머신 러닝 연구 논문집 133:191-205, 2021 NeurIPS 2020 경진대회 및 데모 트랙

진단 질문의 결과 및 인사이트: NeurIPS 2020 교육 챌린지

왕 지차오 [*] [†] JZWANG@RICE.EDU
앵거스 램 [*] [‡] T-ANLAM@MICROSOFT.COM
에브게니 사벨리예프 [§] ES583@CAM.AC.UK
파슈미나 카메론 [‡] PASHMINA.CAMERON@MICROSOFT.COM
요르단 자이코프 [‡] YORDANZ@MICROSOFT.COM
호세 미구엘 에르난데스-로바토 [§] JMH233@CAM.AC.UK
리처드 E. 터너 [§] RET26@CAM.AC.UK
리차드 G. 바라니우크 [†] RICHB@RICE.EDU
크레이그 바튼 [¶] CRAIG.BARTON@EEDI.CO.UK
사이먼 페이튼 존스 [‡] SIMONPJ@MICROSOFT.COM
사이먼 우드헤드 [¶] SIMON.WOODHEAD@EEDI.CO.UK
청 장 [‡] CHENG.ZHANG@MICROSOFT.COM

편집자: 편집자: 휴고 자이르 에스칼란테, 카티아 호프만

초록

이 대회는 오해를 불러일으키는 교육적 효과가 있는 객관식(MCQ) 문제인 교육 *진단 문제*에관한 것입니다. 이러한 질문의 수가 계속 증가함에 따라 교사들은 어떤 질문이 학생들에게 가장 적합한 질문인지 파악하는 것이 부담스러워지고 있습니다. 따라서 우리는 다음과 같은 질문에 답하고자 합니다. 수억 개의 MCQ 답변에 대한 데이터를 사용하여 수동 개인화가 불가능한 대규모 학습 시나리오에서 자동 개인화 학습을 추진할 수 있는 방법은 무엇일까요? MCQ 데이터를 대규모로 성공적으로 활용하면 보다 지능적이고 개인화된 학습 플랫폼을 구축하여 궁극적으로 교육의 질을 전반적으로 향상시킬 수 있습니다. 이를 위해 새로운 대규모 실제 세계 데이터 세트를 도입하고 실제 학습 시나리오를 모방하고 위의 질문의 다양한 측면을 대상으로 하는 MCQ에 대한 4가지 데이터 마이닝 과제를 공식화하여 NeurIPS 2020에서 경연을 펼칩니다. 약 400개 팀이 약 4,000개의 과제를 제출했으며, 각 과제에 대한 다양하고 효과적인 접근 방식이 매우 고무적이었던 NeurIPS 대회에 대해 보고합니다.

키워드

개인화 교육, 진단 질문, 질문 분석, 비지도 학습, 매트릭스 완성, 결측치 예측, 능동 학습

[*] 동등한 기여.
[†] 라이스 대학교
[‡] 마이크로소프트 리서치 캠브리지
[§] 케임브리지 대학교
[¶] Eedi
사이먼 우드헤드와 쳉 장의 답변입니다.

1. 소개

최근 몇 년 동안 전문적으로 제작된 자료와 지침을 저렴한 비용으로 제공하는 대규모 온라인 학습 플랫폼이 점점 더 확산되고 있습니다. 이러한 플랫폼은 전문 학습 리소스에 대한 접근성을 낮추고 대중에게 고품질 학습 경험을 제공함으로써 현재의 교육 관행을 혁신하고 있습니다. 이러한 플랫폼의 핵심 기술 중 하나는 학생의 배경, 관심사, 학습 목표를 고려하여 각 학생에게 자동으로 지침과 교육 활동을 제공하는 개인화 학습 알고리즘입니다. 그러나 *개인화는* 이러한 학습 플랫폼의 핵심 과제로 남아 있으며 여전히 활발한 연구 분야입니다. 학생 개개인이 모두 독특하고 다르기 때문에 각 학생에게 가장 적합한 개별화된 학습 경로가 필요하기 때문입니다. 교사는 학생 한 명 한 명에게 주의를 기울이면서 자연스럽게 학생의 필요에 맞게 교육 방식을 조정하지만, 알고리즘은 전문 교사에 비해 적응력이 떨어집니다.

그렇다면 가장 중요한 질문은 온라인 학습 플랫폼을 개인화하여 특정 학생의 필요에 맞게 조정하는 방법입니다. 이 질문은 너무 크고 모호하기 때문에 이 백서에서는 객관식 *진단 문제의* 선택지를 개인화하는 작은 하위 문제에 초점을 맞춥니다. 이 초점을 신중하게 선택했습니다. 첫째, 잘 만들어진 객관식 문항이 교육적으로 효과적(**?**)이라는 연구 결과가 많이 있습니다. 둘째, 객관식 문항은 학생, 문항, 학생들이 일부 문항에 대한 답변 등 매우 체계적인 형태로 방대한 데이터를 쉽게 수집할 수 있습니다. 셋째, 수십만 명의 사용자를 보유한 실제 배포된 플랫폼(Eedi)과의 적극적인 파트너십을 통해 오늘날의 데이터에 굶주린 머신러닝 알고리즘에 공급할 *많은* 데이터를 수집할 수 있었기 때문입니다.

이 백서에서는 이러한 객관식 진단 문제에 대한 대규모 답변 데이터 세트를 제공하고 참가자들이 여러 가지 과제에 참여하도록 유도한 NeurIPS 대회에 대해 설명합니다. 이 모든 과제는 학습 여정의 특정 시점에서 특정 학생에게 가장 적합한 질문을 식별하는 궁극적인 목표를 향해 진행되었습니다. 좀 더 구체적으로 살펴보면 다음과 같습니다:

- 저희는 교육 분야에서 머신러닝을 적용할 수 있는 비옥한 응용 분야로 판단되는 객관식 *진단 문제에* 대한 경진대회를 NeurIPS 2020에서 개최했습니다. 객관식 진단 문제는 교육적으로 타당할 뿐만 아니라 다양한 머신 러닝 방법에 매우 적합한 형식입니다.
- 현재 공개적으로 이용 가능한 진단 질문에 대한 답변이 담긴 방대한 데이터 세트를 소개합니다(섹션 2).[1] 이는 교육 분야에서 가장 큰 데이터 세트 중 하나입니다. 비교는 표 1을 참조하세요. 또한 이 데이터 세트는 경쟁사의 범위를 넘어 교육 및 머신 러닝 연구를 발전시키는 데 도움이 될 수 있는 잠재력을 가지고 있습니다(섹션 6).
- 진단 문제와 관련된 문제를 해결하는 것을 목표로 하는 네 가지 대회 과제를 소개합니다. 데이터셋을 통해 학생들의 정답을 정확하게 예측하는 것을 목표로 하는 두 가지 일반적인 교육 데이터 마이닝 과제가 포함되어 있습니다(섹션 3.1 및 3.2). 더 중요한 것은 자동 문제 품질 평가에 관한 새로운 작업(섹션 3.3)과 개인화된 문제 선택에 관한 작업도 소개한다는 점입니다.

