



ARTIFICIAL INTELLIGENCE 501

Lesson 5

데이터 수집 및 향상

Learning Objectives

You will be able to:

- 데이터 소스 및 유형.
- 더 많은 데이터 샘플이나 기능이 필요한 상황.
- 데이터 논쟁, 기능 보강 및 기능 엔지니어링.
- 다른 데이터 전처리 방법.
- 데이터에 레이블을 붙이는 방법.
- 데이터 작업시 문제.

Data Collection

데이터를 수집 할 때 고려해야 할 몇 가지 사항이 있습니다.

- 데이터의 출처는 어디입니까?
- 어떤 유형의 데이터가 있습니까?
- 얼마나 많은 데이터와 어떤 속성이 필요합니까?



DATA SOURCES

Data Sources

데이터는 여러 다른 장소에서 제공됩니다.

- 사물의 인터넷 (IoT)과 기계에서 생성.
- 공개 웹 사이트.
- 기존 문서.
- 멀티미디어.



Human Generated Data: Social Media

지난 2 년 동안 세계 데이터의 90 %가 만들어졌습니다.

- 전문가들은 이 데이터의 약 70 %가 소셜 미디어에서 나왔다고 보고 있습니다.
- 우리의 현재 데이터 출력량은 하루에 2.5 엑사 바이트 / 일 입니다.


Human Generated Data: Social Media

데이터에 액세스하는 데는 다양한 API가 있습니다.

- Facebook*: Graph API
- Twitter*: REST API or Streaming API
- Instagram*: Graph API or Platform API
- LinkedIn*: REST API
- Pinterest*: REST API

Internet of Things (IoT) Data

IoT 데이터는 각 환경에서 요구하는 데이터를 수집하고 보내고 응답하는 모든 웹 지원 장치로 구성됩니다.

- 2020 년까지 IoT에는 2,000 억 개의 스마트 연결 장치가 포함될 예정입니다.
 - 생산 된 데이터는 2 년마다 40 제타 바이트 (40 조 기가) 것으로 예상됩니다.
 - 센서, 카메라 및 프로세서를 통해 수집됩니다.
- 



Internet of Things (IoT) Data: Consumer

소비자 IoT는 사용자 경험과 인터페이스를 위한 새로운 통로를 제공합니다.

- Connected Car, 스마트 홈 장치 및 웨어러블



Internet of Things Data: Industrial

산업에서 운영과 운영도구를 모니터링하고 제어하는 데 사용됩니다.

- 외과 붓은 카메라를 통해 수술을 할 수 있습니다.
- 자율 트럭에 설치된 수많은 센서 및 카메라.
- 실시간 피드백을 제공하는 제트 엔진 센서.



Public Websites : How to Access the Data

Webscrapping:

- 주의가 필요합니다 : 크롤러 및 웹 스크랩에 대한 웹 사이트는 일반적으로 사이트의 이용 약관에 텀에 컨디션이 있습니다.
- 크롤러가 웹 사이트 robots.txt 파일에 정의 된 규칙을 따르는 지 확인하십시오.

APIs:

- 종종 웹 스크랩보다 쉽다.

Legacy Documents

여러 산업 분야에서 종이 양식을 사용하여 데이터를 수집하여 디지털로 전환 할 수 있는 기회를 창출했습니다.

- 디지털 방식으로 접근하는 두 가지 주요 산업은 보험과 의약품입니다.
- 의료 기록은 전통적으로 종이와 연필이었습니다.
- 디지털에 대한 요구는 환자에게 더 나은 결과를 가져올 것을 약속합니다.



Multimedia

이전보다 더 많은 미디어 유형에 데이터가 존재합니다.

- 회사는 텍스트, 이미지, 오디오 및 비디오를 수집합니다.
- 이 데이터를 저장하기 위해 새로운 데이터베이스 기술이 발전했습니다.
(MMDB - 멀티미디어 데이터베이스)
- 새로운 ML 기법 (예 : Deep Learning)은 이 데이터를 분석 할 필요성 때문에 부분적으로 발전했습니다.





DATA TYPES

데이터 타입

- 수치
- 숫자
- 범주형
- 서수
- 이진수
- 날짜 시간
- 텍스트
- 영상
- 오디오



Count Data

Integer valued data that comes from counting.

