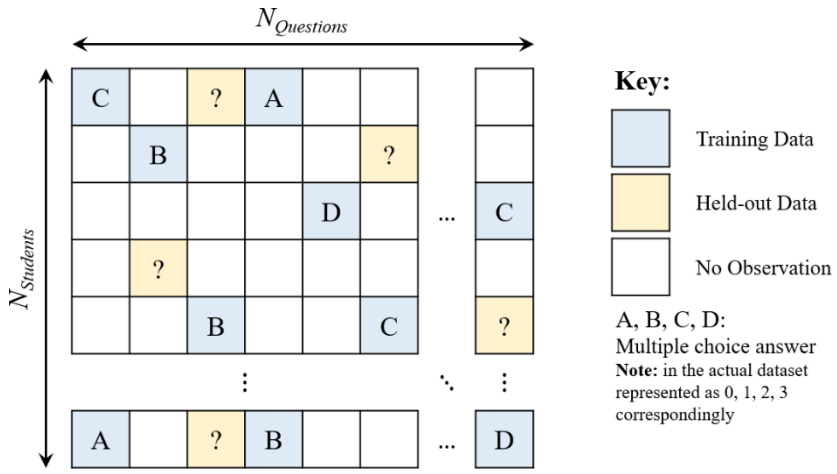

그림 3: 작업 2에 대한 데이터의 스파스 행렬 표현 그림.

3.3 과제 3: 글로벌 문제 품질 평가

세 번째 과제는 데이터 세트에서 찾은 학생의 답변에서 학습한 정보를 기반으로 도메인 전문가 패널(숙련된 교사)이 정의한 문제의 '품질'을 예측하는 것입니다. 이 작업을 수행하려면 전문가의 문제 품질 판단을 모방한 문제 품질을 평가하는 지표를 정의해야 합니다. 결정적으로, 전문가의 판단은 대회 참가자에게 제공되지 않습니다. 따라서 이 과제는 앞의 두 과제와는 성격이 매우 다르며, 비지도 학습 문제로서 혁신적인 사고를 요구합니다.

참가자를 안내하기 위해 전문 교사가 직관 및 사용하는 기준을 포함하여 질문의 품질을 판단하는 방법에 대한 메모를 제시합니다. 참가자는 이 자료를 사용하여 자동 문제 품질 지표를 설계할 수 있습니다. 전문가 데이터 수집에 사용된 프롬프트의 예는 그림 4에 나와 있습니다. 또한 도메인 전문가 중 한 명인 크레이그 바튼은 다음과 같은 양질의 문제 설계에 대한 "황금률"을 확인했습니다.[1]:

- 명확하고 모호하지 않아야 합니다.
- 단일 기술/개념을 테스트해야 합니다.
- 학생은 10초 이내에 답할 수 있어야 합니다.

[1] https://medium.com/eedi/what-makes-a-good-diagnostic-question-b760a65e0320

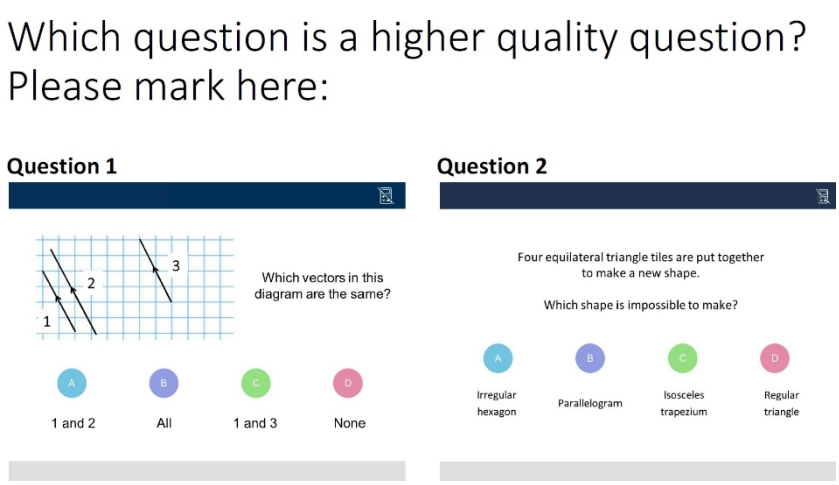

그림 4: 쌍별 상대 평가 문제 품질에 대한 전문가의 판단을 수집하는 데 사용되는 프롬프트의 예입니다. 이 외에도 전문가에게는 다음과 같은 지침이 제공됩니다: *다음 각 슬라이드에는 왼쪽과 오른쪽에 각각 2 개의 질문이 있습니다.*

동점자는 허용되지 않으므로 더 높은 점수를 받은 문제를 선택해 주세요.

- 학생이 설명할 필요 없이 각 오답에서 무언가를 배울 수 있어야 합니다.
- 핵심적인 오해가 남아 있는 상태에서 질문에 올바르게 답할 수는 없습니다.

