AI TRACK

모델 실습: 회귀

1. 데이터 분석이란 ?

: 데이터 분석 단계

2. Boston Housing 데이터를 활용한 회귀 모델 실습

: Linear Regression를 활용한 회귀 예측 모델 생성

강연자 이름
최강규

SeCurious

중앙대학교
산업보안학과 교육봉사동아리

23/02/13 *Mon*

목차

01 데이터 분석이란?

02 예측 : 회귀 (Regression)
VS 분류(Classification)

03 선형회귀(Linear Regression)

04 데이터 설명: Boston Housing

05 실습: 보스턴 주택 가격 예측 회귀 모델

데이터 분석(Data Analysis)

: 데이터 간의 또는 내의 패턴 또는 관계를 찾기 위해 실행

- 데이터의 성격을 파악하고 데이터 간의 패턴 또는 관계를 찾아내서 더 효과적인 모델을 만들기 위해 필수적
- 패턴과 관계는 변수들 사이에 존재함

데이터 분석의 필요성

01 월마트의 "맥주와 기저귀" 사례

맥주와 기저귀는 단순 연관성이 없음

〈 고객들의 구매내역 분석 결과 〉

아이가 있는 20~ 30대의
남성 퇴근 시간:저녁 6시~8시

기저귀를 구매하는 경우 맥주도
함께 잘 구매하는 현상 발견

=> 기저귀와 맥주를 가까이 진열하여 구매를 빠르고
편리하게 하여 구매율 상승시킴

데이터 분석 단계

데이터 분석이란?

01

문제 정의 및 계획 01

문제가 명확해야 문제 해결을 위한 데이터를 추정하고, 어떤 분석기법을 사용할지 계획 할 수 있음

02 원시 데이터 수집

엑셀 파일, 종이 문서, 크롤링 등의 방식으로 필요한 자료를 수집

데이터 전처리 03

수집된 데이터는 바로 분석에 사용하기 어려움

04 탐색적 데이터 분석 (EDA)

데이터 특징과 내재하는 구조 알아내기 위한 통걔적 기법

데이터 수집 -> 시각화 탐색 -> 패턴 도출 -> 인사이트 발견

모델 & 알고리즘 05

머신러닝 기술 등 사용

06 결과 보고

데이터 제품 또는 의사결정 위한 의사소통 보고서

지도 학습

옳고 그름이 있어서 그것을 학습시키는 것

예측할 변수가 존재 ↔ 예측할 변수가 존재 X

결과를 알고 있는 데이터 사용 / 분류화가 되지 않은 데이터 사용

비지도 학습

옳고 그름이 없고 패턴을 파악하는 것

회귀 (Regression)

회귀 - 수치형 데이터 사용

▶ 수치형 데이터: 연속형, 이산형

독립변수(X)와 종속변수(y) 사이에 함수 관계를 파악하여 연속된 값을 예측하는 문제

▶ 대표 알고리즘: Linear Regression(선형 회귀), RandomForestRegression, XgboostRegression

회귀와 분류 둘 다 결국 예측을 위한 것.

분류 (Classification)

분류- 범주형 데이터 사용

▶ 수치형 데이터: 서열형, 명목형

독립변수(X)와 종속변수(y) 사이에 함수 관계를 파악하여 연속된 값을 예측하는 문제

▶ 대표 알고리즘: Decision Tree, RandomForestClassification, XgboostClassification

회귀와 분류 둘 다 결국 예측을 위한 것.

Linear Regression

03 선형회귀

회귀의 대표 알고리즘

단순 선형회귀

1

한개의 독립변수를 사용하여
종속변수를 예측
- f(x) = Wx + b (W: 가중치,
b: 편향성)