1. https://eedi.com/projects/neurips-education-challenge

표 1: 새로운 데이터 세트와 기존 데이터 세트의 비교.

데이터 세트	#학생	#질문	#답변 기록
Algebra2005[2]	569	173,133	607,000
Bridge2006[2]	1,135	129,263	1,817,427
ASSISTments2009[3]	4,151	16,891	325,637
통계2011[4]	333	-	189,297
ASSISTments2012[3]	24,750	52,976	2,692,889
ASSISTments2015[3]	19,840	-	683,801
ASSISTments2017[5]	1,709	3,162	942,816
NAEP2019[6]	1,232	21	438,291
우리	**118,971**	**27,613**	**15,867,850**

(섹션 3.4). 이번 대회에서는 처음 두 가지 과제에 집중했던 이전 대회(**??**)에 비해 더 넓은 관점을 반영하여 과제를 설계했습니다.

- 약 400개 팀이 참가하여 총 4,000개에 가까운 솔루션을 제출한 NeurIPS 경진대회 결과를 보고합니다. 네 가지 과제 각각에 대한 주요 솔루션의 주요 인사이트를 요약하고(섹션 4), 교육 분야에서 머신 러닝의 미래 역할에 대한 경쟁, 데이터 세트 및 제출된 솔루션의 잠재적 영향에 대한 논의도 함께 다룹니다.

2. 데이터 세트

현재 수만 개의 학교에서 사용되고 있는 온라인 교육 플랫폼인 Eedi의 새로운 대규모 실제 데이터 세트를 큐레이션하여 2018년 9월부터 2020년 5월까지 수집된 객관식 진단 문제에 대한 학생들의 응답을 자세히 설명합니다. 이 플랫폼은 초등학생부터 고등학생(대략 7세에서 18세 사이)을 대상으로 크라우드 소싱된 진단 문제를 제공합니다. 각 진단 문제는 객관식 문항으로 정답은 네 가지이며, 그 중 정확히 한 가지가 정답입니다. 현재 이 플랫폼은 주로 수학 문제에 중점을 두고 있습니다. 그림 1은 플랫폼의 문제 예시를 보여줍니다. 대회는 4개의 과제로 나뉘는데, 과제 1과 2는 과제 3과 4와 마찬가지로 하나의 데이터 세트를 공유합니다. 이러한 데이터 세트는 형식은 대체로 동일하지만 서로 다른 문제 세트를 사용합니다. 모든 QuestionId, UserId 및 AnswerId는 익명 처리되었으며 제품에서 찾을 수 있는 것과는 식별할 수 없습니다. 작업 1과 2의 모든 ID는 작업 3과 4의 ID와 별도로 익명 처리되며 이 두 데이터 세트에 사용된 질문은 서로 다른 것입니다. 이는 두 데이터 세트의 익명성을 보장하기 위한 설계입니다.

가 포함되어 있습니다.

2. http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp
3. https://sites.google.com/site/assistmentsdata/home
4. https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=507
5. https://sites.google.com/view/assistmentsdatamining
6. https://sites.google.com/view/dataminingcompetition2019/dataset

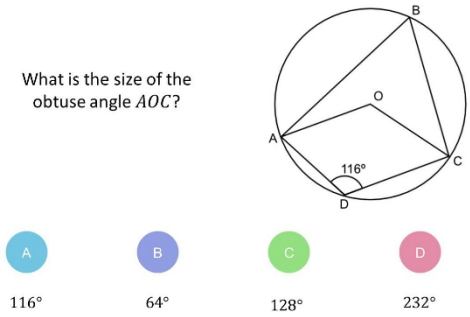

그림 1: 데이터 집합의 질문 예시.

2.1. 기본 데이터

과제에 대한 기본 훈련 데이터는 학생들이 객관식 진단 문제에 다시 답한 기록으로 구성됩니다. 표 2는 데이터 세트에 포함된 일부 데이터 포인트의 형식을 보여줍니다. 각 학생은 일반적으로 가능한 모든 문제 중 극히 일부에만 답하기 때문에 일부 학생과 문제는 너무 적은 수의 답안 기록과 연관되어 있습니다. 따라서 과제 1과 2의 경우 50개 미만의 답안을 받은 문제와 50개 미만의 문제에 답한 학생을 제거합니다. 마찬가지로, 고정된 문제 집합에 관심이 있는 과제 3 및 4의 경우 이러한 문제 중 50개 미만에 답한 모든 학생을 제거합니다. 학생이 동일한 문제에 대해 여러 개의 답안을 제출한 경우 가장 최근의 답안 기록을 사용합니다. 데이터는 각 행이 학생을 나타내고 각 열이 문제를 나타내는 행렬 형태로 변환할 수 있습니다. 그림 2는 이러한 데이터 세트의 표현을 보여줍니다.

작업 1과 2의 경우, 답변 레코드를 80%/10%/10% 훈련/공개 테스트/개인 테스트 세트로 무작위로 분할합니다. 마찬가지로, 작업 3과 4의 경우 사용자 아이디를 80%/10%/10% 훈련/공개 테스트/개인 테스트 세트로 무작위로 분할합니다. 이러한 전처리 단계를 통해 다음과 같은 크기의 훈련 데이터 세트가 생성됩니다:

- 과제 1 및 2: 27,613개의 문제, 11,8971명의 학생, 15,867,850개의 답안
- 과제 3 및 4: 948개 문제, 4,918명의 학생, 1,382,727개의 답변

이러한 훈련 세트의 총 답변 레코드 수는 1,700만 개를 초과하여 수동 분석이 비현실적이며 자동화된 데이터 중심 접근 방식이 필요합니다.

2.2. 질문 메타데이터

데이터 세트의 각 질문에 대해 다음과 같은 메타데이터를 제공합니다.