이 작업에서 설계된 문제 품질 측정 항목은 교사가 제출한 문제의 품질을 평가할 수 있는 확장 가능한 방법을 제공하기 때문에 크라우드 소싱 교육 애플리케이션에서 가장 중요한 요소입니다. 크라우드 소싱된 질문의 품질은 학생과 교사에게 플랫폼의 유용성을 직접적으로 반영합니다. 또한 품질 평가는 교사를 위한 맞춤형 가이드로도 사용할 수 있어 교사가 문제의 질을 개선하는 데 도움이 됩니다.

참가자들은 이 지표를 정의할 때 이전 과제에서 사용한 머신 러닝 모델을 활용하도록 권장됩니다. 참가자는 이 지표를 기반으로 데이터 세트의 질문 ID에 각각 고유하게 매핑되는 1부터 N까지의 순위를 제공해야 합니다. 순위 1은 질문 품질이 가장 높은 질문에 해당해야 하며, 질문 품질이 낮은 순서대로 순위를 매겨야 합니다. 메트릭의 절대값은 필요하지 않습니다. 필요한 출력에 대한 예시는 그림 5를 참조하세요. 평가 절차와 메트릭은 섹션 1.5에 설명되어 있습니다.

이 작업은 명시적인 상위 학습자가 없으므로 비지도 학습 작업으로 볼 수 있습니다.

Example submission:

Quality ranking	*Question ID*
1	Q-2748
2	Q-124
3	Q-3915
⋮	⋮
5,125	Q-10029
⋮	⋮
13,369	Q-6715

Scoring procedure:

Question-pair sampled for quality comparison	*Submission ranking implies*	*Expert judgement of quality*					
		Expert A		*Expert B*		*Expert C*	
Q-124 ≧ Q-10029	>	>	1	>	1	>	1
Q-11092 ≧ Q-3999	<	>	0	>	0	>	0
Q-844 ≧ Q-7491	<	<	1	<	1	<	1
Q-2748 ≧ Q-9882	>	>	1	<	0	<	0
Q-13001 ≧ Q-9115	>	>	1	>	1	<	0
Agreement with expert:		**0.80**		0.60		0.40	
Max agreement (task score):		**0.80** (Expert A)					

Key:

- ≧ — Compare quality between question on the left (L) and question on the left (R)
- > — Question L greater quality than question R
- < — Question R greater quality than question L
- 1 — Submission matches expert judgement
- 0 — Submission doesn't match expert judgement

그림 5: 과제 3의 채점 과정 그림. *왼쪽*: 과제 3에 대한 제출물의 예상 형식 - 문제 ID에 대한 문제 품질 순위(품질이 낮은 순서대로)입니다. *오른쪽*: 성과 지표 산출에 대한 그림으로, 본문의 단계(3.3.1절)를 참조하십시오. 이 예에서는 5개의 문제 쌍과 3명의 전문가를 사용합니다.

비전 레이블을 문제 품질에 사용할 수 있습니다. 정보 이론, 특징 선택, 순위 학습과 같은 영역에서 얻은 인사이트가 이 작업과 관련이 있을 수 있습니다.

3.3.1 평가 지표

참가자는 문제에 대한 품질 순위를 제출합니다(그림 5 *왼쪽*). 그런 다음 보이지 않는 문제 쌍 세트를 사용하여 이 순위의 품질(따라서 기본 지표)을 평가합니다. 각 문제 쌍에서 어떤 문제가 더 높은 품질인지에 대한 전문가들의 판단을 수집했다는 점에 유의하세요. 평가 단계는 다음과 같습니다(그림 5 *오른쪽의 그림*참조):

- 제출된 순위를 기준으로 각 쌍의 문제 중 더 높은 품질의 문제를 결정합니다.
- 이를 각 전문가의 판단과 비교합니다(일치하는 경우 1, 일치하지 않는 경우 0을 할당합니다).
- 각 전문가 i *에*/ 대해 합의 비율을 결정합니다: $_i = \frac{N_{\text{일치하는 쌍}}}{N_{\text{총 쌍}}}$ -
- 다음 합의 분수의 *최대값을* 구합니다. A_{max} = 최대$_i$ A_i . 이것이 이 작업의 최종 평가 지표로 사용됩니다.

전문가의 판단에 매우 근접할 수 있는 지표를 찾고 있으므로 동의의 평균이 아닌 동의 비율의 최대값을 사용합니다.

를 모든 전문가에 대한 분수로 계산합니다. 이러한 접근 방식을 사용하는 이유는 전문가의 품질 지표 자체가 주관적이기 때문이며, 머신 러닝을 사용하여 특정 전문가의 접근 방식이 특히 잘 근사화될 수 있는지 알아내는 것이 흥미롭기 때문입니다.

3.4 작업 4: 맞춤 질문

네 번째 과제는 남은 답변에 대한 모델의 예측 정확도를 극대화하기 위해 학생에게 물어볼 일련의 질문을 대화형으로 생성하는 것입니다. 구체적으로, 참가자의 모델에는 질문에 대한 답변이 완전히 숨겨진 이전에 본 적이 없는 학생 세트와 각 학생에 대해 쿼리할 수 있는 잠재적 질문 세트가 제공됩니다. 그런 다음 모델은 이러한 각 학생에 대해 차례로 쿼리할 개인화된 질문을 선택하면 해당 답변이 모델에 공개됩니다. 이 정보를 바탕으로 모델은 각 학생에 대해 쿼리할 두 번째 질문을 선택하는 식으로 총 10개의 질문이 출제될 때까지 반복합니다.

