15기 정규세션

ToBig's 14기 강의자
이혜린

전처리 및 EDA

Contents

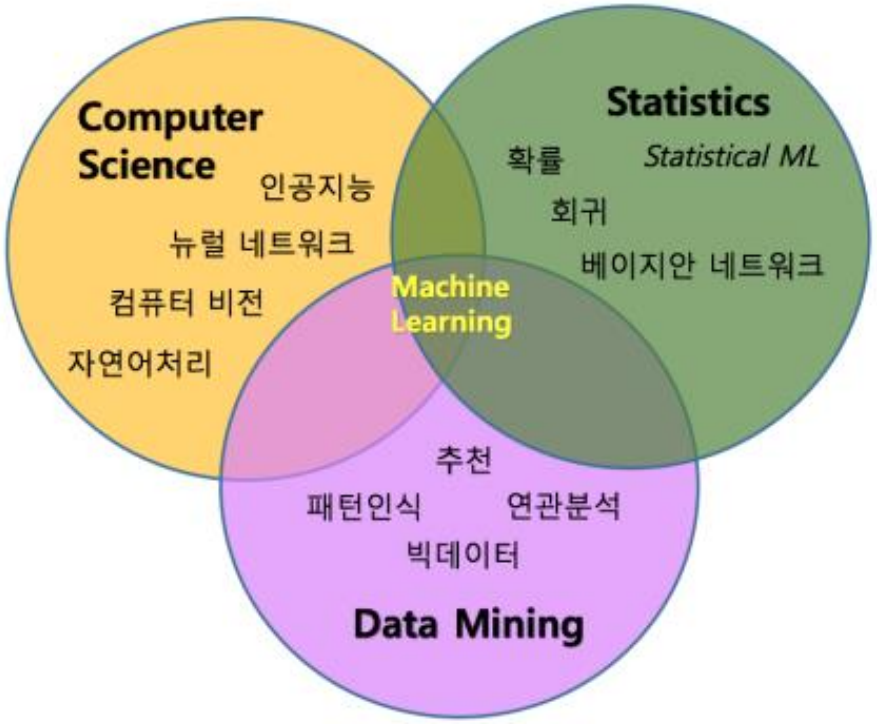
Unit 01 | EDA 란

Unit 02 | 전처리 및 EDA 방법

Unit 03 | 실습 및 과제

Unit 01 | EDA란

✓ 데이터 마이닝, 머신러닝



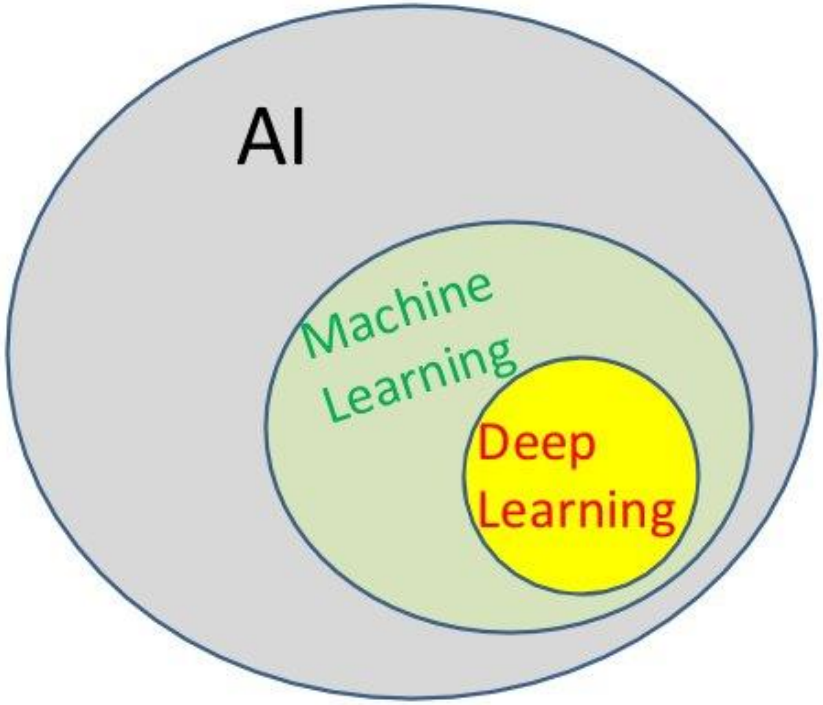
머신러닝 : 컴퓨터가 사전에 미리 프로그램 되어 있지 않고
데이터로 부터 패턴을 학습하여 새로운 데이터에 대해
적절한 작업을 수행하는 일련의 알고리즘이나 처리 과정

