숭실대학교 Soongsil University

알고리즘 기반 의사결정시스템의 한계

숭실대학교 베어드교양대학
서유화 교수
yhsuh@ssu.ac.kr

지난시간 배운 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 편향성이 예측에 미치는 영향 (구글티쳐블머신) 3. 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 타이타닉호 생존자 예측 - 데이터 편향성이 예측에 미치는 영향 - 데이터 왜곡에 따른 예측 결과 비교
5	인공지능과 알고리즘 윤리	1. 알고리즘과 모델링의 개요 2. 알고리즘 기반 의사결정 시스템의 한계 3. 윤리가 필요한 인공지능 4. 오렌지3 설치 및 사용법	1. 오렌지3를 이용한 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	

목차

- 알고리즘의 취약점과 오류
- 알고리즘 설계자와 사용자의 책임
- 인공지능 모델의 특성
- 윤리가 필요한 인공지능

학습 목표

- 인공지능에서 알고리즘의 취약점과 오류에 대해 설명할 수 있다.
- 알고리즘의 사용자와 개발자의 책임을 인식한다.
- 인공지능 모델의 특성을 설명할 수 있다.
- 인공지능에 윤리가 필요한 이유를 설명할 수 있다.
- 인공지능 활용에 유리 조건과 신중해야 할 조건을 구별할 수 있다.

알고리즘의 취약점

- 알고리즘의 취약점
  - 알고리즘은 적절한 인풋이 주어지면 결과를 출력
  - '도로'의 의미, '경로의 길이'가 해석 가능한지 등은 알지 못하고 행동지침만 실행
  - 결과를 해석하는 것은 인간의 몫이며 해석할 수 있으려면 적절한 모델이 구축되어야 함
  - 2017년 미국 웰스파고 은행은 대출연장 거부 사례
    - 2017년 3월 중순~2017년 4월말 사이 870명 고객에 대해 사실은 대출연장을 해주는 것이 은행에 더 유리했음에도 대출연장을 거부
    - 알고리즘이 계산한 부당한 채무연장 거부로 545명의 고객이 집을 잃음
    - 소프트웨어가 공증료를 잘못 계산하여 채무연장이 은행에 매력이 없는 것으로 나타난 오류로 자동적 연장 거부 결과가 나옴

알고리즘의 비윤리적 투입 사례

- 2018년 영국 항공사 비행기 좌석 배치 알고리즘
  - 여러 항공사가 옆좌석에 나란히 앉아 가기를 원하는 승객들을 비교적 자주 갈라놓는 알고리즘 투입
  - 그럼에도 일행이 함께 앉으려고 하면 추가요금 징수
  - 영국 민간항공국 조사에 따르면 라이언에어가 비리에 연루되었으나 라이언에어는 부인

- 2015년 폭스바겐 배기가스 조작 스캔들
  - 차량이 테스트 상황인지 실제 도로에 있는지를 감지하는 소프트웨어 조작
  - 실험상황에서만 여러 시스템이 켜지거나 꺼져서 배출가스 기준을 충족
  - 실제 운행상황에서는 배기가스가 기준치의 40배 발생
  - 복잡한 배기가스 기술을 개발하는 비용을 절약, 엔진 성능을 높여 운전체험을 향상
  - 같은 그룹 산하 아우디에서도 조작이 일어난 것으로 밝혀짐

알고리즘 설계자들의 책임

- 알고리즘 개발자들은 어느정도 책임이 있을까?
  - 채무연장 거부 관련 프로그램과 같은 오류는 악의가 없었더라해도 알고리즘 개발한 회사의 책임
  - 다른 회사를 위해 개발한 알고리즘을 잘 못 사용하는 경우 사용자 책임
  - 사용자들이 알고리즘을 어떻게 이용해야 하는지 잘 아는 상태에서 필요한 정보를 잘못 입력하는 경우 사용자의 책임
  - 기본적인 알고리즘은 추상적인 수학 문제를 해결하기 위해 개발하고 알고리즘 하나의 사용범위는 넓기 때문에 활용을 알지 못하는 경우가 많음
  - 알고리즘 설계자들은 기본 수학의 문제를 어떻게 정의할지, 결과를 해석하기 위해 데이터들이 어떤 전제조건을 충족해야 하는지 명시해야 할 책임이 있음
    - 알고리즘이 특정 상황의 데이터에서만 해석될 수 있다면 그에 대해 명시해야 함