- 예를 들어, 주차장에 있는 차량 수.
- 카운트 데이터의 프로그래밍 용어는 정수입니다.
- 파이썬에서는 int입니다.



Numerical Data: Continuous

Numerical value that can represent any quantity over a continuous range.

- 10 진수 값을 취할 수 있습니다. (예 : 엔진 행정 = 3.40 인치)
- 미세한 정밀도로 줄일 수 있습니다.
(예 : 엔진 행정 = 3.3775 인치)
- 프로그래밍 용어는 부동 소수점입니다.
- Python *에서는 float입니다.



Categorical Data

알려진 카테고리의 유한 집합으로 제한되는 데이터.

- 명목 데이터라고도 합니다.
- 원시 범주형 데이터는 다른 데이터 형식의 형태로 올 수 있습니다. (예 : 텍스트 데이터 (차량 색상 : 빨간색) 또는 수치 데이터 (문 개수 : 4)
- 프로그래밍 용어는 열거 형입니다.



Ordinal Data

순서가 지정된 범주 데이터.

- 카테고리 간 거리를 알 수 없습니다. (예 : 저, 중, 고의 자동차 가격 값 또는 열악한, 양호한, 양호한 고객 리뷰)
- 일반적인 실수는 서수 데이터를 단순히 정수로 변환하는 것입니다. (예 : G, PG 및 R을 1, 2 및 3으로)
- 고유 한 문제점은 카테고리 간 거리가 동일하고 동일하다고 가정한다는 것입니다.

Binary Data

상호 배타적 인 두 가지 범주.

- 이진 데이터는 특히 감독 된 학습 문제에서 매우 일반적입니다.
(예 : 참 / 거짓, 머리 / 꼬리).
- 이진 데이터의 프로그래밍 용어는 부울입니다.
- 파이썬에서는 bool입니다.



Date-time Data

날짜 및 시간 정보를 나타내는 데이터.

- 날짜, 시간 및 24 시간제를 기준으로 한 부분 초가 결합됩니다.
- 기본 유형으로 변환 할 수 있습니다.
- 범주형 데이터로 변환하거나 Unix * 타임 스탬프를 통해 정수로 변환 할 수 있습니다. (예 : 날짜, 월, 연도, 시간 만)
- 대부분의 언어에는 날짜 / 시간 데이터에 대한 기본 제공 데이터 유형이 있습니다.
- 파이썬에서도 datetime입니다.

Text Data

영숫자 문자열, 일련의 문자.

- 일반적으로 사람이 읽을 수 있습니다.
- ASCII 또는 유니 코드와 같은 컴퓨터가 읽을 수 있는 형식으로 인코딩 할 수 있습니다.
- 종종 구조화되지 않은 데이터입니다.
- 텍스트 데이터의 프로그래밍 용어는 문자열입니다.
- 파이썬에서는 str입니다.

Image Data

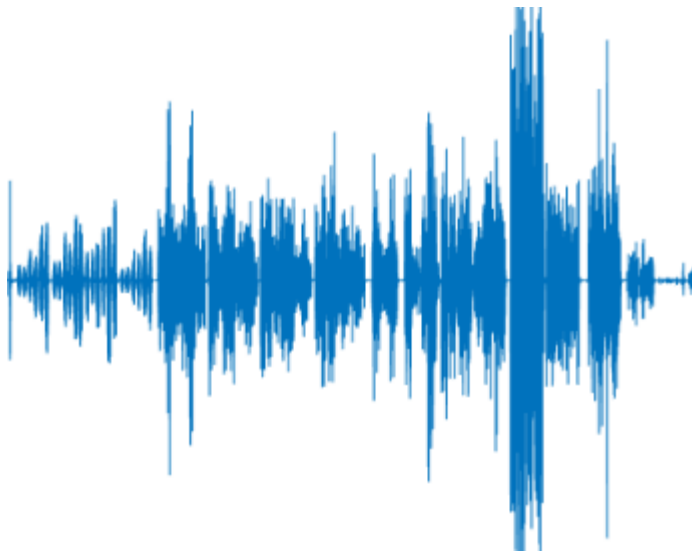
스틸 및 비디오 이미지가 빠른 속도로 생성됩니다.