이 과제의 목표는 각 학생의 10개의 답변에 모델이 노출된 후, 각 학생에 대한 보류된 질문 세트에 대한 참가자의 모델의 예측 정확도를 극대화하는 것입니다. 이 과제는 학생과 교사의 시간을 가장 효율적으로 활용하기 위해 가능한 최소한의 질문을 던지면서 다양한 개념에 대한 학생의 이해 수준을 정확하게 진단하고자 하는 개인 맞춤형 교육에 있어 매우 중요합니다. 또한 이 과제는 참가자가 모델의 불확실성에 대해 효과적으로 추론하고 데이터를 최대한 효율적으로 사용해야 하는 중요한 머신러닝 과제이기도 합니다.

이 작업은 능동 학습, 강화 학습, 밴디트 알고리즘, 베이지안 실험 설계, 베이지안 최적화 등 여러 관련 분야의 렌즈를 통해 볼 수 있으며, 이러한 분야에서 얻은 인사이트는 유용하게 사용될 수 있습니다.

3.4.1 평가 지표

제출된 모델은 보류된 학생 집합의 모든 학생에 대해 10개의 쿼리 문제를 순차적으로 선택하도록 요청받게 됩니다. 각 선택 단계가 끝나면 이러한 학생-문제 쌍에 대한 범주형 답안과 이진 정답 지표가 모두 모델에 비공개로 공개됩니다. 그런 다음 모델에 이 새로운 데이터를 통합하거나 각 문제 후에 재학습할 수 있는 기회가 주어집니다. 각 학생에 대해 10개의 답을 받은 후, 모델은 각 학생에 대한 보류된 시험 세트에 대한 이진 정답률 지표를 예측하는 예측 정확도를 평가받게 됩니다.

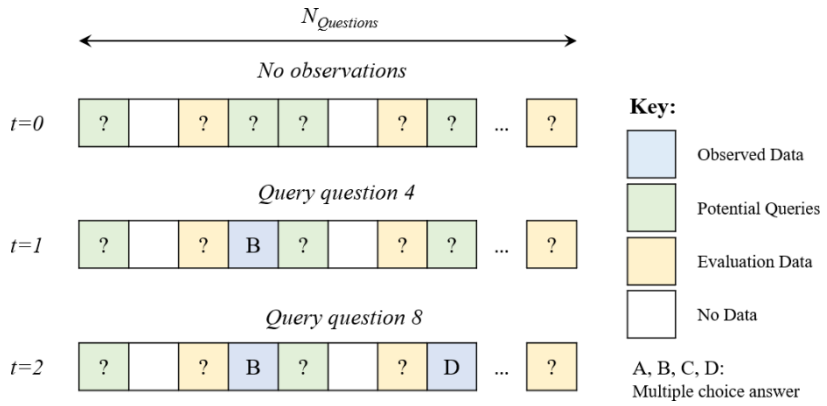

그림 6: 작업 4의 절차 그림. 각 시간 단계에서 모델은 파란색의 데이터에 대해 학습할 수 있으며, 노란색의 보류된 데이터에 대해 예측 성능이 평가됩니다. 그런 다음 알고리즘은 이 새 모델을 사용하여 녹색 질문 집합에서 쿼리할 다음 질문을 선택해야 합니다.

4 제출 프로토콜

각 과제는 공개 평가 단계와 비공개 평가 단계의 두 단계로 구성됩니다. 공개 평가 단계의 결과는 공개 순위표에 표시되어 참가자가 다른 참가자와 비교하여 자신의 제출물 성적을 확인할 수 있습니다. 비공개 평가 단계의 결과는 대회가 끝날 때까지 숨겨집니다. **중요: 각 과제에 대해 참가자는 공개 평가와 비공개 평가에 모두 제출해야 합니다. 단계를 별도로 진행해야 합니다. 공개 평가 단계에만 제출된 제출물은 대회의 최종 심사에 사용되지 않습니다. 각 과제의 비공개 단계에 제출하는 것은 참가자의 책임입니다. 은 최상의 결과를 나타냅니다.** 과제 1~3의 경우 공개 및 비공개 단계 모두에 대한 제출 템플릿 파일이 제공되며, 이를 CodaLab에 제출해야 합니다. 과제 4의 경우, 참가자 모델을 공개 및 비공개 리더보드에 모두 제출해야 하며, 이를 별도로 평가합니다.

제출하기 전에 참가자는 먼저 제출 파일을 하나의 zip 파일로 압축해야 합니다. 코다랩은 하나의 `.zip` 파일만 제출 파일로 허용합니다. 제출 버튼은 대회 웹사이트 상단의 참가 탭과 왼쪽의 제출/결과 보기 탭 아래에 있습니다. 이 페이지 상단의 탭(예: 과제 1 공개)에는 각 과제에 대한 제출 링크가 포함되어 있습니다.

다음은 각 작업에 대한 자세한 제출 지침입니다.

