암기

연속형 vs 범주형

범주형 변수는 사칙 연산의 결과가 의미가 없다는 특징이 있습니다. 순서 의미의 유무에 따라 명목 변수와 서열 변수로 구분합니다.

연속형 변수는 숫자로 표현되며, 연속된 값을 가집니다. 사칙연산이 가능하다는 특징이있습니다. 연속형 범수는 절대적 기준점의 존재 여부에 따라 등간 변수와 비율 변수로 구분된다.

1장

OLTP vs OLAP

OLTP는 컴퓨터를 통해 데이터베이스를 조회, 갱신 등의 단위 작업을 처리하는 방식을 의미한다.

OLAP 사용자가 관심을 가지는 주제를 중심으로 분석하기 위해서 OLAP는 보고서를 생성하고, 복잡한 데이터 분석을 수행하며, 추세를 식별하는 데 사용됩니다.

OLTP는 데이터의 정확도를 중요하게 생각하는 반면, OLAP는 데이터 읽기 작업에 우선순위를 두고 대용량 데이터를 기반으로 복잡한 쿼리를 빠르고, 효율적으로 수행하는것을 중요하게 생각한다.

빌드 파이프라인 - DevOps팀 분야에서 소스코드 작성에서 배포까지 소프트웨어의 전체 진행과정을 설명하는 것을 빌드 파이프라인이라한다. 위를 일반 개발자 분야에서는 코드 파이프라인이라한다.

블록체인이 인공지능의 발전에 중요한이유

블록체인 네트워크에 올라오는 데이터는 여러 사람들에 의해 검증되고 삭제와 변경이 되지 않는 사실에 기반한 데이터이므로 이런 데이터를 인공지능에 활용하면 보다 더 정확한 미래 분석이 가능하다.

사물인터넷 데이터와 인공지능의 관계

사물인터넷은 인간의 개입 없이 정확하게 생성된 데이터이므로 인공지능에서 활용하면 보다 더 정확한 미래 분석이 가능하다.

클라우드와 인공지능의 관계

인공지능을 이용하기 위해서는 수많은 데이터를 저장할 공간과 컴퓨팅 성능이 필요하고 클라우드는 이를 충족시키는 가장 적합한 플랫폼이기 때문에 인공지능 기술이 발전하면 동시에 이를 적용한 클라우드 서비스또한 발전합니다.

2장

재현성의 의미

측정한 결과가 다시 나타나는 성질이다. 재현성을 동일한 분석 과정에 따른 결과를 보장하는 아주 중요한 요소이다. 재현성이 있다는 말은 누가 따라하든 똑같은 결과를 얻을 수 있다는 걸 의미한다.

마크업과 마크다운의 차이점

마크업

- 마크업은 문서나 데이터의 구조를 정의하기 위해 태그 등을 사용하는 언어이다.
- 주로 HTML, XML 등이 있으며, 웹페이지의 레이아웃 및 스타일을 정의하는 데 사용됩니다.

마크다운

- 마크다운은 텍스트 기반의 마크업 언어로, 쉽게 읽고 쓸 수 있으며 HTML로 변환될 수 있습니다.
- 주로 문서 작성이나 README 파일, 블로그 글 작성 등에 사용됩니다.
- 간결하여 별도의 도구 없이 작성이 가능함, 텍스트로 저장되어 용량이 적음
- 표준이 없어 도구에 따라 변환 방식과 결과가 다름, 모든 HTML 마크업을 대체할 수 없다.

마크업은 문서의 구조를 세밀하게 제어하기 위해서 태그를 사용하고, 마크다운은 간편한 문서 작성을 위해 일반 텍스트 기반의 서식을 지정하는 데 사용됩니다.

파이썬의 장점
언어가 간소해서 좋고 클린한 코드를 작성할 수 있습니다.

또한 파이썬은 풍부한 라이브러리를 가지고 있습니다.

장고, 플라스크를 이용하여 웹 어플리케이션을 개발할 수 있습니다.

3장

폰노이만 구조

폰 노이만 구조 방식은 중앙처리장치(CPU)와 한 개의 메모리를 사용하여 처리하는 현대의 범용 컴퓨터들이 사용하는 구조 모델이다.

폰노이만 그림 →

변수의 할당과 컴퓨터 구조의 관련성을 설명하시오.