SubjectId 각 문제에 대해 해당 문제와 관련된 과목 목록을 제공합니다. 각 과목은 다양한 수준의 세분화된 수학 영역을 다룹니다. "대수", "데이터 및 통계", "기하와 측정" 등이 예시 과목입니다. 이러한 과목은 트리 구조로 배열되어 있습니다. 즉, "인수분해" 과목은 "단일 대괄호로 인수분해"의 상위 과목입니다.

표 2: 데이터 세트의 답안 레코드 예시. 각 행은 하나의 레코드를 나타냅니다.

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

문제 내용. 과제 3과 과제 4에서는 주제 외에도 그림 1에 표시된 것처럼 각 문제와 관련된 이미지도 제공합니다. 각 이미지에는 문구, 그림 및 표를 포함한 문제 세부 정보가 포함되어 있습니다.

2.3. 학생 메타데이터

데이터 세트의 각 학생에 대해 다음과 같은 메타데이터를 제공합니다.

UserId. 학생을 고유하게 식별하는 ID입니다.

성별. 학생의 성별입니다. 0은 지정되지 않음, 1은 여성, 2는 남성, 3은 기타입니다.

DateOfBirth. 학생의 생년월일(월 1일로 반올림)입니다.

PremiumPupil. 학생의 재정적 필요 상태, 즉 값이 1이면 학생이 무료 학교 급식 또는 학생 프리미엄을 받을 자격이 있음을 나타냅니다.

2.4. 답변 메타데이터

데이터 세트의 각 답변 레코드에 대해 다음과 같은 메타데이터를 제공합니다.

UserId. 답변 레코드를 고유하게 식별하는 ID입니다.

날짜응답. 학생이 질문에 답변한 시간 및 날짜입니다.

자신감. 백분율 신뢰도 점수는 학생이 질문에 답할 때 부여하는 점수입니다. 0은 무작위 추측을 의미하고 100은 완전한 자신감을 의미합니다.

GroupId. 학생이 문제를 할당받은 클래스(학생 그룹)입니다.

QuizId. 학생이 답한 질문이 포함된 할당된 퀴즈입니다.

작업 계획 ID. 학생에게 문제가 할당된 작업 계획입니다. 작업 계획은 퀴즈를 포함하는 일련의 주제입니다. 학습 계획은 일반적으로 한 학년도 동안 지속됩니다.

3. 경쟁 과제

이 섹션에서는 네 가지 경쟁 과제와 평가 지표에 대해 설명합니다. 처음 두 과제는 데이터 세트의 모든 질문에 대한 학생의 응답을 예측하는 것을 목표로 합니다. 이 두 과제는 추천 시스템 과제(**???**) 또는 결측치 대입 과제(????)와 같이 여러 가지 방식으로 공식화할 수 있습니다. 세 번째 과제는 교육에서 필수적이며 여전히 미해결 과제로 남아 있는 문제 품질을 평가하는 데 중점을 둡니다.

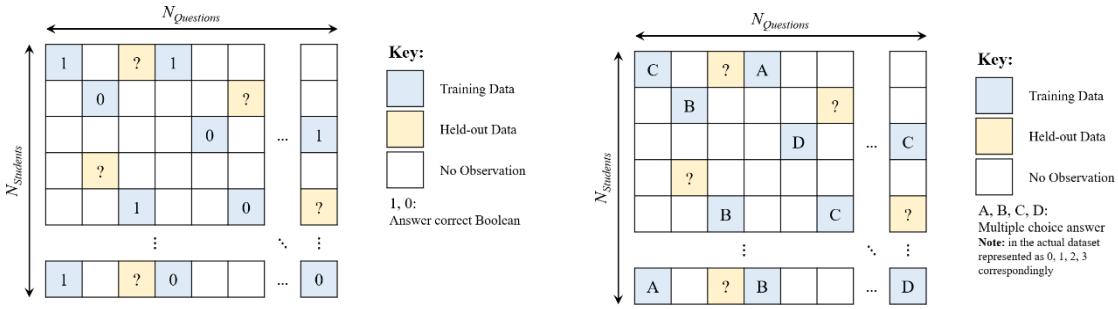

그림 2: 작업 1(**왼쪽**)과 작업 2(**오른쪽**)에 대한 데이터의 스파스 행렬 표현 그림.

도메인(**?**). 네 번째 과제는 개인화된 동적 의사결정(**??**)이 필요한 개인화 교육이라는 과제를 직접적으로 다루고 있습니다.

데이터셋으로 해결할 수 있는 실제 교육 데이터 마이닝 문제의 폭과 다양성을 반영하기 위해 대회를 네 가지 과제로 구성했습니다. 특히 앞의 두 과제는 기존 교육 데이터 마이닝 경진대회에서 흔히 볼 수 있는 학생의 성적 예측이라는 공통점이 있지만, 이전 대회보다 훨씬 더 크고 풍부한 데이터 세트를 활용했습니다(표 1 참조). 후자의 두 과제는 참신하고 혁신적인 접근 방식과 솔루션을 필요로 합니다. 네 가지 과제 모두 의미 있는 실제 교육 문제를 반영합니다. 전반적으로, 기존의 교육 데이터 마이닝 대회에 이미 익숙하지만 새로운 대규모 데이터 세트로 작업하고자 하는 참가자와 교육 환경에서 새로운 머신 러닝 문제를 해결하고자 하는 참가자에게 어필할 수 있는 과제 설정입니다.

3.1. 과제 1: 학생의 반응 예측 - 맞거나 틀림

첫 번째 작업은 학생이 질문에 올바르게 답했는지 예측하는 것입니다. 이 작업에 사용되는 기본 데이터는 학생이 문제를 올바르게 답했는지 여부를 나타내는 이진 지표인 마지막 열이 있는 레코드 테이블(StudentId, QuestionId, IsCorrect)입니다. 이 데이터의 스파스 행렬 표현은 그림 2에 나와 있습니다. 구체적으로, 각 학생에 대해 사용 가능한 기록의 일부가 평가가 수행된 숨겨진 시험 집합으로 유지되었습니다.

평가 지표. 예측 정확도, 즉 실제 정확도 지표와 일치하는 예측의 수를 총 사전 받아쓰기 수(보류된 테스트 세트에서)로 나눈 값을 지표로 사용합니다:

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

중요성. 정답이 없는(또는 새로 도입된) 질문에 대한 학생의 정답을 예측하는 것은 실제 개인별 맞춤 교육 플랫폼에서 학생의 기술 수준을 추정하는 데 매우 중요하며 고급 작업의 기초를 형성합니다. 이 작업은 행렬 완성 클래스에 속하며 이진 데이터의 경우 추천 시스템 영역에서 흔히 볼 수 있는 문제를 연상시킵니다.

