딥러닝 : Multi-layer machine learning

머신러닝 자체는 오래된 기술.

1. KoreaIT 빅데이터 분석 전문가

- Day1

- 20-08-08

빅데이터 개념.

분석방법론

환경셋팅

R프로그래밍 기본 / 기본문법, 데이터타입, 함수

수업방식 : 주어진 명령어를 여러 상황에서 어떻게 사용하고,

동작원리는 어떻고 등을 알아본다

우리가 사용하는 수준 : 엑셀 사용하는 수준

머신러닝 :

Tensorflow. hp -> ?

데이터 엔지니어, IT전문가, ...

대학원수준은 접근 X

기초는 탄탄히 잡고 간다

ppt - 따로 제공

이론 : day1까지만

모집단 분석 - 일반화

표본은 어떻게 추출할건지?

'데이터분석 기획', 시각화

데이터를 알아야 분석 가능

관련된걸 알아야 분석

데이터는 어떻게 수집할것인지?

쿠팡은 제공

제공하지 않는 사이트도 많음

예복습

반복해서 타이핑/실행이라도

크롤링 스크래핑

GB TB PB EB ZB YB

정형/비정형 Data

Pivot Table

Data Vs Informaiton

Graph

데이터 : raw data(관측한, 관측가능한 data),

객관적 사실, 값, 참/거짓

e.g. 엑셀 각 셀의 값

정보 : 가공된 데이터

e.g. 특정 분류를 이용해 묶은 데이터의 집합 & 특정 의미를 가지고 있는 데이터 묶음

e.x. "배송정보" ...

지식 : "정보 이용에 대한 노하우"

Wisdom : 지식 활용에 관한 것

날짜/요일 : "데이터가 아님" 나라마다 다르니까?

일반 IT정보 : 행으로 정리

- 데이터분석은 열 데이터를 분석한다?

- ex) 마트 A, B에서 판매중인 여러개의 각 품목 가격을 비교. 만약 A가 전반적으로 저렴하다면, "마트 A"가 더 합리적인 곳이라 판단"

=> "지식"

지혜 : A마트가 더 저렴하다는 지식은 있으나, 위치 등 기타 변수(정보, 지식)까지 판단 => "지혜"

지식과 지혜가 빅데이터의 영역. 스스로 학습시키고 결과를 출력하는것이 목표!

"예측"

- A마트가 B마트보다 10만개의 품목에 대해서 저렴해도, 어느 하나는 비쌀 수 있다.

현실에서는 항상 일어나는 일.

예측 : 90%만 넘어도 대단한 것

빅데이터 분석을 했을 시 정확도가 너무 높게나온다면 상식을 벗어난 것. 오류가 있을 수 있음.

빅데이터 : 일반적인 DB가 저장/관리/처리가 어려울 수준의 데이터(과거)

지금은 과거와 다름. 막대한 데이터를 저장할 수 있음. / CPU도 빨라짐

처리속도 증가 -> 다양한 데이터 수집

3V(Volume, Variety, Velocity)

- 데이터가 다양한 매체에서 막대한 양으로 생산되고 있음.

- 여러 형태로 데이터가 발생함. 영상, 사진, 통꼐치, 워드, 엑셀, 음성인식, ...

- 데이터 저장/분석/로드 속도 증가

빅데이터/머신러닝으로 매출 극대화 -> 아마존, 넷플릭스

데이터를 통해 성공

'Digital Transformation'

I. 데이터 엔지니어

CRM, EDW, ERP, Social Media, Text, Hadoop ...

- Data engineering

II. 데이터 분석 전문가

1. 데이터 수집

2. 데이터 가공(전처리)

3. 분석 - 시각화 반복(정보가 나올때까지)

4. 예측 및 활용(Ph.D, 능력이 증명된 case)

Modeler가 필요한 알고리즘을 만들어냄

네이버 웹툰 - 채색 AI...?

배달의민족 : AI회사.

Koex 배민식당(서빙 로봇)

라이더스

자동 배달기계

비즈니스 시장 판단 : 빅데이터로도...

명확한 설계를 찾고 방향성을 찾자

------

정형, 반정형, 비정형 데이터

- 구분기준 : 데이터를 한번에 통칭할 수 있는 '열 이름'이 정해져있다면 정형 데이터.

없으면 비정형.

통계청에서는 정형화시켜서 제공.

즉, Categorized data -> 정형 data

반정형 : 정형과 비정형 중간정도.

카테고리는 없지만 key와 value로 데이터를 정리.

- 정형 : 데이터가 Table의 형태를 띔.

- 반정형은 반드시 Table의 형태를 따르지는 않음. 정보마다 데이터 개수가 다를 수 있다!

-> 반정형이든 비정형이든, 정형으로 데이터를 '변경'해야 해석할 수 있다. 이것이 '데이터 가공!'