4.1 과제 1-3 제출

처음 세 가지 과제는 압축된 CSV 파일을 코다랩에 제출하여 평가합니다:

- **과제 1**: 참가자에게는 답이 보이지 않는 (UserId, QuestionId) 쌍이 포함된 CSV 파일(폴더 스타터 키트/제출 템플릿에 있음)이 제공됩니다. 참가자는 다음을 채워야 합니다.
  를 사용하여 학생이 각 사례에서 정답을 맞출지 예측할 수 있습니다. **중요: 제출 파일은** 제출 과제 `1.csv`라는 **파일이 포함된** `.zip` **파일이어야 합니다**.
- **과제 2**: 참가자에게는 답이 표시되지 않는 (사용자 ID, 문제 ID) 쌍이 포함된 CSV 파일(폴더 스타터 키트/제출 템플릿에 있음)이 제공됩니다. 참가자는 다음을 채워야 합니다.
  를 사용하여 학생이 각 경우에 어떤 대답을 할 것인지 예측할 수 있습니다.
  중요: 제출 파일은 다음과 같은 이름의 파일이 포함된 `.zip` **파일이어야 합니다.**
  제출 작업 `2.CSV`.
- **과제 3**: 참가자에게는 과제 3의 데이터 세트에 제공된 모든 QuestionId의 목록이 포함된 CSV 파일(폴더 시작 키트/제출 템플릿에 있음)이 제공됩니다. 참가자는
  각 문제에 부여한 순위(1-948)를 나타내는 '순위' 열을 입력해야 하며, 여기서 1은 최고 품질의 문제이고 948은 가장 낮은 품질의 문제입니다. **중요: 제출 파일은** 제출 과제 `3.csv`라는 **파일이 포함된** `.zip` **파일이어야 합니다**.

4.2 과제 4 제출

과제 4는 코드 제출을 통해 평가됩니다. 참가자에게는 템플릿 파일 제출 모델 과제 `4.py`가 제공되며, 이 템플릿 파일은 평가 스크립트가 제출된 모델과 인터페이스할 수 있는 간단한 API 래퍼를 제공합니다. 참가자가 구현해야 하는 방법은 다음과 같습니다:

- **init**: 이 파일 또는 주변 파일(예: 모델이 정의된 별도의 `model.py` 파일)에서 참가자의 모델을 로드하고 필요한 초기화를 수행합니다.
- **문제를 선택합니다**: 지금까지 모델에서 관찰한 데이터(이진 정답 표시기와 특정 객관식 답안 모두)와 각 학생에 대해 쿼리할 수 있는 문제를 나타내는 배열을 고려하여 각 문제에 대해 쿼리할 다음 문제를 선택합니다.
- **모델을 업데이트합니다**: 선택적으로 새 답을 공개한 후 공개된 새 데이터를 기반으로 모델을 업데이트합니다.

- **예측**: 예측: 학생의 답을 관찰하지 않은 모든 문제에 대해 각 학생에 대해 이진 정답률 지표를 예측합니다.

이러한 함수의 특정 서명에 대한 자세한 내용은 템플릿에서 확인할 수 있습니다. 제출 모델 작업 4.py 파일에 저장합니다. 평가 절차는 다음과 같습니다:

1. 를 사용하여 모델을 초기화합니다. `init ()`으로 모델을 초기화합니다.
2. 10단계의 경우, 먼저 `select questions()`로 새 기능을 선택하고 선택한 값을 표시한 다음 새 데이터를 `update model()`에 전달합니다.
3. 10개의 특징 선택 단계가 끝나면 예측()을 호출하여 보류된 대상 요소에 대한 이진 예측을 수행하고 모델의 예측 정확도를 평가합니다.

이 과제의 제출 파일에는 제출 모델 과제 4.py가 포함되어야 합니다. 사용자는 모델 정의 또는 학습된 모델 가중치와 같은 추가 파일을 제출물에 자유롭게 포함할 수 있습니다. 제출하려면 제출 모델 과제 4.py 템플릿 파일과 모든 추가 파일을 하나의 .zip 파일로 압축한 후 CodaLab에 제출해야 합니다. **Im-**
중요: 제출 모델 작업 4.py에서 **클래스 메서드 이름을 변경하지 마세요.** **또한 저장된 모델 파일의 경우 참가자는 제출할 파일 이름의 접두사로** 모델 과제 4를 **사용해야 합니다**. 예를 들어, 모델 이름이 `my pytorch.pt`인 경우 제출하려면 이 모델 파일의 이름을 모델 과제 4 `my pytorch.pt`로 변경해야 합니다.

평가 과정에서 과제 4의 일부로 포함된 모든 학습 데이터 및 메타데이터 파일은 제출 디렉토리의 루트에 추가되며, 제출자는 이러한 파일을 원하는 대로 사용할 수 있습니다. 로드할 데이터 세트의 이름을 지정하는 것만으로도 파일을 로드할 수 있습니다(예: 제출된 모델은 "`./train task 3 4.csv`" 경로에서 학습 데이터에 액세스할 수 있습니다). 제출 컴퓨팅 워커의 시간 제한으로 인해 평가 전 학습에 너무 많은 시간이 소요될 수 있으므로 사용자는 학습된 모델을 업로드하여 제출해야 한다는 점에 유의하세요.