폰노이만 구조를 통해서 변수를 메모리에 할당합니다. 메모리는 유한하기 때문에 지나치게 큰 메모리 낭비는 치명적입니다. 따라서 필요 이상으로 크거나 너무 부족한 메모리 할당은 적절치 않습니다. 따라서 변수를 적절한 자료형에 담아서 사용합니다.

수치 데이터 vs 비수치 데이터

수치 데이터는 주로 산술 논리 연산 과정으로 사용되며 주로 정수와 실수로 구성됩니다.
비수치 데이터는 우리가 일상생활에서 표현하는 문장과 같은 텍스트 데이터와 소리, 영상과 같은 멀티미디어 데이터로 구성되어있다.
빅데이터 시대가 도래함에 따라 비수치 데이터에 관심이 생기게 됐다.

4장

정형 데이터

정형 데이터는 미리 정해진 형식으로 구조화된 데이터이다. 행과 열로 표현되는 데이터를 저장하는 엑셀 시트나 RDBMS 테이블이 정형 데이터의 대표적인 예라고 할 수 있다.

반정형 데이터

테이블과 같이 규격화된 형식에 저장되어 있지 않으나 내부 형식을 가지는 데이터

반정형 데이터는 특정한 형식에 따라 저장된 데이터이지만 정형 데이터와 달리 형식에 대한 설명을 함께 제공해야한다. 따라서 구조를 해석하는 파싱과정이 필요하며 파일 형태로 저장된다. XML, JSON등이 대표적인 예이다.

비정형 데이터

규격화된 형식에 저장되어 있지 않은 자유로운 데이터이다. 빅데이터와 인공지능 세상이 시작되며 가파른 증가를 보이는 데이터이다.

데이터소스

미디어, 클라우드, 웹, 사물인터넷, 데이터베이스, 오픈 데이터 등이있다.

미디어 - 구글, 페이스북, 트위터, 유튜브, 인스타그램 같은 **SNS 및 이미지 비디오 오디오 같은 데이터**

정형 및 비정형 데이터를 가지고 있고 비지니스에 실시간 정보를 제공하는 **클라우드**

방대한 양과 사용성을 보장하는 **웹**(인터넷)데이터

인간의 개입 없이 정확한 정보를 생성하는 **사물인터넷** 데이터

ETL과정으로 정형 데이터를 확보하는 **데이터베이스** 데이터

누구나 사용할 수 있는 **공공데이터**

ETL이란? (데이터베이스 중에서...)

Extract, Tranform, Load

추출

추출하는 동안 ETL은 데이터를 식별하고 해당 소스에서 복사하므로 데이터를 대상 데이터 저장소로 전송할 수 있습니다. 데이터는 문서, 이메일, 비즈니스 애플리케이션, 데이터베이스, 장비, 센서, 타사 등 구조적 및 비구조적 소스에서 가져올 수 있습니다.

변환

추출된 데이터는 원래 형식의 원시 데이터이므로 최종 데이터 저장소에 맞게 준비하려면 매핑하고 변환해야 합니다. 변환 프로세스에서 ETL은 결과 데이터를 신뢰할 수 있고 질의할 수 있는 방식으로 데이터를 검증, 인증, 중복 제거 및/또는 집계합니다.

로드

ETL은 변환된 데이터를 대상 데이터 저장소로 이동합니다. 이 단계는 모든 소스 데이터의 초기 로드를 수반할 수 있거나 소스 데이터의 증분 변경 로드일 수 있습니다. 데이터를 실시간으로 또는 예약된 배치로 로드할 수 있습니다.

HTTP 기본 기능과 역할

GET 방식

GET를 통해 해당 리소스를 조회하고 해당 document에 대한 자세한 정보를 가져온다.

GET 방식은 URL를 통해 데이터를 전달하기 때문에 같은 URL를 전달하면 서버에는 항상 데이터를 같이 전달합니다.

속도가 POST 방식보다 빠릅니다.

긴 데이터를 보내면 전체 주소가 길어지게되어서 URL의 길이를 제한해야 한다.

POST 방식

데이터의 제한이 없고 눈에 보이지 않는 방식이 POST 방식입니다. POST를 통해 해당 URL를 요청하면 리소스를 생성합니다.

속도가 GET방식보다 느리다. 주소창에 전송하는 데이터의 정보가 노출되지 않아Get방식에 비해 보안성이 높다.