- 그들은 다양한 방법으로 저장됩니다. (예 : PNG 및 JPEG)
- 이미지는 매우 크고 계산 집약적 일 수 있습니다.
- 3D 바디 스캔에서와 같이 다차원적일 수 있습니다.
- 다양한 분야의 다양한 장치 유형에서 생성됩니다. (예 : 인공위성,자가운전용 자동차, 감시 카메라, 휴대 기기)
- 의료 영상은 점점 더 중요한 분야입니다.

Audio

오디오 데이터는 콜센터와 같은 산업 환경뿐 아니라 Alexa *와 같은 소비자 장치에 사용됩니다.

- 다양한 압축 및 비 압축 방식으로 저장됩니다.
(예 : WAV 및 MP4)





THE SHAPE OF DATA

The Shape of Data

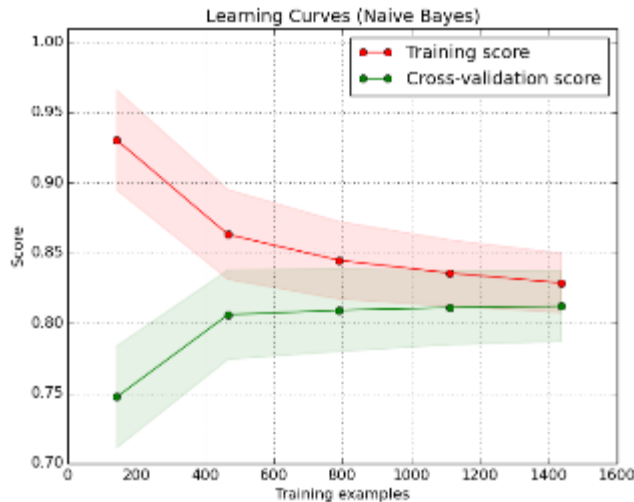
문제를 해결하는 데 필요한 샘플 및 기능의 수를 확인하십시오.

- 더 많은 기능을 사용할수록 일반적으로 더 많은 샘플이 필요합니다.
- 기능에 충분한 정보가 포함되어 있지 않으면 샘플을 더 추가하면 도움이 되지 않으며 추가 기능이 필요합니다.
- 샘플 수가 너무 적 으면 추가 기능을 추가하면 overfitting이 발생할 수 있습니다.
- 일반적으로 어려운 문제는 더 많은 기능과 추가 샘플이 필요합니다.

How Many Samples?

There are several ways to determine the number of sample needed.

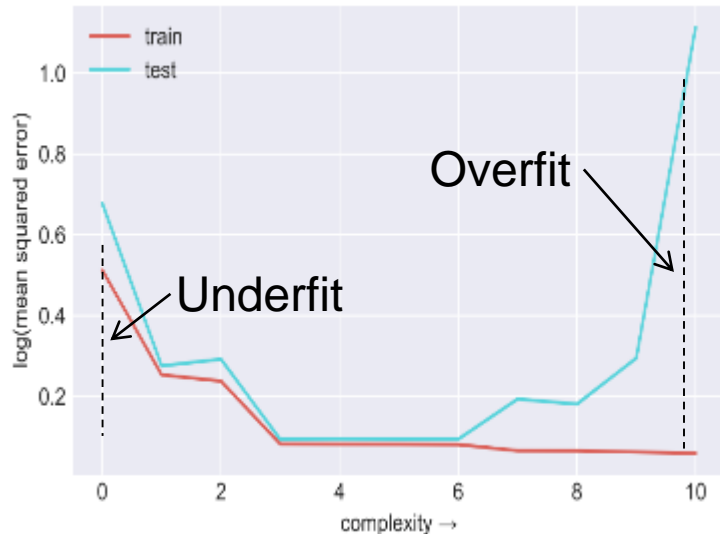
- 학습 곡선은 샘플 수로 실적 점수를 표시합니다.
- 이 곡선을 사용하여 샘플 개수를 늘리면 모델에 도움이되는지 판단 할 수 있습니다.
- 통계적 발견 적 방법 : 자유도의 10 배에 달하는 샘플. 다른 성공과 실패로부터 배우기 위해 유사한 연구를 정독하십시오.