각 작업에 대해 Codalab의 해당 제출 페이지에 지정된 일일 및 총 제출 한도가 있습니다. 오류로 인해 실패한 제출은 이 총합에 포함되지 않습니다. **제출물은 리더보드의 공개 및 비공개 구성 요소 모두에 개별적으로 제출해야 하며,** 최종 대회 결과는 비공개 리더보드만을 기준으로 합니다.

4.3 리더보드

제출된 결과는 대회의 각 공개 단계에 대한 공개 순위표에 표시됩니다. 공개 순위표에는 각 과제에 대한 공개 평가 데이터에서 참가자가 다른 참가자들과 어떻게 비교되는지 표시됩니다.

순위표에는 모든 과제에 대해 동일한 "점수" 열만 표시되며, 참가자는 "점수"의 의미가 과제마다 다르다는 점에 유의해야 하며, 섹션 3의 각 과제별 평가 지표에 대한 세부 사항을 참조하시기 바랍니다. 그럼에도 불구하고, 보다 자세한 결과는 대회 웹사이트의 "결과" 탭 아래 표의 가장 오른쪽 열에 있는 "상세 결과" 부분에서 확인할 수 있습니다.

대회의 모든 비공개 단계에서는 공개 순위표가 표시되지 않으며 결과는 대회 주최자만 볼 수 있습니다.

4.4 계산 환경

CodaLab의 평가는 대부분의 데이터 과학, 머신 러닝 및 딥 러닝 패키지가 포함된 기성 도커 이미지를 사용하여 수행됩니다. 참가자는 제출물이 이 환경에서 실행될 수 있는지 확인해야 합니다. 자세한 내용은 `https://github.com/ufoym/deepo` 을 참조하세요.

5 시작하기: 샘플 모델, 로컬 평가 및 제출 준비

5.1 빠른 시작

경진대회를 위한 공개 데이터와 스타터 키트는 모두 경진대회 홈페이지의 참가/데이터 받기/ 탭에서 확인할 수 있습니다. 스타터 키트에는 대회에 쉽게 참여할 수 있도록 다양한 유틸리티 스크립트와 샘플 모델이 포함되어 있습니다. 각 과제에 맞는 형식으로 제출물을 준비할 수 있는 제출 템플릿은 제출 템플릿 디렉토리에 포함되어 있습니다.

대회를 빠르게 시작하려면 다음 지침을 따르세요:

1. 대회 홈페이지에서 훈련 데이터와 스타터 키트를 다운로드하세요.
2. 다운로드한 데이터 디렉토리를 스타터 키트 디렉토리의 루트에 배치합니다.
3. (선택 사항) 로컬 모델 평가를 위한 유효성 검사 세트를 생성하기 위해 각 작업에 대해 로컬 데이터 `split.py` 파일을 실행합니다.
4. 과제 1-3에 제공된 샘플 모델 파일을 실행하여 대회용 샘플 하위 미션을 생성합니다. 과제 4는 참가자가 모델 코드를 직접 제출해야 하므로 생성할 필요가 없습니다.

과제 1-3에 솔루션을 제출하려면 참가자는 완성된 제출 템플릿 파일(제출 과제 n.csv, 여기서 n은 과제 번호)이 포함된 .zip 파일을 제출해야 합니다. 그런 다음 이 파일을 해당 과제의 공개 및 비공개 단계에 **모두** 업로드해야 합니다.

과제 4에 솔루션을 제출하려면 참가자는 학습된 모델을 실행하는 데 필요한 추가 모델 파일 또는 아티팩트와 함께 완성된 제출 API 래퍼 파일 제출 모델 과제 4.py가 포함된 .zip 파일을 제출해야 합니다.

각 작업에는 일반적으로 로컬 모델 평가를 위한 로컬 '유효성 검사 세트'를 만들기 위한 스크립트, 로컬 모델 평가를 수행하기 위한 스크립트, 작업 시작에 도움이 되는 예제 모델이 포함된 자체 포함된 디렉터리가 있습니다.

각 작업의 리소스에 대한 자세한 내용은 다음 섹션에서 확인할 수 있습니다.

5.2 작업 1

사용 가능한 파일 작업 1에 사용할 수 있는 파일은 다음과 같습니다:

- 트레이닝 데이터: 데이터/트레이닝 데이터 폴더의 트레이닝 작업 1 2.csv
- 제출 템플릿: 제출 과제 1 2.csv, 폴더에 있음
  스타터 키트/제출 템플릿
- 문제 메타데이터: 데이터/메타데이터 폴더에 있는 문제 메타데이터 작업 1 2.csv
- 학생 메타데이터: 데이터/메타데이터 폴더의 학생 메타데이터 작업 1 2.csv
- 답변 메타데이터: 데이터/메타데이터 폴더에 있는 답변 메타데이터 작업 1 2.csv
- 로컬 평가 스크립트: 로컬 데이터 split.py, 로컬 평가.py, 폴더 내
  스타터 키트/작업 1
- 샘플 기준 모델: 샘플 모델 majority.py, 시작 키트/작업 1 폴더에 있습니다.