3.2. 과제 2: 학생 반응 예측 - 답안 예측:

두 번째 작업은 학생이 특정 문제에 대해 어떤 답을 답할지 예측하는 것입니다. 이 작업에 사용되는 기본 데이터는 레코드 테이블(StudentId, QuestionId, AnswerValue, CorrectAnswer)이며, 마지막 두 열은 [1, 2, 3, 4]의 값을 갖는 범주형입니다(각각 객관식 답안 옵션 A, B, C 및 D와 상관관계가 있음). 스파스 행렬 표현은 그림 2(오른쪽)에 나와 있습니다. 데이터 세트의 질문은 모두 객관식이며, 각각 4개의 잠재적 선택지와 1개의 정답이 있기 때문에 이 작업을 행렬 완성 공식의 다중 클래스 예측 문제로 취급합니다. 이 문제 공식은 과제 1의 문제와 유사하지만, 학생들의 정답/오답 지표와 같은 이진 데이터 대신 정렬되지 않은 범주형 데이터(즉, 학생들의 실제 선택지)를 사용합니다. 일반적으로 응답이 이진 또는 서수(예: 별 1~5개)인 추천 시스템 영역에서는 이러한 정렬되지 않은 범주형 데이터가 드물다는 점에 유의하십시오.

평가 메트릭. 이제 정답이 이진형이 아닌 범주형이라는 점을 제외하고는 과제 1과 동일한 지표 **예측 정확도를** 사용합니다.

중요도. 학생의 답에 대한 실제 객관식 옵션을 예측하면 학생이 한 주제에 대해 가질 수 있는 일반적인 오해를 분석할 수 있습니다. 예를 들어, 상관관계가 높은 문제-답변 쌍의 클러스터는 동일하거나 관련된 오해에 해당할 수 있습니다. 오해 간의 관계를 이해하는 것은 커리큘럼 개발을 위해 해결해야 할 중요한 문제이며, 이는 주제를 가르치는 방식과 주제 순서를 결정할 때 중요한 정보를 제공할 수 있습니다.

3.3. 과제 3: 글로벌 문제 품질 평가

세 번째 과제는 학생의 답안 기록에서 학습한 정보를 바탕으로 도메인 전문가 패널(숙련된 교사)이 정의한 문제의 '품질'을 예측하는 것입니다. 전문가 교사가 문제의 질을 어떻게 판단하는지는 알 수 없고 정량화하기 어렵기 때문에 이 과제는 전문가의 문제 질 판단을 모방한 문제 질 평가 지표를 정의하고 개발해야 합니다. 이 과제는 문제 품질에 대한 명시적인 감독 레이블이 없으므로 비지도 학습 과제로 볼 수 있습니다.

평가 지표. 제안된 품질 지표를 사용하여 자동으로 계산된 문제 품질 순위와 전문가 순위 간의 일치도를 평가합니다. 평가용 데이터를 수집하기 위해 5명의 서로 다른 전문가 평가자로부터 질문의 하위 집합에 대한 쌍별 질문 품질 순위를 수동으로 수집합니다. 데이터 수집 프로세스에서 사용되는 프롬프트의 예는 그림 3에 나와 있습니다. 또한, 도메인 전문가 중 한 명인 크레이그 바튼이 품질 순위를 결정하는 몇 가지 질문 설계의 "황금률"을 확인했습니다,[7] 품질 순위를 결정하는 데 도움이 됩니다. 구체적으로, 양질의 질문은 다음과 같아야 합니다.

- 명확하고 모호하지 않아야 합니다;
- 단일 기술/개념을 테스트합니다;
- 학생들이 10초 이내에 답할 수 있도록 합니다;

7. https://medium.com/eedi/what-makes-a-good-diagnostic-question-b760a65e0320

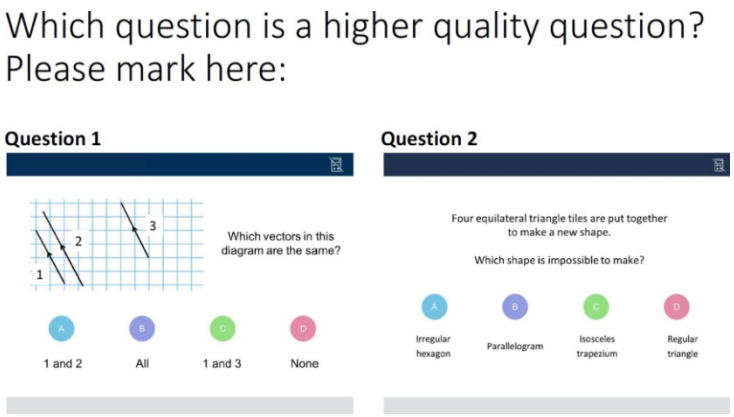

그림 3: 쌍별 상대 평가 문제 품질에 대한 전문가의 판단을 수집하는 데 사용되는 프롬프트의 예입니다. 이 외에도 전문가에게는 다음과 같은 지침이 제공됩니다: *다음 각 슬라이드에는 왼쪽과 오른쪽에 각각 2 개의 질문이 있습니다. Please*

동점자는 허용되지 않으며, 어떤 문제가 더 높은 점수를 받을지 결정합니다.

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 4: 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 계산의 그림(자세한 내용은 섹션 3.3 참조). 이 예에서는 5개의 문제 쌍과 3명의 전문가를 사용합니다.

- 오답을 쉽게 식별할 수 있도록 답안을 신중하게 설계했습니다;
- 질문에서 확인하려는 주요 오해가 있는 학생에게는 답하기가 어려울 수 있습니다.

총 40개의 질문 쌍을 무작위로 선택한 후 각 평가자에게 각 쌍에 대한 이원적 순위, 즉 쌍 중 어느 질문의 품질이 더 높은지 개별적으로 평가해 달라고 요청합니다. 평가 단계는 다음과 같습니다(그림 4 참조). 먼저, 질문 품질 메트릭이 모든 질문에 대한 전체 순위를 내림차순으로 계산합니다. 둘째, 이 전체 순위에서 40쌍의 질문에 대한 쌍별 순위를 추출합니다. 셋째, 각 전문가 i*에* 대해 동의 비율을 결정합니다: $_i = \frac{N \text{ 일치하는 쌍 수}}{N \text{ 총 쌍}}$. 마지막으로 다음을 찾습니다.