Under-fitting vs Overfitting

학습 곡선은 모델이 과소 적합 또는 과잉 여부를 진단하는데도 도움이 됩니다.

- 모델 복잡성으로 성능 점수를 그려보십시오.
- 더 많은 기능은 더 많은 복잡성을 의미합니다.
- 열차와 시험 성능이 비슷하지만 낮을 때 모델이 적합하지 않습니다.
- 테스트 성능이 떨어지면 열차 성능이 향상되지만 모델은 지나치게 적합합니다.



How to Increase the Number of Features?

피쳐 수를 늘리면 모델이 잘 맞지 않을 때 도움이됩니다.

- 데이터 파이프 라인을 다시 방문하고 이전에 제거 된 기능을 추가하십시오.
- 새로운 기능 설계 -이를 기능 엔지니어링이라고합니다.
- 다항식 피쳐는 일반적인 방법입니다. 이는 기존 피쳐를 곱하여 새로운 피쳐를 작성합니다. (예 : $x_1 * x_2$ 또는 $x_2 * x_2$)
- 이미지 / 오디오 : 변형을 추가하십시오. (예 : 델타 피쳐 - 기존 이미지에 픽셀에서 픽셀로의 색상 변경 방법을 추가 할 수있는 방법)
- 컴퓨터 비전을 위한 이미지 확대.
- 기존 이미지를 이동, 뒤집기, 회전하여 새 이미지를 만듭니다.

Reducing Features: Feature Selection

모델의 복잡성을 줄이고 중복 기능이나 약한 기능을 제거하려면 기능의 하위 집합을 선택하십시오.

- 필터 방법 : 통계 방법 (p 값)을 통해 기능을 필터링합니다.
- 정규화 방법 : 모델이 생성되는 동안 어떤 기능이 정확도에 가장 기여하는지 알아보십시오.
- 래퍼 메서드 : 기능의 하위 집합을 검색 문제로 선택합니다. 여기서는 기능 간의 상호 작용을 캡처하기 위해 여러 조합을 평가합니다.

Reducing Features: Dimensionality Reduction

차원성 감소는 특징 수를 줄이기 위해 사용되는 감독되지 않은 학습 기법입니다.

- 지나치게 많은 정보를 잃지 않고 차수를 줄이거나 압축하십시오.
(예 : 공분산 행렬의 고유 벡터를 축소 된 차원 공간으로 사용)



DATASETS

Datasets for AI

AI에 사용되는 표준 데이터 유형은 없습니다.

- 모양과 크기가 다를 수 있습니다.
- 그것들은 다양한 데이터 유형으로 구성 될 수 있습니다.
- 이들은 작업에 따라 다릅니다.

Open Source Repositories

오픈 소스 데이터 세트를 얻는 데는 여러 저장소가 있습니다.

UCI Machine Learning Repository: ML 커뮤니티 내에서 알고리즘을 분석하는 데 사용되는 많은 대중적인 데이터 세트를 포함합니다.

- **ImageNet:** 이미지의 대형 데이터베이스.
- **KDD Cup:** 컴퓨팅 기계 협회 (Association for Computing Machinery)가 주최하는 연례 경쟁입니다. 연간 경쟁 데이터 세트가 보관됩니다.
- **Kaggle*:** 데이터 과학 경진 대회를 위한 플랫폼. 경쟁 데이터 세트 및 기타 데이터 세트를 사용할 수 있습니다.
- **Data.gov:** 미국 정부의 공개 데이터.
- And many more...

ImageNet

ImageNet은 AI 커뮤니티에서 널리 사용되는 대형 이미지 데이터베이스입니다.

- 2009 년 컴퓨터 비전 및 패턴 인식 컨퍼런스에서 처음 발표되었습니다
- 현재 1,400 만 개가 넘는 이미지를 보유하고 있습니다.
- 연간 개체 탐지 및 분류 경쟁을 개최합니다.
 - 딥 학습 모델은 현대 AI에 대한 심층 학습에 초점을 맞춘 획기적인 결과를 가져 왔습니다.
 - 가장 인기있는 심층 학습 모델 중 상당수가 이 경쟁에서 나왔습니다.