코다랩에 제출 코다랩에 제출하려면 참가자는 제출 템플릿 파일 제출 과제 1 2.csv에 *모든* (UserId, QuestionId) 쌍에 대한 예측 결과를 포함하는 결과를 제출해야 합니다. 제공된 샘플 모델 대다수.py 스크립트를 실행하면 이 파일의 예제가 생성되며, 생성된 파일은 기본적으로 ../submissions 디렉터리에 생성되며, 생성된 파일의 이름은 제출 과제 1.csv로 지정되어 압축하여 제출할 준비가 되어 있습니다.

중요: 제출물에서 예측 열의 이름이 IsCorrect인지 확인합니다.

로컬 평가학습 데이터에서 생성된 유효성 검사 세트에 대해 로컬로 평가하려면 아래 단계를 따르세요:

1. 폴더 스타터 키트/작업 1로 이동합니다.
2. 데이터 분할. 이를 위해 로컬 데이터 split.py 스크립트를 제공했으며, 파티션은 다음을 실행하여 데이터 분할을 수행할 수 있습니다.
   파이썬 로컬 데이터 split.py
   를 호출하여 훈련 데이터를 사용하여 훈련 및 검증 세트를 생성합니다. 기본적으로 분할된 파일은 훈련 작업 1 2.csv 및 유효 작업 1 2.csv로 이름이 지정되며 데이터/테스트 입력 폴더에 저장됩니다.
3. 모델을 실행하고 예측을 해보세요. 샘플 모델을 제공했습니다.
   샘플 모델 majority.py를 사용하여 시작할 수 있습니다. 로컬 평가 데이터 분할에서 이 모델을 실행하려면 "Codalab 제출을 위한 기본 인수"로 표시된 코드 블록을 주석 처리하고 "로컬 평가를 위한 기본 인수"로 표시된 블록을 주석 처리하지 않아야 합니다. 이 모델을 사용하여 예측을 얻으려면 다음을 실행하세요.
   파이썬 샘플 모델 majority.py 이 모델은 (데이터/테스트 입력/테스트 제출 작업 1.csv)에 결과가 포함된 .csv 파일을 출력합니다.
4. 평가. 실행
   파이썬 로컬 평가.py
   를 적절한 옵션과 함께 사용할 수 있습니다(argparse 인수를 참조하세요). 이 명령은 예측 및 유효성 검사 집합을 사용하여 점수를 계산하고 저장합니다. 점수와 혼동 행렬은 기본적으로 데이터/테스트 출력인 출력 디렉터리에 저장됩니다. 이 작업에 대한 공식 평가는 로컬 평가.py의 평가 메트릭과 동일한 평가 메트릭을 구현합니다.

5.3 작업 2

사용 가능한 파일 작업 2에 사용할 수 있는 파일은 다음과 같습니다:

- 트레이닝 데이터: 데이터/트레이닝 데이터 폴더의 트레이닝 작업 1 2.csv
- 제출 템플릿: 제출 과제 1 2.csv, 폴더에 있음
  스타터 키트/제출 템플릿
- 문제 메타데이터: 데이터/메타데이터 폴더에 있는 문제 메타데이터 작업 1 2.csv
- 학생 메타데이터: 데이터/메타데이터 폴더의 학생 메타데이터 작업 1 2.csv

- 답변 메타데이터: 데이터/메타데이터 폴더에 있는 답변 메타데이터 작업 1 2.csv
- 제목 메타데이터: 데이터/메타데이터 폴더에 있는 제목 메타데이터.csv
- 로컬 평가 스크립트: 로컬 데이터 split.py, 로컬 평가.py, 폴더 내 스타터 키트/작업 2
- 기준 모델 샘플: 샘플 모델 대다수.py, 폴더 작업 2에 있습니다.

로컬 평가 및 제출 준비 과제 2는 과제 1과 성격이 매우 유사하므로 학습 데이터 및 제출 템플릿이 동일하며 로컬 평가 스크립트도 유사합니다. 유일한 차이점은 과제 2에서는 학생이 문제에 대한 실제 응답(각 객관식 문제에 대해 각각 1, 2, 3 또는 4로 인코딩된 답안 A, B, C 또는 D)을 예측한다는 점입니다.

중요: 제출 파일에 있는 예측 열의 이름이 AnswerValue.

5.4 작업 3

사용 가능한 파일 작업 3에 사용할 수 있는 파일은 다음과 같습니다:

- 트레이닝 데이터: 데이터/트레이닝 데이터 폴더의 트레이닝 작업 3 4.csv
- 제출 템플릿: 제출 작업 3.csv, 폴더 내 스타터 키트/제출 템플릿
- 질문 이미지: 데이터/이미지/ 폴더 내
- 문제 메타데이터: 데이터/메타데이터 폴더에 있는 문제 메타데이터 작업 3 4.csv
- 학생 메타데이터: 데이터/메타데이터 폴더의 학생 메타데이터 작업 3 4.csv
- 답변 메타데이터: 데이터/메타데이터 폴더에 있는 답변 메타데이터 작업 3 4.csv
- 제목 메타데이터: 데이터/메타데이터 폴더에 있는 제목 메타데이터.csv
- 샘플 기준 모델: 시작 키트/작업 3 폴더에 있는 샘플 모델 엔트로피.py