데이터 마이닝 : 인간에게 **인사이트를 제공**하는 데 초점

머신러닝 : 학습된 알고리즘으로 **새로운 데이터를 처리**하는 데 초점

Unit 01 | EDA란

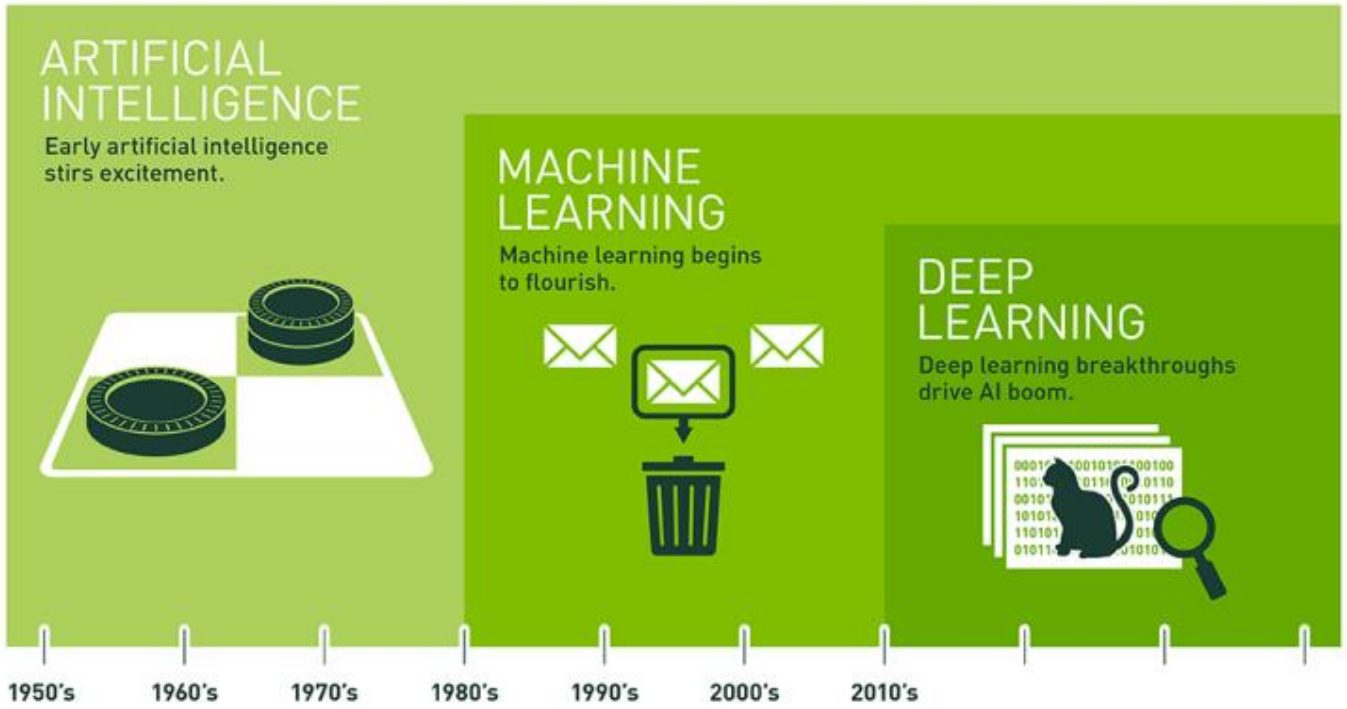
✓ 인공지능, 머신러닝, 딥러닝



딥러닝 : **인공신경망(Neural Network) 알고리즘**을 이용한 머신러닝의 한 분야

Unit 01 | EDA란

✓ 인공지능, 머신러닝, 딥러닝



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Unit 01 | EDA란

✓ 앞으로의 정규세미나 일정

2월 10일은 설날 연휴로 인해
쉬어갑니다!