이러한 합의 분수의 *최대값* A_{max} = 최대$_i$ A_i . 이 A_{max} 점수는 이 작업에 대한 최종 평가 지표로 사용됩니다.

저희는 모든 전문가의 판단에 가장 근접할 수 있는 지표를 찾고 있었기 때문에 모든 전문가에 대한 합의율의 평균이 아니라 합의율의 최대값을 사용했습니다. 이러한 접근 방식을 채택한 이유는 전문가의 품질 지표 자체가 주관적이기 때문이며, 머신러닝을 통해 특정 전문가의 접근 방식이 특히 잘 근사화될 수 있는지 알아내는 것이 흥미롭기 때문입니다.

과제 3의 중요성. 질문 품질은 종종 주관적인 척도로 간주되기 때문에, 즉 사람마다 질문 품질에 대한 정의가 다를 수 있기 때문에 신뢰할 수 있는 질문 품질 측정을 생성하는 확장 가능한 증거 기반 메커니즘은 아직 해결되지 않은 과제로 남아 있습니다. 따라서 하나의 통일된 객관적인 지표를 만드는 것은 어려운 일입니다. 또한 전문가조차도 언뜻 보기에는 좋은 것 같지만 품질이 좋지 않은 질문을 작성하는 경우가 있습니다. 질문이 크라우드 소싱되는 상황에서는 질문의 양은 많지만 품질이 크게 다를 가능성이 높기 때문에 수동 검사가 어렵고 고품질 질문을 식별하기 위한 자동화된 기술이 필요합니다. 또한 자동화된 문제 품질 판단은 교사가 더 높은 품질의 문제를 작성하는 데 도움이 되는 유용한 가이드가 될 수 있습니다.

3.4. 작업 4: 맞춤 질문

네 번째 과제는 학생의 나머지 질문에 대한 답변에 대한 예측 정확도를 최대화하기 위해 학생에게 물어볼 일련의 질문을 대화식으로 생성하는 것입니다. 이 과제 설정은 능동 학습 및 베이지안 실험 설계와 유사합니다. 구체적으로, 참가자는 선택한 문제가 0개인 상태에서 시작합니다. 그런 다음, 참가자는 시험 세트의 각 학생에 대해 첫 번째 문제를 선택하고 이에 대한 학생의 답을 관찰한 다음, 이 관찰을 사용하여 나머지 문제에 대한 학생의 답을 예측합니다. 이 예측을 바탕으로 참가자의 방법은 미리 지정된 선택 문제 수에 도달할 때까지 두 번째 문제를 선택하는 등의 방식으로 진행됩니다. 참가자는 i) 문제 순서를 선택하고 ii) 나머지 문제에 대한 학생의 답을 예측해야 하는 모델을 제출해야 합니다. 그림 5는 이 작업의 설정을 보여줍니다.

평가 지표. 제출된 모델은 보류(시험) 학생 집합의 모든 학생에 대해 10개의 쿼리 질문을 순차적으로 선택하도록 요청받았습니다. 각 선택 단계가 끝나면 이러한 학생-질문 쌍에 대한 범주형 답안과 이진 정답 지표가 모델에 비공개로 공개되었습니다. 그런 다음 모델이 이 새로운 데이터를 통합하거나 각 질문 후에 재학습할 수 있도록 했습니다. 각 학생에 대해 10개의 답을 받은 후, 답을 조회할 수 없는 각 학생의 보류된(테스트) 답 집합에 대한 이진 정답을 예측하는 모델의 예측 정확도를 평가했습니다.

과제 4의 중요성. 이 과제는 가능한 최소한의 질문을 던지면서 학생의 기술 수준과 다양한 개념에 대한 이해도를 정확하게 진단하고 평가하고자 하는 적응형 시험에 있어 매우 중요합니다. 이를 달성하면 시험 시간을 줄이면서도 시험 결과의 신뢰성을 유지하여 기존 시험에 혁명을 일으킬 수 있습니다. 예를 들어, 듀오링고는 최근 영어 시험에 적응형 시험 기술을 적용하여 토플 시험에 일반적으로 4시간이 소요되는 시험 시간을 40분으로 단축했습니다. [8]. 이 작업이 더 많은 영감을 주길 바랍니다.

8. https://englishtest.duolingo.com/

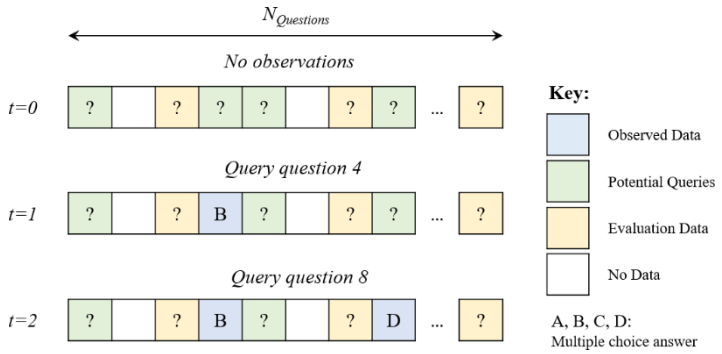

그림 5: 작업 4의 절차 그림. 각 시간 단계에서 모델은 파란색 데이터(관찰 데이터)로 학습할 수 있으며, 노란색의 보류된 데이터(평가 데이터)로 예측 성능을 평가합니다. 그런 다음 알고리즘은 이 새 모델을 사용하여 녹색 질문 집합(잠재적 쿼리)에서 쿼리할 다음 질문을 선택해야 합니다.

혁신을 통해 적응형 테스트를 더욱 개선했습니다. 이 작업은 머신러닝 모델의 불확실성에 대해 효과적으로 추론하고 데이터를 효율적으로 사용해야 하는 중요한 머신러닝 과제이기도 합니다.

4. 대회 결과

2020년 7월부터 2020년 10월까지 2개월 동안 진행된 이번 공모전에는 전 세계 382개 팀에서 총 3,696개의 과제가 제출되었습니다. 네 가지 과제 모두에서 시간이 지남에 따라 유의미한 개선이 이루어졌습니다. 표 3은 각 과제별 참여 및 제출 내역을 보여줍니다. 다음 섹션에서는 각 과제별 최우수 팀의 솔루션에 대해 간략하게 설명합니다. 그런 다음 대회를 통해 얻은 인사이트와 결과를 요약합니다.