알고리즘 설계자들의 책임

- 알고리즘 개발자들은 어느정도 책임이 있을까?
  - 개발자 팀이 알고리즘의 정확한 활용에 대해 더 많이 알고 영향을 끼칠 수록 알고리즘의 행동에 대한 책임이 커짐
    - 알고리즘을 정확히 어디에 사용할지(맥락)을 잘 알고 사용을 감독할 수 있는 경우 그 결과에 대한 책임이 커짐
    - 알고리즘이 정확이 내가 알고있는 맥락에 사용되면 그것은 프로그래밍한 대로 행동하고 결과는 설계자 책임

알고리즘의 오류

- 인공지능에서는 알고리즘의 정확한 행동을 추상적 규칙으로의 예측이 어려움
  - 알고리즘 내에 조절요소(파라메터)가 어떻게 결합되는지에 따라 결과가 너무 다양하기 때문
  - 인공지능에서는 알고리즘의 행동을 추상적인 규칙으로 묘사할 수 없는 현상이 더 많고 알고리즘 오류를 찾는것도 어렵게 만듦
  - 해답의 특성이 명확히 정의된 경우는 검증이 쉽고 오류를 찾기가 쉽지만 해답을 명확히 정의 할 수 없는 경우가 많음
  - 많은 머신러닝은 휴리스틱한 접근을 취하기 때문에 최적의 해인지에 대해 정의하기 어려움 예) 추천시스템
  - 알고리즘이 의도적으로 비윤리적으로 투입된 경우는 더욱 어려움
    - 예)입법자가 의도한 것과 다르게 조절되는 알고리즘

알고리즘 사용자들의 책임

- 최상의 모델 선택을 위한 다양한 접근의 시도와 검증이 필요
  - 모델 및 알고리즘이 오류가 없는지 해답이 정확성에 대한 기준과 평가가 필요함
    - 해답의 특성이 명확히 정의하기 위한 실측자료를 확보
    - 실측자료는 정확해야하고, 대표성을 지녀야 하고, 차별을 포함하고 있지 않아야 함
    - 해답이 실측자료로 규정된 이상과 어느정도 거리에 있는지 품질척도로 측정가능 해야함
    - 예) 성공적인 입사지원자는 성공과 성공적이지 않은 것을 어떻게 구분해서 실측자료를 만들까?
  - 알고리즘 사용자는 최상의 모델을 구축하고 도출된 해답에 대한 품질평가과 공정성을 기반으로한 해석과 행동이 요구됨
  - 인공지능 기반 의사결정 과정에 있는 많은 관계자들의 협력을 통해 실측자료와 품질척도를 명확히 하는 것이 중요함

인공지능 모델의 특성

- 머신러닝은 많은 부분은 휴리스틱에 가깝고 인간에 의한 결정에 의해 이루어짐
  - 예를 들어 의사결정 나무는 알고리즘이라기 보다 휴리스틱에 가까움
    - 구축한 의사결정 나무가 최상의 것인지 알 수 없음
    - 테스트데이터 세트와 품질 척도로 그 품질을 기술
  - 트레이닝 세트와 테스트 세트의 규모, 데이터 학습의 하이퍼파라메터(설계변수) 조절
    - 데이터 선택에 있어서도 차별이 없도록 구성
    - 모델이 언제 만족하고 더 이상 모델을 개량하지 않을지 선택
    - **피처 엔지니어링(feature engineering)**
      - 입력 데이터의 정확한 구성을 결정하는 모든 단계
      - 인공지능의 예측에 커다란 영향을 미침
  - 테스트 데이터 세트가 양질의 것이고 품질 척도, 공정성의 척도가 의미있게 정해져야 함
    - 무엇이 좋은 결정인지는 측정가능하게 끔 되어야 하고 인간이 정의해야 함
    - 예) 채용시스템에서 어떤 지원자를 잘못해서 면접에 부르는 것과 잘못해서 면접에 부르지 않는 것이 동일하게 안 좋은 것일까?