Example Image Datasets

다음은 인기있는 이미지 데이터 세트 예제입니다.

- **MNIST:** 많은 기계 학습 모델에서 널리 사용되는 벤치 마크입니다.
 - 필기 이미지 숫자들이며
 - 70,000 개의 28 x 28 픽셀의 흑백 이미지들 입니다 .
- **Cifar-10:** 컴퓨터 비전 연구에 널리 사용되는 데이터 세트.
 - 60,000 32 x 32 픽셀 컬러 이미지들이며
 - 분류를위한 10 개의 다른 오브젝트, 각 오브젝트별 6,000 개의 이미지가 있습니다.
- **ILSVRC:** 매년 개최되는 ImageNet 시합에는 각각 자체 데이터 세트가 포함 된 여러 구성 요소가 있습니다.
 - Object localization: 1,000 개의 오브젝트에 대한 120 만 개의 트레이닝 이미지
 - Object detection: 200 개의 오브젝트에 대한 456,567개의 트레이닝 이미지

Example Natural Language Datasets

다음은 인기있는 언어 데이터 집합입니다.

- **Common Crawl:** 웹을 일 년에 네 번 크롤링하고 아카이브 및 데이터 세트를 일반인에게 무료로 제공합니다.
 - 1.8 billion 웹페이지에서 145TB 이 데이터를 크롤 (in 2015.)
- **Stanford Question Answering Dataset:** A dataset of over 100,000 이상의 질문에 대한 답변 데이터셋.
- **Project Gutenberg:** 56,000 이상의 free eBooks.

Example Datasets

다른 작업에 사용할 수 있는 많은 데이터 세트가 있습니다.

- **YouTube*-8M Dataset:** 약 7 백만 비디오 URLs 과 450,000 시간의 비디오.
- **MovieLens* Dataset:** 약 27,000 비디오와 138,000 사용자.
- **OpenStreetMap:** 세계의 무료지도를 만드는 crowdsourcing 프로젝트.
 - 2 백만 유저로 부터 survey, GPS, aerial photographs 등을 이용해서 데이터를 얻습니다.
- **1000 Genome Project:** 인간 유전자형 및 변이 데이터(2008-2015) 수집
 - 1,000 개의 게놈은 2,504 개체에서 8440 만 가지의 변종을 가지고 있습니다.



DATA PREPARATION

Data Preprocessing

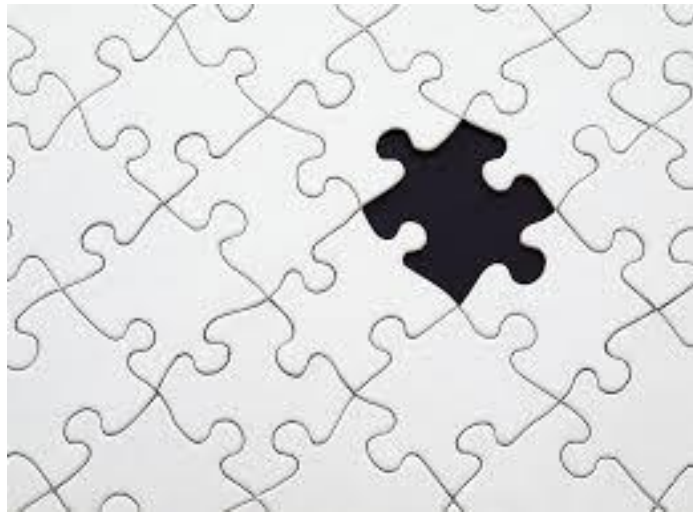
데이터의 품질 및 구조가 종종 분석 할 준비가되지 않아서 데이터 전처리가 필요합니다.

- 웹 스크랩이나 API를 통해 얻은 데이터는 일반적으로 구조화되지 않으며 분석을 위해 수치 행과 열로 조작해야 합니다.
- 이것은 종종 낮은 수준의 데이터 객체 조작을 포함하여 문자열을 숫자 데이터로 또는 그 반대로 변환합니다.
- 데이터 값이 누락되었을 수 있습니다.
- 다른 ML 모델은 다른 데이터 형식을 필요로 합니다.