4.1. 수상 솔루션

과제 1의 최우수 솔루션. 우승 솔루션(?)은 77.29%의 예측 정확도를 달성한 ㈜에이데미(Aidemy)의 솔루션입니다. 이 솔루션은 유용한 특징 인코딩, 특징 선택 및 앙상블 방법을 활용합니다. 이들이 사용하는 주요 기법은 특징의 세그먼트를 기반으로 목표 통계(이 경우 학생의 이진 정답/오답 기록)를 계산하는 *목표 인코딩,* 즉 학생의 인구 통계에 따라 데이터를 분할하고 각 하위 집합에 대한 주요 통계를 계산하는 기법을 포함합니다. 저자들은 많은 솔루션에서 이 기법을 사용하며 특히 이 작업에서 그 효과를 입증했습니다.

또한 다양한 *메타데이터에서* 여러 가지 기능을 구성하는데, 그 중 세 가지 기능이 예측 성능을 개선하는 데 특히 유용하다고 합니다. 첫 번째 유형의 기능은 학생의 답변 타임스탬프 및 학생의 생년월일(나이)과 같은 *시간 관련 기능입니다.* 두 번째 유형의 기능은 답변한 질문 및 수강한 수업 수와 같은 학생의 현재까지의 경험을 포함하는 *사용자 기록 기능입니다.* 세 번째 유형의 기능은 문제 과목을 포함한 *과목* 기능입니다. 학생-문제 답안 행렬에서 특이값 분해(SVD)를 수행하여 얻은 벡터를 통해 학생과 문제 간의 중요한 상호 작용을 파악할 수 있는 몇 가지 유용한 기능이 추가로 제공됩니다.

표 3: 각 작업의 활성 팀 수 및 총 제출 건수.

작업	작업 1	작업 2	작업 3	작업 4
#팀	80	33	16	17
#제출	1401	448	1061	786

패턴. 이러한 기능을 예측 모델의 입력으로 사용하는 것 외에도 타깃 인코딩과 함께 사용하여 학생들을 연령대별로 그룹화하거나 주제에 따라 질문을 그룹화하는 데 사용하기도 했습니다.

예측 파이프라인은 세 단계를 거칩니다. 첫 번째 *특징 선택* 단계에서는 모델 앙상블을 사용하여 각각 상위 100개의 특징을 선택합니다. 두 번째 *메타 특징 구성* 단계에서는 또 다른 모델 앙상블을 사용하고, 선택한 특징에 대해 학습한 후 메타 특징을 생성합니다. 마지막 단계에서는 메타 피처를 입력으로 사용하여 다양한 회귀 모델을 다시 앙상블 방식으로 훈련하여 최종 예측을 생성합니다.

과제 2의 최고 솔루션. 우승 솔루션(**?**)은 68.03%의 예측 정확도를 달성한 중국과학기술대학 팀이 내놓은 솔루션입니다. 이 팀은 질문에 답할 때 학생들의 *집중력을* 파악하는 아이디어에서 착안한 새로운 *순서 인식 인지 진단 모델*(OCD)을 솔루션으로 제안했습니다. 이들은 컨볼루션 신경망을 구현하고 슬라이딩 컨볼루션 창을 사용하여 학생들의 집중력을 포착합니다. 구체적으로, 먼저 학생의 답안 기록을 $\{(q_i, c_i)\}_i$ 의 튜플 목록으로 구성합니다. 여기서 q_i 와 c_i 는 각각 질문과 학생의 임베딩입니다. 둘째, 한 학생의 모든 기록을 하나의 벡터로 연결합니다. 마지막으로, 스택형 1차원 슬라이딩 컨볼루션 신경망(CNN)을 적용하여 각 학생 기록을 특징으로 변환합니다. 이 구현은 학생의 답안 기록이 답안 타임스탬프 순서대로 정렬되기 때문에 '순서 인식'이 가능합니다. 또한 다양한 슬라이딩 창 크기를 가진 CNN 앙상블을 사용하여 다양한 규모의 주의 집중 시간을 모델링합니다. 실험을 통해 단일 주의 집중 시간만 모델링할 때와 비교하여 다중 주의 집중 시간 모델링의 이점을 경험적으로 입증했습니다.

과제 3의 상위 솔루션들. 우승 솔루션은 세 개의 다른 팀(**??**)에서 나왔는데, 모두 인간 평가자의 판단과 최대 80%의 일치도를 달성했습니다. 이는 수상 기준이 평가자 중 누구와도 최대로 일치하는 것을 기준으로 삼았기 때문입니다. 표 4에는 전체 결과, 즉 각 평가자와 각 우승 팀의 제출물 간의 일치도가 표시되어 있습니다. 아래에서는 우승한 세 가지 솔루션 각각에 대해 간략하게 설명합니다.

의 팀(**?**)은 양질의 문제는 적절히 어렵고, 가독성이 좋으며, 답안 선택의 균형을 이룬다는 가설에 기반한 솔루션을 제시합니다. 그들은 가설로 설정된 각 속성에 대한 문제 특징을 계산합니다. 특히 엔트로피를 사용하여 문제의 난이도와 답안 선택 간의 균형을 측정합니다. 또한 문제 이미지에 광학 문자 인식(OCR)을 수행하여 가독성 특징을 추출하여 문제 텍스트 정보를 추출합니다. 실험 결과에 따르면 위의 모든 기능을 결합하는 것이 일부 기능만 사용하는 것보다 공개 및 비공개 데이터 모두에서 가장 우수한 성능을 발휘하는 것으로 나타났습니다. 문제 품질 순위를 계산하기 위해 먼저 각 기능을 사용하여 문제의 순위를 매긴 다음 평균 순위에 따라 순위를 다시 매깁니다.

표 4: 각 평가자와의 합의를 포함한 과제 3의 수상 솔루션에 대한 자세한 결과.

팀	평가 #1	평가 #2	평가 #3	평가 #4	평가 #5
Aidemy, Inc	0.64	0.6	0.6	0.6	0.8
TAL	0.72	0.68	0.6	0.6	0.8
U. 시드니	0.72	0.6	0.6	0.6	0.8

TAL 교육 그룹(**?**)의 팀은 위의 솔루션과 어느 정도 유사한 솔루션을 제시합니다. 구체적으로는 정답 간의 엔트로피와 정답과 오답 간의 엔트로피도 계산합니다. 또한 문제 그룹과 퀴즈 그룹에 따라 세분화된 정답-오답 엔트로피를 추가로 계산하여 다양한 지식 습득 수준에 따른 문제 난이도를 파악합니다. 또한 학생의 정답 신뢰도 점수를 지표에 포함합니다. 위의 솔루션과는 달리, 이 솔루션은 최종 순위를 계산하기 위해 특징의 가중 평균을 취하여 지표를 구성합니다.