다중 선형회귀

2

여러개의 독립변수를 사용하여
종속변수를 예측
f(x, y, z) = W0 + W1x + W2y+ W3z

선형회귀의 목표

학습을 통하여 최적의 가중치와
편향성을 찾아냄

3

선형회귀 장점

4

1. 상대적으로 학습이 빠름
2. 쉽고 간단한 설명 가능
3. 데이터가 많고 적음에
따른 큰 영향이 없음

Linear Regression

03 선형회귀

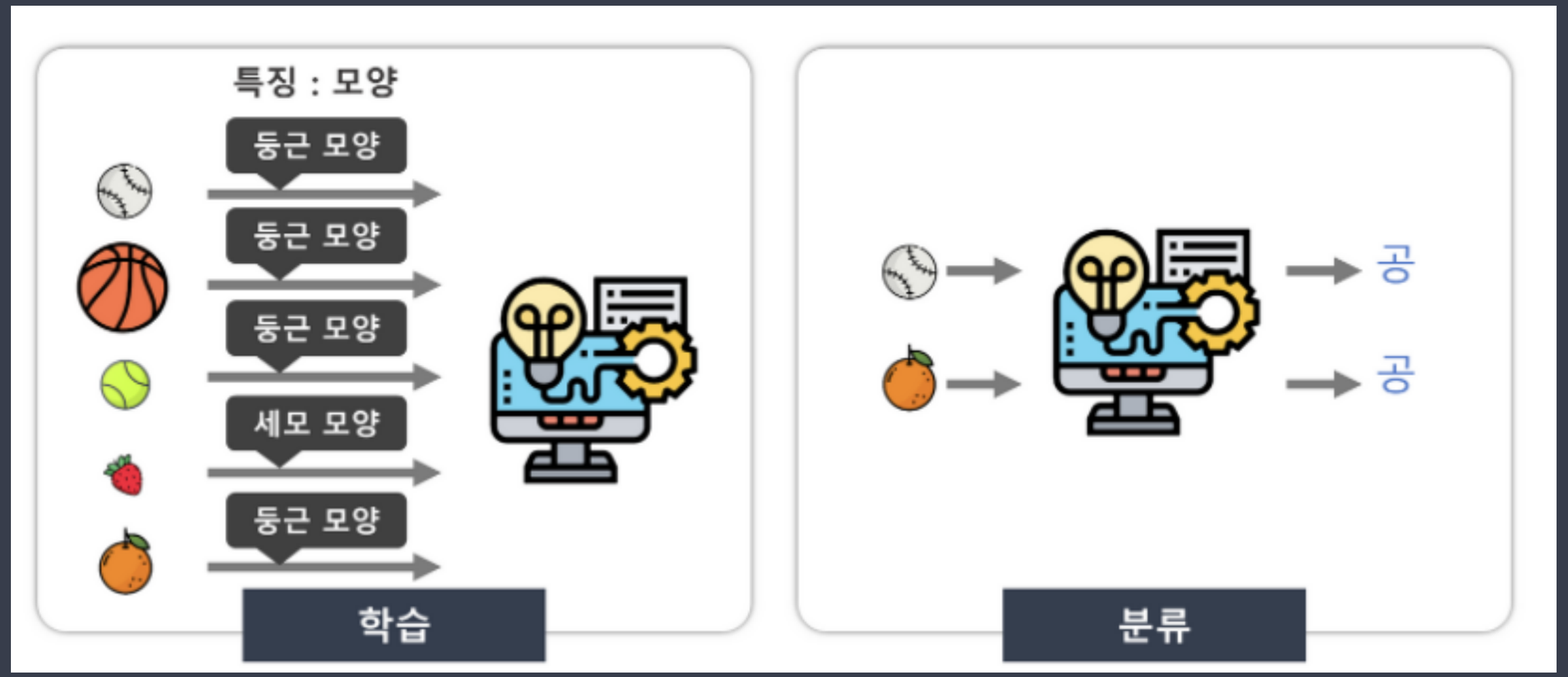
회귀의 대표 알고리즘

선형회귀의 예시

- 부모의 키와 자녀의 키의 관계 조사
- 연령에 따른 실업률 예측
- 공부시간과 중간고사 점수간의 관계
- 키에 따른 몸무게 예측

과소 적합 (Underfitting)