Missing Values

종종 데이터의 특정 부분이 누락됩니다.

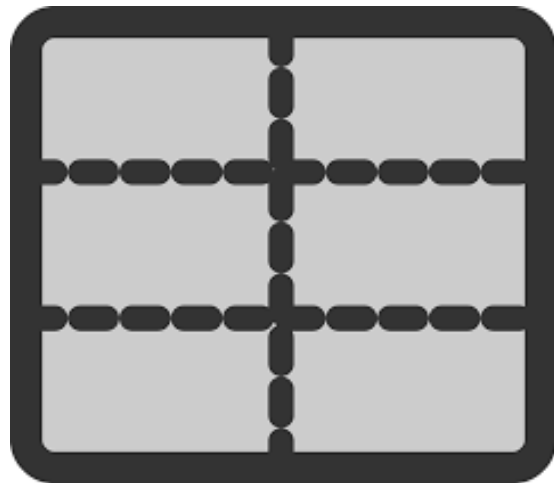
- 누락 된 값을 적당한 기본값으로 바꾸십시오.
- 대칭 분포 기능 : 누락 된 값을 피쳐의 평균으로 입력하십시오.
- 대칭 적이거나 범주 적이지 않으면 중앙값 또는 모드를 각각 입력하는 것이 가장 좋습니다.
- 너무 많은 피쳐가 없는 경우에는 관찰을 중단하는 것이 가장 좋습니다.



Data Preprocessing

데이터를 다르게 포맷해야하는 ML 모델의 몇 가지 예는 다음과 같습니다.

- 기능이 비슷한 스케일을 필요로 하는 모델.
 - 옵션 1 : 최대 값이 1이고 최소값이 0이 되도록 각 기능의 크기를 조정하십시오.
 - 옵션 2 : 평균 및 표준 편차가 각각 1로 되도록 각 기능을 표준화합니다.
- 다른 사람들은 종속물이나 표적을 변형시켜 극단적으로 왜곡되지 않도록 합니다. (예 : 대수를 사용하여)





LABELING DATA

Labeling Data

우리는 supervised learning task 를위해 분류된 데이터를 필요로하지만, 일반적으로 분류되지 않은 데이터는 분류된 것보다 많습니다.

- 데이터는 직원이 손으로 분류 할 수 있지만 비용이 많이 들고 시간이 오래 걸릴 수 있습니다.
- 다른 옵션에는 Amazon Mechanical Turk * 및 반 감독 학습 기술이 포함됩니다.

Amazon Mechanical Turk*

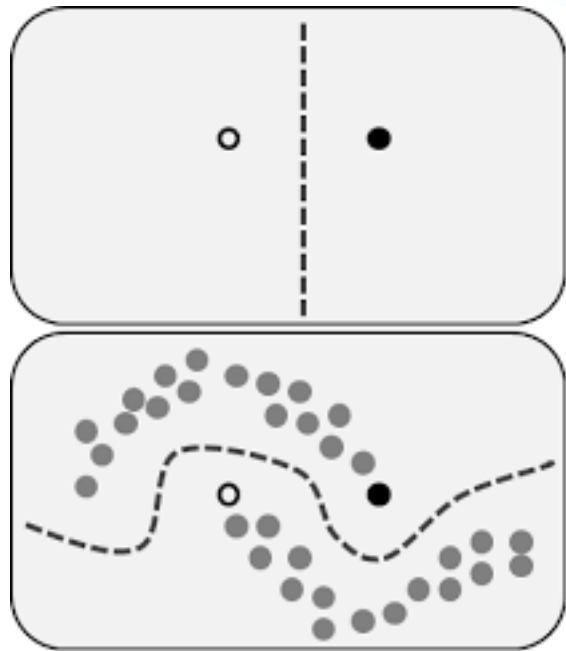
Amazon Mechanical Turk *는 작업을위한 온라인 마켓 플레이스입니다.

- 조직은 웹 사이트에 "Turkers"가 수행하도록 선택할 수 있는 작업을 게시 할 수 있습니다.
- 주요 오픈 소스 데이터 세트는 Amazon Mechanical Turk를 사용하여 작성되었습니다.
- Microsoft COCO (Common Objects in Context) 데이터 세트에는 "Turkers"의 간단한 설명 5 개가 포함 된 이미지가 포함되어 있습니다.