-> 시간과 노력이 매우 많이 필요하다.

- 난잡한 데이터를 가공해야, 함수/메소드 등을 활용하여 원하는 정보를 추출할 수 있다

- 열로 정리되지 않은 데이터를 1열로 정리하는 것도 일종의 가공! 실제적인 예시.

이것이 우리가 해야하는 것

- 데이터를 새로 쌓는건 상대적으로 쉽다.

- 기존의 데이터를 양질의 데이터로 정리하는건 매우 어렵다.

- 데이터들중에는 잘못된 데이터가 포함될수밖에 없다. 상식범위를 초과하거나 누락된 데이터.

Had oop -> 비, 반정형 데이터 -> 정형 데이터.

사람은 같은 데이터를 오래 볼수록 능률이 떨어짐.

컴퓨터는 빠르고 신뢰도가 높으나 주어진 데이터셋을 제대로 가공하지 않으면 잘못된 결과가 나올 수밖에 없다

- 통계분석

데이터 요약 및 기술하는 기술통계,

데이터에서 전체 집단의 의미를 추정하는 통계적 추론.

- 머신러닝

기계가 데이터를가지고 스스로 예측, 의사결정 할 수있도록 만드는것

지도학습/비지도학습, 딥러닝 강화학습, ...

지도학습보다, 비지도학습이 중요.

Mission-clitical? critical? 때문에 상용화, 신뢰도 측면에서 떨어짐.

- 딥러닝

CNN -> 이미지 분석(거의 대안이 없다)

RNN -> 연속된 텍스트, 연속성 찾기

- Recommendation System

협업필터링(정보 필터링) 기술의 일종

-> 아마존, 넷플릭스

AI ⊃ Big Data ⊃ Machine Learning ⊃ Deep Learning

정보필터링 : B.D.⊃정보필터링

D.L.~ M.L. ⊃ some 정보

- 경계선 찾기는 중요한 M.L.

-> Classification

- 좋은 예측은 무엇인가?

\* 더 잘 나타내는것.

\* 얼마나 합리적인가.

\* 낮은 비용, 낮은 오류

딥러닝 - Theory - Perceptron!

배타 연산(exclusive)

1 0

0 1 -> 해결책 : 머신러닝을 두번. 2개의 Layer.

Multilayer Machinelearning... -> Deeplearning.

(기술의 생존을 위한 작명..!)

뉴런의 구조를 쉽게 만들 수 있는 알고리즘이 텐서플로(구글) -> 인간의 뇌를 효율적으로 모방

어떻게 데이터를 빠르게 쌓을 것인가?

빅데이터 에코 시스템 / Had oop

- 데이터 엔지니어가 관리

Apache hadoop

Java : "Bigdata에서 가장 좋은 언어"

한국에서는 무조건 Java.

이유 :

1. 법으로 Java를 보호 / 공공기관 99.9% 전자정보

2. Hadoop이 Java로 만들어져있음.

단, 분석할 때는 Java가 비효율적.

- R, Python 사용.

R : 데이터분석에 최적화

-> 웹이 구조적으로 사용 불가.

-> 시각화에서 Python과 비교 불가능할 정도로 좋은 ggplot이 있다. Python에도 있으나 R에서 훨씬 효율적으로 동작

Python : 가장 범용. 통신, 타분야 활용 등.

엑셀처럼 사용하겠다 -> R

데이터분석 방법론 요구사항

- 컴퓨터사이언스 / 수학,통계학 / 비즈니스 분석능력

=> 데이터 사이언스

개인이 커버하기 어려움. 현재는 각 분야별 팀으로 구성

R -> 데이터 수집, 정제, 분석, 정보화, 정보활용

Model 만들기 실습

r-project.org, rstudio, java JRE

R-project -> Mac은 한글폰트관련 오류 있음. 설정해야함

프로그램 : 보통 최신버전에서 1~2단계 전버전으로 설치!

인트라넷 :

192.168.7.254/

**-> 어떻게 접속??**

weekend\_bigdata -> 교안... 소스코드

-> Java 8 다운

admin, pw없음

수업자료는 github

조성진, vianimusi@naver.com, IT

연락처 필요하면 멘토 거쳐서.

Wifi - (pw)korea123a~g

1. 자습실이용

- 4층. 주중, 주말 모두 가능. 조교 있을 수 있음.

2. 학원 행사(코로나...)

- 내부행사 : 정수기 위

3. 외부행사(코로나..)

- 단체 외부행사

- 홈페이지 이용(?), 멘토샘한테 질문

- KOEX...

4. 먹을것 : 탕비실.

5. 멘토링 : 학원비에 포함.

피어슨 : 시험이 있는...

토렌트/불법다운로드 관련 소송 기사

R-studio : 설치할 32비트 불필요 -> 에러 원인

R programming : 통계계산 및 그래픽을 위한