숭실대학교
Soongsil University

윤리가 필요한 인공지능

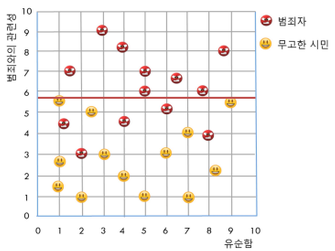
윤리가 필요한 인공지능

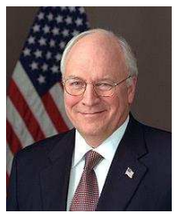
▸ 품질 척도의 선택에는 언제나 도덕적 숙고가 들어간다

▸ 인공지능 재범율 예측 결과의 예

법학자 윌리엄 블랙스톤

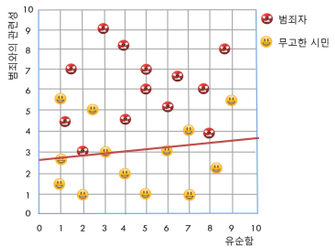
"무고한 한 사람이 고통당하느니 열 사람의 죄인을 놓치는 것이 더 낫다."





미국 제 46대 부통령 딕 체니

"나는 몇몇 무고한 사람들이 구속되는 상황보다 반대로 범죄를 저지른 사람들이 풀려나는 일이 발생하는 것이 더 우려스럽다."



Soongsil University

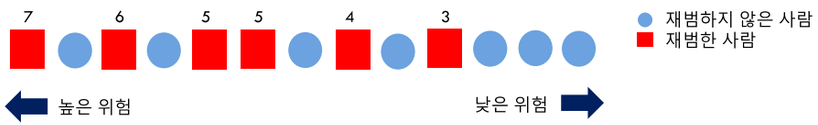
윤리가 필요한 인공지능

▸ 예측 품질 척도의 중요성

▸ 1998년 석방된 범죄자들이 재범률 위험 평가 소프트웨어(COMPAS) 사례

- 해당 소프트웨어의 평가는 지원대책 승인 결정의 자료로 사용됨
- 전과자를 위험값에 따라 분류하고 두 개의 기준값을 설정
- 위험값이 두 개의 기준값 사이에 위치하면 '중등위험군 ' 으로 분류
- 위험값이 두 번째 기준값보다 높은 사람은 '고위험군'으로 분류
- 품질척도를 규정하기 위해 테스트 데이터 세트에 짝을 지어 고려 (둘 중 하나는 재범이고 하나는 아님)
- 알고리즘 기반 의사결정 시스템이 얼마나 자주, 재범을 한 사람에게 그 짝보다 더 높은 위험값을 매겼는지 테스트 결과 70%

7개의 동그라미 좌측에 위치, 7개의 좋은 쌍 가능
7+6+5+5+4+3 = 30
30/42 *100 = 71%



Soongsil University

윤리가 필요한 인공지능

▸ 예측 품질 척도의 중요성

▸ 1998년 석방된 범죄자들이 재범률 위험 평가 소프트웨어(COMPAS) 사례

- 오늘날에는 재범 예측 시스템이 사전심리 절차에도 사용
- 사전 심리에서는 고위험군에만 관심이 있어 고위험군 중 몇 명이 재범을 할 것인가를 기준으로 평가
- 결과 일반적인 범죄에서는 고위험군은 테스트 데이터세트의 70% 이상 재범, 폭력범죄에서는 고위험군 중 25%만 재범

▸ 테러리스트 색출 시스템 스카이넷(SKY NET)

- 휴대전화 단말기의 메타데이터를 활용해 특정 행동 패턴을 지닌 단말을 색출
- 5,500만 유저 핸드폰 데이터에서 테러조직들 사이를 오가는 전달책 색출이 목적
- 사용 핸드폰 단말기 데이터 활동 시간, 활동 장소, 접촉 사람, 의사소통 방법, 유심칩 교환횟수, 여행횟수, 교류가 없는 그룹과의 커뮤니케이션, 밤의 활동시간 등

Soongsil University

윤리가 필요한 인공지능

▸ 예측 품질 척도의 중요성

▸ 테러리스트 색출 시스템 스카이넷(SKY NET)

- 전달책의 특성을 학습하기 위한 학습데이터 : 법적으로 유죄판결을 받은 7명의 테러 전달책, 셀렉터(의심스러운자, 혹은 이미 판결받은자와 접촉한 사람)
- 5,500만명에 대해 위험값 부여 기준값은 테러리스트(판결받은 자+셀렉터)의 50%는 기준값 왼쪽, 50%는 오른쪽에 오게끔 자의적으로 결정
- 알고리즘이 무고한 사람들 중 0.008%를 의심자로 분류 (0.008%는 5,500만명 중 4,400명)

- **기준값의 결정은 윤리적 결정이고 알고리즘이 알아서 내릴 수 없으며, 알고리즘 사용자들이 알아서 내릴 수 있는 결정이 아니다.**
- **알고리즘은 그러한 감수성이 없고 균형감각도 없고 자신의 품질 척도만 알고 있다.**

Soongsil University

실측 자료의 정의는 어떻게 할까?