코다랩에 제출하기 이 과제는 참가자에게 훈련 데이터 훈련 과제 3 4.csv에 있는 질문의 품질 순위를 내림차순으로 매기도록 요청합니다(예: 1등급이 가장 높은 품질, 2등급이 두 번째로 높은 품질 등). 실측 데이터가 제공되지 않아 로컬 평가가 불가능하며, 대신 엔트로피 추정치를 기반으로 질문의 순위를 매기는 기준 모델(샘플 모델 entropy.py)을 제공하여 과제 3에 제출하기에 적합한 결과를 생성합니다. 계산에 대한 자세한 내용은 이 파일의 구현을 참조하세요. 제출 파일에는 2개의 열(QuestionId, 순위)이 포함되며, 두 번째 열은 각 문제의 순위입니다. 각 문제에는 고유한 순위가 있어야 합니다. 즉, 두 순위가 동일해서는 안 됩니다.

과제 3과 과제 4에 제공되는 훈련 데이터 및 메타데이터는 과제 1과 과제 2의 형식과 동일하지만, 더 작은 질문 세트를 사용합니다. 섹션 2에서 설명한 것처럼, 과제 1과 과제 2에 사용된 무작위 ID는 과제 3과 과제 4에 사용된 ID와 독립적으로 생성되므로, 참가자는 과제 1과 과제 2의 데이터를 과제 3과 과제 4에 사용하려고 시도해서는 안 됩니다.

제출물을 준비하려면 제출물 작업 3.csv 파일에 있는 모든 문제 ID의 순위를 생성하는 모델을 실행합니다. 제공된 엔트로피 기반 모델은 이 프로세스를 보여 줍니다.

파이썬 샘플 모델 random.py

그러면 예측 파일 테스트 제출 작업 3.csv가 ../submissions에 생성됩니다. 폴더에 기본적으로 저장됩니다. 이 파일을 압축하여 제출할 준비가 되었습니다.

중요: 참가자의 예상 품질 순위를 올바르게 읽으려면 순위 열의 이름을 순위로 지정해야 합니다.

5.5 작업 4

사용 가능한 파일 작업 4에 사용할 수 있는 파일은 다음과 같습니다:

- 트레이닝 데이터: 데이터/트레이닝 데이터 폴더의 트레이닝 작업 3 4.csv
- 제출 템플릿 파일: 제출 작업 4.py, 제출 템플릿 폴더에 있습니다.
- 질문 이미지: 데이터/이미지/ 폴더 내
- 문제 메타데이터: 데이터/메타데이터 폴더에 있는 문제 메타데이터 작업 3 4.csv
- 학생 메타데이터: 데이터/메타데이터 폴더의 학생 메타데이터 작업 3 4.csv
- 답변 메타데이터: 데이터/메타데이터 폴더에 있는 답변 메타데이터 작업 3 4.csv
- 제목 메타데이터: 데이터/메타데이터 폴더에 있는 제목 메타데이터.csv

- 로컬 평가 스크립트 및 모델:
  - 로컬 데이터 split.py는 스타터 키트/작업 4 폴더에 있습니다.
    - 로컬 평가.py, 훈련 모델.py, 제출 모델 작업 4.py, 모델_작업_4.pt, 스타터 키트/작업 4/pytorch_폴더에 있는 모델.py
  - 로컬 평가.py, 훈련 모델.py, 제출 모델 작업 4.py, 모델 작업 4_가장 인기 있는.npy, 모델 작업 4 num answers.npy_모델.py 폴더에 starter kit/task 4/numpy

과제 4 제출에 설명된 대로 참가자는 완성된 제출 모델 과제 4.py 파일과 함께 추가 모델 코드 또는 가중치가 포함된 .zip 파일을 비공개로 평가되는 CodaLab에 제출해야 합니다. 과제에 대한 모든 학습 데이터 및 메타데이터 파일은 제출 파일과 동일한 디렉토리(예: ./)에 포함될 것으로 가정하여 제출할 수 있습니다.

코다랩에 제출할 수 있는 샘플 제출물을 생성할 수 있는 샘플 모델과 스크립트를 제공했습니다. 샘플을 생성하려면 아래 단계를 따르세요:

1. 스타터 키트/작업 4/넘피 또는 스타터 키트/작업 4/토치로 이동합니다.
2. 실행

   파이썬 트레인_모델.py

   를 호출하면 모델 파일이 NumPy 또는 PyTorch 폴더에 저장됩니다.
3. NumPy 모델을 사용하는 경우, model.py, 제출 모델 과제 4_.py, 모델 과제 4_가장_인기_있는 과제.npy 및 모델_과제_4_num_answers.npy를 .zip 파일로 압축합니다. PyTorch 모델을 사용하는 경우, model.py, 제출물 모델 과제 4.py 및 모델_과제 4.pt를 .zip 파일로 압축합니다. 파일 이름 지정 및 제출물 템플릿 파일 제출물 모델 작업 4.py 내에서 변경할 수 있는 항목에 대한 자세한 내용은 섹션 4.2에 설명되어 있습니다.