시드니 대학교(**?**)의 연구팀은 학생들이 각 질문에 대한 답변에 대해 보고한 평균 신뢰도 순으로 순위를 매기고, 평균 신뢰도가 높을수록 문제의 질이 높은 것으로 해석하는 훨씬 더 간단한 해결책을 제시했습니다. 저자들은 학생의 신뢰도가 높다는 것은 문제가 명확하고 모호하지 않다는 것을 의미하며, 질문이 이러한 오해를 명확하게 해결하면 주요 오해를 가진 학생이 오답에 대해 높은 신뢰도를 보고하는 더닝-크루거 효과(**?**)가 발생할 수 있다고 제안합니다.

특히 이러한 모든 솔루션은 완전히 결정론적이며 머신 러닝 요소를 포함하지 않습니다.

과제 4의 최고 솔루션. 우승 솔루션은 74.74%의 예측 정확도를 가진 새로운 *메타러닝* 프레임워크를 제안한 미국 애머스트 매사추세츠 대학교(**?**)의 아리트라 고쉬(Aritra Ghosh)의 작품입니다. 직관에 따르면 모델은 각 학생의 답변 기록 몇 개만 관찰한 후 각 학생에게 개인화된 질문 순서를 선택하는 등 각 학생에게 빠르게 적응해야 합니다. 시험 시간 동안의 이러한 소수 샷 학습 설정은 훈련 절차도 이러한 소수 샷 학습 설정을 따르는 것이 가장 좋다는 것을 의미합니다. 따라서 그들은 메타 학습 문제를 공식화하여 모델이 먼저 각 학생에 대해 훈련 데이터에서 분할된 서로 다른 작은 답변 하위 집합을 최적화한 다음, 훈련 데이터와 구별되는 또 다른 답변 하위 집합을 최적화하는 것을 목표로 합니다. 또한 이미 관찰된 답을 바탕으로 다음 문제를 선택하는 샘플링 기반 방법도 제안합니다. 이들은 강화 학습의 행위자-비평가 네트워크와 같이 더 복잡하고 잠재적으로 더 강력한 문제 선택 정책을 프레임워크에 통합할 수 있다고 언급합니다.

훈련 중에는 훈련 데이터의 여러 파티션을 사용하여 모델을 훈련합니다. 각 문제마다 답안 기록을 분할하는 것 외에도 모든 문제를 훈련 및 로컬 테스트 세트로 분할하는 등 여러 가지 다른 훈련 및 데이터 분할 기법이 사용됩니다. 이러한 기법을 함께 사용하면 모델이 과적합하지 않고 초기 답안 기록이 거의 없는 다른 학생에게도 일반화할 수 있습니다.

5. 관찰 및 인사이트

교육 분야 지식의 중요성. 모든 과제에서 성공적인 솔루션은 교육 도메인 지식의 중요성을 보여줍니다. 예를 들어, 과제 1에 대한 최상의 솔루션은 학생과 질문에 대한 여러 관련 메타데이터를 활용하고 이를 특징 선택 및 앙상블 방법과 결합합니다. 이러한 관찰은 성공적인 예측 및 분석 모델을 구축하려면 단순히 더 강력한 모델을 구축하는 것 외에도 교육 도메인 지식을 잠재적으로 블랙박스 머신러닝 방법에 창의적이고 독창적으로 통합하는 것이 중요하다는 것을 시사합니다. 경험적으로 딥러닝과 같은 최첨단 모델에만 의존하는 방법은 교육적 도메인 지식과 직관을 활용하는 방법에 비해 성능이 떨어지는 것으로 나타났습니다. 또한 도메인 지식이 내장된 모델은 예측 작업에 유용한 기능에 대한 귀중한 인사이트를 제공하여 데이터 수집 프로세스를 안내하고 최적화하는 데 도움이 될 수 있습니다.

엔트로피 기반 질문 품질 측정 지표의 잠재적 유용성. 최고 성능의 솔루션 중 다수는 문제 품질을 측정하는 방법으로 엔트로피를 활용합니다. 엔트로피는 문제의 답안 선택지 간, 그리고 정답과 오답 학생의 답안 간의 균형을 측정하는 적절한 척도입니다. 제출된 일부 솔루션의 직관에서 알 수 있듯이, 이러한 균형이 잘 잡힌 문제는 너무 쉽거나 어렵지 않은 적당한 난이도를 가지며, 지식의 숙달 여부에 따라 학생을 구분할 수 있습니다. 따라서 이러한 문제는 높은 수준의 문제여야 합니다.

위의 관찰을 통해 두 가지 흥미로운 인사이트를 얻을 수 있습니다. 첫째, 앞서 언급한 직관은 전문 인간 평가자의 판단 기준과 밀접하게 일치하는데, 이러한 직관에 기반한 지표로 평가한 질문의 품질은 전문 인간 평가자가 평가한 품질과 상당 부분 일치하기 때문입니다. 둘째, 엔트로피는 위의 직관을 정량화하는 데 적합한 방법입니다. 우리의 인사이트에 따르면 엔트로피 기반 메트릭은 전통적으로 지나치게 주관적이고 계산하기 어렵다고 여겨졌던 질문 품질을 객관적으로 정량화할 수 있는 유망한 방향입니다.

교육을 위한 새로운 머신러닝 기법의 가능성. 모든 출품작에서 최신 머신러닝 기법을 창의적으로 응용한 사례가 많았습니다. 예를 들어, 과제 4에서 우승한 솔루션은 강화 학습, 소수의 샷 학습, 이중 수준 최적화 등의 기법을 활용하는 메타 학습 프레임워크를 개발했습니다. 이러한 새로운 방법은 부스팅 방법이나 앙상블과 같이 데이터 과학 대회에서 전통적으로 성공한 방법보다 뒤떨어지기도 하지만, 교육 데이터 모델링에 새로운 아이디어를 도입하고 교육 이외의 다른 실제 데이터 과학 문제에도 기여할 수 있는 잠재력을 가지고 있습니다.

6. 경쟁 영향

이 섹션에서는 경쟁이 교육용 AI에 미칠 수 있는 잠재적 영향에 대해 설명합니다. 또한 다른 교육 및 광범위한 영향에 대해서도 논의합니다.