Semi-supervised

Semi-supervised 학습은 레이블이 붙은 데이터와 레이블이 없는 데이터를 모두 사용하여 작업을 해결하는 것을 포함합니다.

- 분류된 데이터에 분류자를 만들고, 분류되지 않은 데이터의 일부를 분류하기 위해 이를 사용하고, 모든 데이터가 분류 될 때까지 반복적으로 진행하십시오.



Learned Divisions



CHALLENGES

Challenges

성능 저하로 이어질 수있는 데이터 수집, 구조 및 사용에는 많은 어려움이 있습니다.

- 이 모든 것이 지나치게 낙관적이고 오도 된 결과 또는 모델 실패로 이어질 수 있습니다.
- 데이터의 편향과 이상치, 부적절한 검증과 테스트,
- 분류 문제의 클래스 불균형.



Biases in Data

알고리즘은 훈련에 사용된 데이터를 반영합니다.

- 편향된 데이터에 대해 교육을 받으면 이러한 편향을 반영합니다.
- 모델러가 존재할 수 있는 편견을 해결하기 위해 데이터를 수집 한 방법을 아는 것이 중요합니다.
- 이것은 일반적으로 분석 단계에서 극복 할 수 없습니다. 특수한 경우에 대한 해결책이 있습니다.

Data Leakage

Data leakage is any time the train/test split is violated.

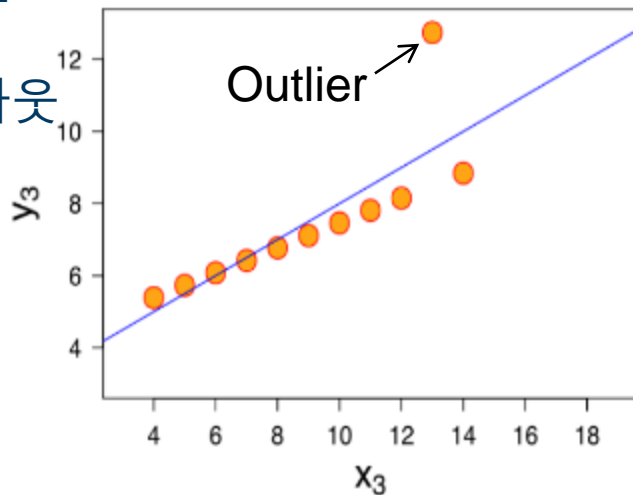
- 모델은 평가 될 때까지 테스트 데이터를 보지 않습니다.
- 데이터 누출은 모델이 보이지 않는 예제를 일반화하는 정도를 과대 평가하는 결과를 가져올 수 있습니다.
- 데이터 누출에 대한 단일 솔루션은 없습니다. 모델러는 유효성 검사 전략을 개발하고 데이터 세트를 분할 할 때주의해야 합니다.



Outliers

Outliers are data points far away from other data points.

- 성능이 실제보다 훨씬 낮다고 제안하여 모델 결과를 왜곡 할 수 있습니다.
- 모델러는 아웃 라이어를 탐지하고 EDA 단계에서 아웃 라이어가 존재하는 이유를 이해해야 합니다.
- 이상치 탐지 방법에는 여러 가지가 있습니다.
- 아웃 라이어와 싸울 수 있는 방법은 다양합니다.
- 종종 이상치는 데이터 세트에서 제거되거나 나머지 데이터 세트의 패턴에 맞게 조정되어야 합니다.
- 특이치에 덜 민감한 모델 및 / 또는 지표를 사용하십시오.



Imbalanced Datasets

불균형 한 수업은 모델을 훈련하고 평가하는 데 어려움을 겪을 수 있습니다.

- 예를 들어 99 %의 레이블이 하나의 클래스에 속하면 항상 다수 클래스를 예측하는 모델은 기본적으로 99 % 정확합니다.
- 이 문제를 해결하기위한 모델링 기법에는 더 큰 클래스를 다운 샘플링하거나 정확성 이외의 채점 척도를 사용하여 모델을 평가하는 방법이 있습니다.