웹크롤링

웹크롤링은 웹 페이지를 원본 그대로 불러와 웹페이지 내에 데이터를 추출하는 기술이다.

웹소켓을 이용하여 원하는 웹 사이트에 연결 요청을 진행해야한다. 연결 요청의 응답으로 웹 서버가 응답을 보내면 HTML 또는 JSON 형식으로 반환한다. 이렇게 반환된 HTML, JSON 데이터를 BeautifulSoup로 파싱하는 것을 크롤링이라고 한다.

웹크롤링하는 방법

from bs4 import BeautifulSoup as bs

from pprint import pprint

import requests로 임포트해서 라이브러리를 가져온다.

html = request.get(”웹 크롤링 할 주소”)의 반환을 변수에 담고 이를 구문분석(파싱)하여 내가 원하는 웹페이지 내의 데이터를 추출한다.

soup = bs(html.text, ‘html.parser’)로 html소스를 파이썬으로 파싱해 저장한다.

find 함수로 태그 또는 속성에 접근하고 .get_text()를 사용해 데이터를 추출합니다.

파싱

의미는 구문 분석, 즉, 문장을 이루고 있는 구성 성분을 분해하고 그들 사이의 관계를 분석하여 문장의 구조를 결정하는 것이다.

REST API는 HTTP 요청을 통해 통신함으로써 리소스 내에서 레코드(CRUD 라고도 함)의 작성, 읽기, 업데이트 및 삭제 등의 표준 데이터베이스 기능을 수행하는 인터페이스이다.

공공데이터 이용방법

공공 데이터 포털에 접속한다.

1. 회원가입
2. 데이터 검색
3. 오픈 API 활용 신청 후 API 인증키 발급을 확인합니다.
4. api key를 입력한 후 REST API를 요청해서 데이터를 가져옵니다.

파일데이터 vs OPEN API

파일데이터는 공공 데이터 포털에 접속해 간편하게 받을 수 있는 데이터이며 편리한 장점이 있지만 일회성이라는 단점이 있습니다.

API는 Application Programming Interface의 약자로 다양한 응용 프로그램에 사용할 수 있는 운영 체제, 혹은 프로그래밍 언어가 제공하는 기능을 제어할 수 있는 인터페이스 이며 오픈 API는 누구나 사용할 수 있는 API입니다. 장점으로는 지속적인 업데이트가 있지만 이를 위해선 API 프로그램을 개발해야한다는단점이있습니다.

5장

시리즈가 여러개 합쳐지면 어떤 데이터 구조가 되는지 설명하시오.

시리즈는 모든 데이터 유형을 저장할 수 있는 1차원 레이블이 지정된 배열이고 시리즈가 여러개 합쳐지면 행과 열 구조로 이루어진 데이터프레임이됩니다.

데이터 사이언티스트에게 행과 열의 증가가 어떤 의미인가

행의 증가는 처리할 데이터의 양이 증가해 이를 위해서는 컴퓨터의 파워 또한 증가해야 함을 의미한다.

열의 증가는 데이터의 특성, 분석 대상이 증가했다는 의미이며 이를 통한 다양한 해석이 가능하다.
하지만, 너무 많은 변수는 분석을 어렵게 할 수있어 주성분 분석과 같은 변수 축소를 고민해봐야한다.

행 데이터

loc - 인덱스 기준으로 행 데이터를 읽는다. 인덱스 이름이 '1번' : '4번' 이면 인덱스가 '1,2,3,4번'인 행을 읽는다.

iloc - 행 번호를 기준으로 행 데이터를 읽는다. 행 번호 [0:3]을 읽으려고 하면 행번호 0,1,2인 행을 읽어온다.

data.**drop(인덱스) -** 인덱스를 기준으로 삭제

인덱스와 행 번호에 대해 설명하시오.

데이터프레임에서 0열에 해당하는 부분이 인덱스이다. 인덱스는 문자열이나 숫자로 지정할 수 있다.

행 번호는 보이지 않는 일련번호이다.

결측치 처리단계

1. 결측치를 확인한다. → df.isnull(), df.isnull().sum()
2. 결측치를 제거 또는 대체한다. →df[열값].dropna(), axis=1하면 열이 삭제됨,
   df.fillna(df.mean())
   df[고치려는 열].fillna(값, inplace=True)

결측치 제거는 결측치를 포함한 행, 열을 삭제하는 것이고, 대체는 평균 값같은 대푯값으로 변환하는것

3. 결측 데이터 작업 후 반영이 되었는지 확인한다.