	강의명
1주차	전처리및EDA(이혜린) 오픈소스와git(정민준)
2주차	회귀분석(강재영) Optimization(김상현) 환경구축(정민준)
3주차	NB&DT(정재윤) KNN&Clustering(김민경)
4주차	SVM(정재윤) Ensemble(장예은)
5주차	차원축소(박준영) 크롤링(강의정)
6주차	NN기초(박지은) 클래스(민거홍)
7주차	NN심화(한유진) 프레임워크(서아라) 시계열분석(이원도)
8주차	CNN기초(이정은) NLP기초(정세영) 강화학습(장혜림)
9주차	모델심화1(장예은) 모델심화2(고경태)
10주차	음성(민거홍) 추천시스템(이원도) 비지도생성모델(김상현)

통계 & 머신러닝

딥러닝 & 강화학습

Unit 01 | EDA란

✓ 앞으로의 정규세미나 일정

2월 10일은 설날 연휴로 인해
쉬어갑니다!

	강의명	
1주차	전처리및EDA(이혜린) 오픈소스와git(정민준)	→ 기본 중의 기본!
2주차	회귀분석(강재영) Optimization(김상현) 환경구축(정민준)	통계 & 머신러닝
3주차	NB&DT(정재윤) KNN&Clustering(김민경)	
4주차	SVM(정재윤) Ensemble(장예은)	
5주차	차원축소(박준영) 크롤링(강의정)	
6주차	NN기초(박지은) 클래스(민거홍)	딥러닝 & 강화학습
7주차	NN심화(한유진) 프레임워크(서아라) 시계열분석(이원도)	
8주차	CNN기초(이정은) NLP기초(정세영) 강화학습(장혜림)	
9주차	모델심화1(장예은) 모델심화2(고경태)	
10주차	음성(민거홍) 추천시스템(이원도) 비지도생성모델(김상현)	

Unit 01 | EDA란

✓ 앞으로의 정규세미나 일정

2월 10일은 설날 연휴로 인해 쉬어갑니다!

모든 과정은 데이터에 대한 충분한 이해가 수반되어야 함

1주차	EDA(이혜린) 오픈소스와git(정민준)	기본 중의 기본!
2주차	회귀분석(강재영) Optimization(김상현) 환경구축(정민준)	통계 & 머신러닝
3주차	NB&DT(정재윤) KNN&Clustering(김민경)	
4주차	SVM(정재윤) Ensemble(장예은)	
5주차	차원축소(PCA)(장예은) 크롤링(강의정)	
6주차	NN기초(박지은) 클래스(민거홍)	딥러닝 & 강화학습
7주차	NN심화(한유진) 프레임워크(서아라) 시계열분석(이원도)	
8주차	CNN기초(이정은) NLP기초(정세영) 강화학습(장혜림)	
9주차	모델심화1(장예은) 모델심화2(고경태)	
10주차	음성(민거홍) 추천시스템(이원도) 비지도생성모델(김상현)	

EDA

Unit 01 | EDA란

✓ EDA란?

Exploratory Data Analysis

탐색적 데이터 분석

데이터를 분석하고 결과를 내는 과정에 있어서

지속적으로 해당 데이터에 대한 탐색과 이해를 기본적으로 가져야 한다는 의미

Unit 01 | EDA란

✓ 올바른 EDA의 흐름

처음 로우 데이터를 접할 때부터 데이터에 대한 충분한 이해 & 파악

데이터에 대한 가설 설정

가설에 따라 여러 feature로 필터링하고 값을 출력

표와 그래프로 시각화 하며 인사이트 도출(사전 검증)

Valuable information 창출

Unit 02 | 전처리 및 EDA 방법

✓ 전처리 (Data pre-processing)

1. 결측치
2. Categorical 변수 처리

Unit 02 | 전처리 및 EDA 방법

✓ 결측치 확인 및 제거

1. 결측치 확인

isnull(), isna()

: Generate a Boolean mask indicating missing values

notnull(), notna()

: Opposite of isnull

Unit 02 | 전처리 및 EDA 방법

✓ 결측치 확인 및 제거

2. 결측치 제거

- 결측치가 있는 행을 모두 제거
- 결측치가 있는 변수 자체를 제거
- 결측치가 있는 변수를 다른 값으로 대체
  - 수치형 변수 : 0, 평균, 최솟값, 중앙값 등으로 대체
  - 범주형 변수 : None, 모름 등의 별도의 범주를 생성하여 대체