- 재범예측 사례에서 재범여부는 어떻게 정의할 수 있을까?
  - 재범여부는 어떤행위를 하는 것만으로는 불충분, 검거되어 그 행위에 대해 기소되고 형을 선고 받아야 함
    - 미국에서는 경찰이 더 주목하고 검열하는 그룹과 특히 적발되기 쉬운 범법행위들이 존재
    - 유죄판결 비율도 인구집단 사이에 차이가 존재
  - 시스템이 최적화되지 않은 품질에서 출발하여 피드백을 통해 개선되는 것은 바람직
    - 예) 유저클릭을 통해 상품추천시스템에 대한 빠른 피드백이 가능
  - 인간행동의 예측은 일방적인 피드백인 경우가 많아 피드백으로 활용이 어렵다
    - 어떤 사람이 고위험군으로 잘 못 분류되어 수감기간이 길어지는 경우 올바른 피드백이 어려움
    - 무고한 사람이라도 형을 살고 나오면 다른 범법행위를 저지를 확률이 높아져 범죄를 저질러야 예측을 확인할 수 있거나 재범을 하지 않더라도 알고리즘이 틀렸다고 적용될 수 없음
    - 고위험군으로 분류된 사람에 대해 예측과 그의 행동을 비교해 알고리즘을 쉽게 수정할 수 없음

윤리적 측면이 더욱 중요한 인공지능 시스템

- 인공지능의 내부 역학을 감독하고 조절할 필요가 있는 시스템
  - 인간에 대해 결정하는 시스템
  - 인간에 관계된 자원에 대해 결정하는 시스템
  - 인간의 사회참여 가능성을 변화시킬 결정을 내리는 시스템
- 인공지능을 통한 결정의 질을 좌우하는 요소
  - 투입되는 데이터의 질과 양
  - 사안의 특성에 대한 기본적인 가정
  - 사회가 '좋은' 결정이라고 여기는 모델을 구축하는 것

인공지능의 활용이 유익한 조건

- 인공지능이 성공적일 수 있는 조건
  - 양질의 방대한 학습데이터(인풋)가 있을 때
  - 측정가능한 실측자료, 즉 예측할 수 있는 것(아웃풋)이 있을 때
  - 인풋과 예측할 수 있는 아웃풋 사이에 인과관계가 있을 때
- 인공지능 결과를 신뢰할 만한 조건
  - 인풋과 예측되는 아웃풋 사이에 인과관계가 알려져 있어 관계자들이 쉽게 합의할 수 있는 명확한 인풋 데이터가 존재할 때
  - 알고리즘이 예측한 것과 실측 데이터가 맞지 않는 경우에 대해 가급적 많은 피드백이 있을 때 그로써 지속적으로 품질을 측정해 개선할 수 있을 때
  - 모든 관계자들이 쉽게 동의할 수 있는 명확한 품질 척도가 있을 때
- 위의 조건을 충족한다는 이유로 인공지능이 상업적으로 활용되어 우수한 성과를 보인다고 해서 인간행동에 대한 의사결정에 투입하는 것은 문제가 있고 정확한 기술적 감독이 요구된다.

다음시간에 배울 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성,) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 편향성이 예측에 미치는 영향 (구글티처블머신) 3. 지도학습(SVM)을 이용한 타이타닉호 생존자 예측	1. 타이타닉호 생존자 예측 - 데이터 편향성이 예측에 미치는 영향 - 데이터 왜곡에 따른 예측 결과 비교
5	인공지능과 알고리즘 윤리	1. 알고리즘과 모델링의 개요 2. 알고리즘 기반 의사결정 시스템의 한계 3. 윤리가 필요한 인공지능 4. 오렌지3 설치 및 사용법	1. 오렌지3를 이용한 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	