교육용 AI에 미치는 영향. 섹션 3에서 설명한 것처럼 각 대회 과제는 실제 교육 문제에 뿌리를 두고 있습니다. 과제 1과 2를 성공적으로 해결하면 각각 더 정확한 학생 분석과 오개념 식별이 가능해집니다.

이러한 개선 사항은 교사와 개인화된 학습 알고리즘이 학생의 학습 방식을 더 잘 파악할 수 있도록 지원하여 보다 효과적인 개인화된 학습으로 이어질 수 있습니다. 과제 3에 대한 성공적인 솔루션은 문제 품질을 정량화할 수 있는 새로운 아이디어를 제공합니다. 이러한 아이디어는 문제 품질 정량화에 대한 새로운 연구를 촉발하고 이를 위한 예비적인 방법을 제공할 것입니다. 마지막으로, 과제 4에 대한 성공적인 솔루션은 보다 효과적인 문제 순서 선정으로 이어질 것입니다. 이는 잠재적으로 적응형 시험 알고리즘의 성능을 향상시키고, 온라인 평가의 효율성을 개선하며, 교사가 수동으로 문제를 선택하는 시간을 절약할 수 있게 해줄 것입니다.

우리의 경쟁은 우리가 소개한 과제 외에도 교육용 AI에 잠재적으로 도움이 될 수 있습니다. 규모가 크고 메타데이터가 풍부한 우리의 데이터 세트는 교육용 AI의 다른 많은 연구 문제에 기여할 것입니다. 예를 들어, 우리 데이터 세트에는 답변의 타임스탬프가 포함되어 있어 학생들의 시간 경과에 따른 진도를 추적하는 교육 데이터 마이닝의 근본적인 문제 중 하나인 지식 추적**(????)의** 설정에 완벽하게 부합합니다. 또한 데이터 세트에는 각 질문이 테스트하려는 주제/기술이 포함되어 있어 세분화된 오개념 식별 및 분석에 사용할 수 있습니다**(**?????). 마지막으로, 데이터 세트에는 문제 텍스트가 포함된 문제 이미지가 포함되어 있습니다. 이러한 텍스트와 이미지는 다양한 교육 데이터 마이닝 작업에서 모델링 성능을 향상시키기 위해 이미지와 자연어 등 멀티모달 데이터 통합에 대한 연구를 가능하게 할 수 있습니다.

광범위한 영향력. 이번 대회는 해결해야 할 여러 가지 근본적인 머신 러닝 과제를 포함하고 있기 때문에 교육을 넘어 더 광범위한 영향을 미칩니다. 예를 들어, 처음 두 과제에 대한 솔루션은 딥러닝 기반 방법과 광범위한 피처 엔지니어링을 사용하는 앙상블 기반 방법 모두 실행 가능한 솔루션이라는 것을 보여줍니다. 이러한 관찰은 향후 매트릭스 완성 및 추천 시스템 설계에 두 가지 방법의 장점을 더 잘 활용할 수 있는 개발에 영감을 줄 수 있으며, 이는 과제 1과 과제 2의 문제를 일반화한 것입니다. 과제 3의 솔루션은 정보 검색과 같이 순위와 관련된 다양한 애플리케이션에서 새로운 엔트로피 기반 순위 지정 방법으로 유용하게 사용될 수 있습니다. 과제 4의 솔루션은 일반적으로 능동 학습 및 베이지안 실험 설계를 위한 메타 학습 및 이중 수준 최적화와 같은 새로운 방법론을 도입합니다. 전반적으로 모든 과제가 기술적으로 도전적이고 흥미로웠으며, 이번 대회에 제출된 창의적인 아이디어가 향후 NeurIPS 커뮤니티와 그 너머에 미칠 영향을 지켜보게 되어 기대가 큽니다.

7. 결론

유니티는 교육 분야에서의 AI 적용에 초점을 맞춘 포괄적인 챌린지인 NeurIPS 2020 교육 챌린지를 개최합니다. 오늘날 대규모 온라인 교육 플랫폼이 직면하고 있는 실질적이고 시급한 문제를 대표하는 네 가지 과제를 소개합니다. 학생과 문제 모두에 대한 풍부한 메타데이터가 포함된 현존하는 최대 규모의 교육 데이터 세트를 큐레이팅합니다. 이 대회는 대규모 교육 데이터 마이닝에 참신한 아이디어를 제시하고 각 과제에 대한 유망한 미래 연구 방향을 제시하는 솔루션으로 전 세계 참가자들의 큰 관심을 끌었습니다. 마지막으로, 대규모의 실제 데이터 세트는 대회에서 지정한 과제보다 교육용 AI와 머신 러닝 분야에서 훨씬 더 광범위하게 응용될 수 있다는 점에 주목합니다. 우리의 대회는 의도적으로 객관식 진단 문제에 대한 학생의 응답이라는 집중된 대상을 정확히 찾아냅니다. 하지만 교육은 매우 복잡하고 다양한 분야가 존재합니다,

다른 많은 중요한 문제가 있습니다. 예를 들어, 학생의 학습 성과를 정의하고 정량화하기는 어렵습니다. 현재 환경에서 학생들의 목표는 자신에게 주어진 모든 문제에 정답을 맞추는 것입니다. 이는 교사가 중요하게 생각하는 실제 학습 결과의 근사치가 아닐 수 있습니다. 또한 현재로서는 데이터를 수집한 *후에야* 학생의 성취도를 분석할 수 있습니다. 학생들의 학습 과정 중에 분석을 수행하고 *개입하는* 것은 매우 바람직하지만 어려운 일이지만, 이는 윤리적 문제를 야기할 수 있습니다. 또한 학습에 큰 영향을 미칠 수 있는 협업, 감정 상태, 자신감 등 다른 유형의 학습 활동 데이터는 수집하기 어렵습니다. 마지막으로, 학생과 교사의 개인정보를 보호하면서 AI 애플리케이션을 위한 대규모 교육 데이터를 수집하고 저장하는 최선의 방법을 결정하는 것은 현재 진행 중인 연구 과제입니다. 우리는 이러한 어려운 문제를 해결하기 위해 계속 노력하고 있으며, 데이터 세트와 대회를 통해 얻은 인사이트가 교육 실무자와 다양한 연구 커뮤니티에 장기적으로 긍정적인 영향을 미칠 것으로 믿습니다.

감사

대회 기간 동안 기술 지원을 아끼지 않은 코다랩 팀과 기여해 주신 모든 대회 참가자에게 감사드립니다. ZW와 RGB는 NSF 보조금 1842378 및 1937134와 ONR 보조금 N0014-20-1-2534의 지원을 받았습니다.