Unit 02 | 전처리 및 EDA 방법

✓ 변수 종류 확인

1. Numeric vs Categorical

수치형 변수일 경우

- outlier(이상치) 확인 (ex. sns.pairplot)
- mean, std, range 확인 (ex. dat.describe)

범주형 변수일 경우

- 범주의 개수 확인 (ex. unique)
- 범주별 빈도 확인 (ex. value_counts)

```
dat.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 117 entries, 0 to 116
Data columns (total 15 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0    case_id                  117 non-null    int64
 1   province_x                117 non-null    object
 2   city                      117 non-null    object
 3   group                     117 non-null    bool
 4   infection_case            117 non-null    object
 5   confirmed                 117 non-null    int64
 6   latitude_x                117 non-null    object
 7   longitude_x               117 non-null    object
 8   elementary_school_count   117 non-null    int64
 9   kindergarten_count        117 non-null    int64
 10  university_count          117 non-null    int64
 11  academy_ratio             117 non-null    float64
 12  elderly_population_ratio  117 non-null    float64
 13  elderly_alone_ratio       117 non-null    float64
 14  nursing_home_count        117 non-null    int64
dtypes: bool(1), float64(3), int64(6), object(5)
memory usage: 13.0+ KB
```

✓ 변수 종류 확인

2. Target variable

Classification이 목적일 경우

- 분류의 대상이 되는 feature를 target variable 로 설정. 이 때 target variable은 categorical.
- 분류의 사용할 feature와 그렇지 않은 feature를 필터링

Regression이 목적일 경우

- 다른 변수들에 의해 얼마나 영향을 받을 지에 대해 알고 싶은 feature를 target variable로 설정.
  이 때 target variable은 numeric.
- Target variable에 영향을 줄 것 같지 않은 feature는 필터링

Unit 02 | 전처리 및 EDA 방법

✓ 범주형 변수 처리

범주형 변수를 **숫자형 벡터**로 만들기 위해

Ordinal encoding, One-hot encoding, Dummy variable encoding 등을 사용

Ordinal Encoding	One-hot Encoding	Dummy variable Encoding
[1, 2, 3]	[1, 0, 0] [0, 1, 0] [0, 0, 1]	[0, 0] [1, 0] [0, 1]

Unit 02 | 전처리 및 EDA 방법

✓ EDA

1. Feature engineering

- 도메인에 대한 충분한 이해 후, 데이터에 대한 가설 설정
- 가설에 맞는 파생변수 생성

2. 시각화

- numpy, pandas, seaborn, matplotlib 이용하여 표와 그래프 생성

Unit 02 | 전처리 및 EDA 방법

✓ 유의할 점

결국 데이터에 대해 충분한 시간과 많은 노력을 쏟는다면,

분석 프로젝트에서 좋은 결과를 얻을 수 있음 !

(특히 통계 모델, 그리고 머신러닝 모델의 경우)

결론 : 데이터를 어떤 방식으로든 하나하나 뜯어보자 !!

실습

Unit 03 | 실습 및 과제

✓ 과제

- 데이터 : 과제 데이터.csv
- 파이썬을 이용하여 전처리 및 EDA를 진행해주세요
  - 결측치 처리
  - 유의미한 시각화 10개 이상
  - 수치형 변수 간 상관관계 파악
- 위에서 도출한 인사이트를 통해 유의미한 Feature를 5개 이상 만들어주세요

Unit 03 | 실습 및 과제

✓ 과제 데이터셋 설명

datetime - hourly date + timestamp
season - 1 = spring, 2 = summer, 3 = fall, 4 = winter
holiday - whether the day is considered a holiday
workingday - whether the day is neither a weekend nor holiday
weather - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp - temperature in Celsius
atemp - "feels like" temperature in Celsius
humidity - relative humidity
windspeed - wind speed
casual - number of non-registered user rentals initiated
registered - number of registered user rentals initiated
count - number of total rentals

Unit 00 | 참고자료

투빅스 11기 유기윤님 1주차 EDA 강의자료

투빅스 13기 김현선님 1주차 EDA 강의자료

Kaggle

https://tensorflow.blog/

들어주셔서 감사합니다.