로컬 평가 과제 4의 스타터 키트에서는 참가자가 자신의 제출물이 평가 절차에 따라 작동하는지 확인할 수 있도록 대회 API의 일부로 모의 테스트 환경이 제공됩니다. 위에서 설명한 NumPy와 PyTorch 모델 모두 이러한 방식으로 로컬에서 평가할 수 있습니다.

1. 폴더 시작 키트/작업 4로 이동합니다.
2. 데이터 분할. 이를 위해 로컬 데이터 split.py 스크립트를 제공했으며, 참가자는 다음을 실행하여 데이터 분할을 수행할 수 있습니다.

   파이썬 로컬 데이터_split.py

   를 호출하여 학습 데이터를 사용하여 학습 및 유효성 검사 집합을 생성합니다. 기본적으로

분할된 파일은 훈련 작업 4.csv 및 유효한 작업 4.csv로 이름이 지정되며 데이터/테스트 입력 폴더에 저장됩니다.

3. 모델을 훈련하세요. NumPy와 PyTorch를 모두 포함하는 샘플 모델 파일과 학습 스크립트를 제공해 드립니다. 샘플 모델을 훈련하려면 스타터 키트 또는 작업 4/넘피 또는 스타터 키트/작업 4/파이토치 및 실행

   파이썬 트레인 모델.py

   NumPy 모델에서는 train model() 메서드의 학습 데이터 경로가 새로 생성된 "로컬 테스트" 분할을 가리키도록 업데이트해야 합니다. 이 스크립트는 모델 파일을 위의 NumPy 또는 PyTorch 폴더에 저장합니다. 자세한 내용은 train model.py 스크립트를 참조하세요. 제공된 PyTorch 모델 클래스는 다음을 구현하지 않습니다.
   모델을 학습시키는 방법을 제공합니다;

4. 평가. 실행

   파이썬 로컬 평가.py

   이 명령은 선택한 문제의 순서와 최종 정답률을 계산하여 인쇄합니다.

참조

[1] Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas 등. 다중 세계 테스트 결정 서비스. *arXiv preprint arXiv:1606.03966*, 7, 2016.

[2] 제임스 베넷, 스탠 래닝 외. 넷플릭스상. *KDD 컵 및 워크샵 자료집*, 2007, 35페이지. New York, 2007.

[3] 기드온 드로르, 노암 코닉스타인, 예후다 코렌, 마르쿠스 바이머. 야후 뮤직 데이터세트와 KDD-CUP'11. *KDD 컵2011 회의록,* 3-18쪽, 2012.

[4] W. Gong, S. Tschiatschek, R. Turner, S. Nowozin, J. M. Hern´andez-Lobato, 및 C. Zhang. 쇄빙선: 베이지안 심층 잠복 가우시안 모델을 사용한 요소별 능동 정보 획득. In *Proc. 신경 정보 처리 시스템의 발전*, 2019.

[5] F 맥스웰 하퍼와 조셉 A 콘스탄. 영화 데이터 세트: 역사와 맥락. *대화형 지능형 시스템에 대한 Acm 트랜잭션(TIIS)*, 5(4):1-19, 2015.

[6] 리홍 리, 웨이 추, 존 랭포드, 로버트 E 샤파이어. 개인화된 뉴스 기사 추천을 위한 컨텍스트-밴딧 접근 방식. *제19회 월드와이드웹 국제 컨퍼런스 논문집*, 661-670페이지, 2010.

[7] J. Little, E. Frickey, and A. Fung. 객관식 질문에 답할 때 검색의 역할. *실험 심리학 저널: 학습, 기억, 인지*, 2018.

[8] 로더릭 JA 리틀과 도널드 B 루빈. 누락 된 *데이터를 사용한 통계 분석*, 볼륨 793. 존 와일리 앤 선스, 2019.

[9] C. Ma, S. Tschiatschek, K. Palla, J. M. Hern´andez-Lobato, S. Nowozin, 및 C. Zhang. EDDI: 부분 VAE를 사용한 고가치 정보의 효율적인 동적 검색. *국제 기계 학습 컨퍼런스*, 97권, 4234-4243쪽, 2019년 6월.

[10] 다니엘 J 스테크호벤과 피터 불만. 혼합형 데이터에 대한 미스포레스트-비모수적 결측치 대입. *생물정보학*, 28(1):112-118, 2012.

[11] 지차오 왕, 세바스찬 치치첵, 사이먼 우드헤드, 호세 미겔 에르난데스-로바토, 사이먼 페이튼 존스, 쳉 장. 부분 변형 자동 인코더를 사용한 대규모 교육 문제 분석. *arXiv 사전 인쇄물 arXiv:2003.05980*, 2020.

[12] C. 와일리와 D. 윌리엄. 진단 질문: 단 한 가지에 가치가 있을까요? *미국 교육 연구 협회(AERA)와 전국 교육 측정 위원회(NCME)의 연례 회의에서 발표된 논문에서 다음과 같이 발표되었습니다.*

2006년 4월 6일부터 12일까지 캘리포니아주 *샌프란시스코에서 개최되었습니다.*

[13] 윤진성, 제임스 조던, 미하엘라 반 데르 샤르. Gain: 생성적 적대 네트워크를 사용한 누락 데이터 대입. *arXiv 사전 인쇄물 arXiv:1806.02920*, 2018.