이상 데이터 처리단계

1. 이상 데이터 확인

2. 결측 데이터로 대체 or 제거

   단순 삭제 or 다른값으로 대체 or 변수화, 리샘플링, 케이스 분리분석 등이있다.

3. 이상 데이터 처리 확인

filt = df[’열’] > max

df.dropna(index=df[filt].index) #이상치 제거하기

```
plt.boxplot(df['열']) # 이상치 확인하기
plt.show() # 이상치 확인하기

QR1 = df.열.quantile(q=0.25) # 박스 Q1
QR3 = df.열.quantile(q=0.75) # 박스 Q3
QR2 = QR3 - QR1
max = QR2 + (QR3 * 1.5)
min = QR2 - (QR1 * 1.5)

filter1 = df['열'] > max
filter2 = df['열'] < min
df.dropna(index=df[filter1].index) # 매우 큰 이상치 제거
df.dropna(index=df[filter2].index) # 매우 작은 이상치 제거
plt.boxplot(df['열']) # 처리 확인하기
plt.show() # 이상치 처리 확인하기
```

중복 데이터 처리 단계

1. 중복 데이터 확인
   df.duplicated(열 값)
2. 중복 데이터 처리 - 유일한 1개 키만 남기고 나머지 제거하기
   DataFrame.drop_duplicates()
   df.drop_duplicated([열 값], keep=’first) # 처음만 남기고 삭제
3. 중복 데이터 처리 확인
   df.duplicated(열 값)

7장

파이썬 재구조화 함수를 설명하시오.

데이터를 구간화하는 pd.cut(), pd.qcut()함수

데이터를 전치해주는 T함수

데이터 형태를 재구성하는 pivot(), pivot_table()함수

열, 행을 전환하는 melt() 함수

행, 열 인덱스를 전환하는 stack(), unstack()함수

범주형 데이터를 원핫인코딩해주는 pd.get_dummies

데이터 구간화를 설명하고 판다스 구간화 방법을 설명하시오.

연속형 변수를 일정 구간으로 변형하여서, 연속형 변수를 범주형 변수로 만드는 것을 데이터 구간화라고한다.

pd.cut()은 전체 데이터의 구간을 동일한 길이로 나눠서 범주화 해주고 pd.qcut()은 전체 데이터를 동일한 개수로 맞추고 나눠서 범주화 한다.

데이터 피봇 중 df.pivot()과 df.pivot_table()의 차이를 설명하시오.

피봇 테이블이란 많은 양의 데이터에서 필요한 자료만 뽑아 새롭게 데이터를 재구성한다.

첫번째 인수로는 행 인덱스로 사용할 열 이름, 두번째 인수로는 열 인덱스로 사용할 열 이름, 마지막으로 데이터로 사용할 열 이름을 지정한다.

pivot()과 pivot_table() 모두 데이터 프레임의 형태를 재구성할 때 사용하지만, pivot()의 인자는 index, columns, values 이고 pivot_table의 인자는 index, columns, values, aggfunc이다. 즉, pivot_table은 데이터를 요약하고 집계하는 기능을 가지고 있다.

**** pivot은 중복된 index와 column을 처리하지 못하고 pivot_table은 처리할 수 있다.

시계열 데이터란 무엇인지 설명하시오.

시간의 흐름에 따라 불확실성을 가지고 변하는 시간 변수를 시계열 데이터라 한다. 시간 변수의 흐름(x축)에 따른 종속변수(y축)의 움직임을 이해하고 예측하는 것을 목표로하는 분석 방법을 사용한다.

실습

이상치 제거하기

```
qr1 = df['tip'].quantile(q=0.25)
qr3 = df['tip'].quantile(q=0.75)
qr2 = qr3 - qr1
max = qr2 + (qr3 * 1.5)
min = qr2 - (qr1 * 1.5)

filter1 = df['tip'] > max
filter2 = df['tip'] < min
x = pd.concat([filter1, filter2], axis=1) # 필터링 여부 확인

df.loc[filter1 , 'tip'] = round(max, 2) # 너무 큰 이상치를 박스플롯
df.loc[filter2 , 'tip'] = round(min, 2) # 너무 작은 이상치를 박스플
```