: 모델 학습 시, 충분하지 못한 특징만으로 학습되어, 특정 특징으로 편향되게 학습되는 것

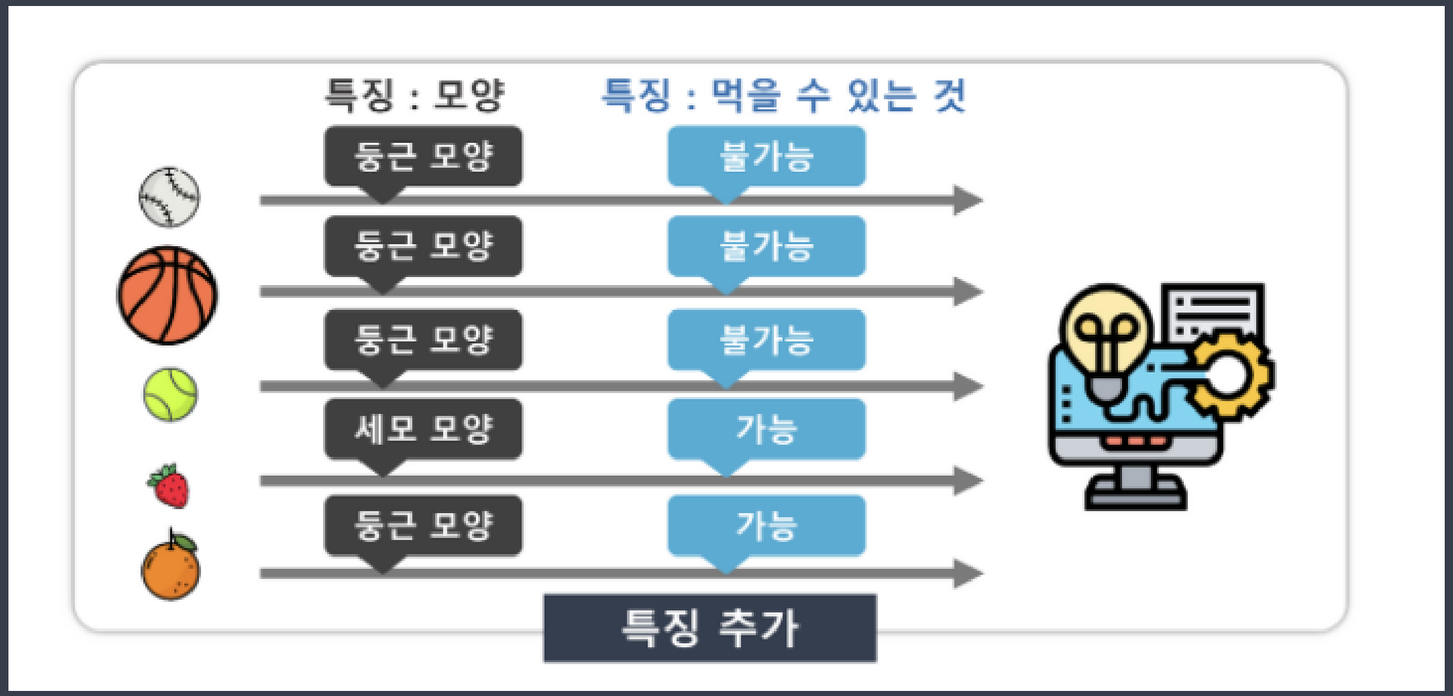


둥근 모양을 학습시키려 했지만, 데이터로 전부 공을 사용했기 때문에 둥근 모양의 어떤 물체를 모두 공이라 학습하게 됨

과소 적합 (Underfitting)

: 테스트 데이터 뿐만 아니라 학습 데이터에 대해서도 정확도가 낮게 나올 경우 과소적합된 모델

=> 학습에 사용된 특징의 개수를 늘리는 것을 통해 개선



둥근 모양을 학습시키려 했지만, 데이터로 전부 공을 사용했기 때문에 둥근 모양의 어떤 물체를 모두 공이라 학습하게 됨

과대 적합 (Overfitting)

: 학습 데이터에 대한 정확도는 매우 높지만 테스트 데이터 정확도가 낮게 나옴

=> 너무 많은 특징들을 알려줘서 학습 데이터에 해당하는 것만 정확도가 높음

04
데이터 설명 :Boston Housing

```
from sklearn import datasets

boston = datasets.load_boston()
```

보스턴 주택 가격: 1978년 발표된 데이터로 미국 보스턴 지역의 주택 가격에 영향을 미치는 요소들을 정리함.

- CRIM: 자치시(town) 별 1인당 범죄율
- ZN: 25,000 평방피트를 초과하는 거주지역의 비율
- INDUS:비소매상업지역이 점유하고 있는 토지의 비율
- CHAS: 찰스강에 대한 더미변수(강의 경계에 위치한 경우는 1, 아니면 0)
- NOX: 10ppm 당 농축 일산화질소
- RM:주택 1가구당 평균 방의 개수
- AGE: 1940년 이전에 건축된 소유주택의 비율
- DIS: 5개의 보스턴 직업센터까지의 접근성 지수
- RAD: 방사형 도로까지의 접근성 지수
- TAX: 10,000 달러 당 재산세율
- PTRATIO: 자치시(town)별 학생/교사 비율
- B: 1000(Bk-0.63)^2, 여기서 Bk는 자치시별 흑인의 비율을 말함.
- LSTAT: 모집단의 하위계층의 비율(%)
- CMEDV: 본인 소유의 주택가격(중앙값) (단위: $1,000)

보스턴 주택 가격 예측하기

예시 파일 참조

1. 문제 정의 및 계획

보스턴 지역 주택 가격에 영향을 미치는 요소들을 살펴보고,
보스턴 지역 주택 가격을 회귀 방식으로 예측하여 실제값과의 오차 비교해보기

2. 원시 데이터 수집

생략

3. 데이터 전처리

데이터 결측값, 중복값 확인

4. EDA

가설 설정 후 시각화, 패턴 도출, 인사이트 얻기
가설 설정: 사회적 통념 상 흑인 비율이 많은 지역에 범죄율이 높을 것이라 보기에 해당 부분을 확인하고, 범죄율이 높을 경우 주택가격이 하락하기에 흑인 비율이 높은 지역의 주택 가격이 낮을 것이다

5. 모델 & 알고리즘

1) Train, Test 데이터 셋 나누기

```
from sklearn.model_selection import train_test_split
feature_columns = list(df.columns.difference(['Target']))
X = df[feature_columns]
y = df['Target']
X_train, X_test, y_train, y_test = train_test_split(X,y, random_state=42)
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

Pandas

import pandas as pd
데이터 구조를 제공하는 Python 패키지

Matplotlib

import matplotlib.pyplot as plt
데이터를 시각화하는 Python 패키지

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Numpy

import numpy as np
수학, 과학 연산을 위한 Python 패키지

Seaborn

import seaborn as sns
데이터를 시각화하는 Python 패키지,
matplotlib보다 더 예쁘게 시각화 가능

Thank you

강의 열심히 들어주셔서 감사합니다 :)

강연자 이름
김중앙

SeCurious

중앙대학교
산업보안학과 교육봉사동아리

23/01/01 Sun