숭실대학교 Soongsil University

선형회귀를 이용한 보스톤 집값 예측

숭실대학교 베어드교양대학
서유화 교수
yhsuh@ssu.ac.kr

목차

- 보스톤 집값 데이터 읽어오고 정리하기
- 데이터를 이용한 단순 선형회귀 모델 생성과 학습
- 생성한 단순 선형회귀 모델 분석과 시각화
- 다중선형회귀 모델 분석 및 시각화

학습목표

- 지도학습의 선형회귀의 원리를 설명할 수 있다.
- 보스톤 집값 데이터를 이용해 선형회귀 모델을 생성할 수 있다.
- 생성된 모델의 결과를 분석하고 시각화할 수 있다.
- 생성된 모델을 통해 데이터로부터 새로운 인사이트를 발견할 수 있다.
- 생성된 모델 이용해 새로운 데이터를 예측할 수 있다.

지난시간 배운 내용

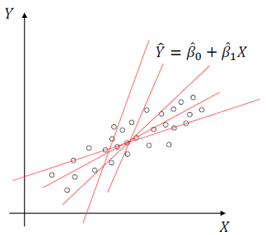
주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성,) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 왜곡에 따른 분석과 예측 결과 비교	1. 데이터 편향성이 예측에 미치는 영향 (구글 티쳐블 머신) 2. 데이터 왜곡에 따른 예측 결과 비교 - 타이타닉호 생존자 예측
5	인공지능과 알고리즘 윤리	1. 알고리즘 기반 의사결정 시스템의 한계 2. 윤리가 적용된 인공지능 알고리즘	1. 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	

선형회귀 (온라인)

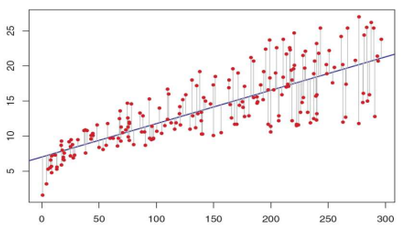
- **단순선형회귀**
  - 변수가 1개인 경우
  - 실제 데이터의 식 : $Y = \beta_0 + \beta_1 X + \varepsilon$
    - β_0는 절편, β_1는 기울기이며 합쳐서 회귀계수(coefficients)로도 불림
  - 우리 추정해야 하는 식 : $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X$
  - 즉, 선형회귀는 학습데이터를 이용해서 회귀계수 $\widehat{\beta_0}$, $\widehat{\beta_1}$를 추정하는 작업
- **다중선형회귀**
  - 변수가 여러 개인 경우
  - 우리가 추정해야 하는 식 : $\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \cdots + \widehat{\beta_p} X_p$

선형회귀 (온라인)

- **선형회귀의 목표**
  - 직선과 데이터의 차이가 평균적으로 가장 작아지는 직선 찾기



- **어떻게 찾을까?**
  - 잔차(residual) : 실제 값과 추정한 값의 차이
  - 잔차를 최소화하는 방향으로 직선 추정
    - 일반적으로 잔차의 제곱합을 최소화하는 방향으로 추정



1. 데이터 읽어와서 필요한 데이터 가져오기


```
import pandas as pd
import statmodels.formula.api as smf  # 다양한 통계 분석 모듈 제공
```

1) 데이터 파일 읽어오기

- boston = pd.read_csv('파일경로명' , sep= ',' , encoding= '인코딩방식')
  - sep 옵션은 생략하면 ',' 로 인식
  - encoding 옵션은 생략하면 'utf-8' 로 인식

2) 원본데이터에서 필요한 열데이터만 가져오기

- boston_data = boston[['Target' , 'CRIM' , 'RM' , 'LSTAT']]

2. Target~CRIM 단순선형회귀 분석

1) target과 crim 선형회귀모델 만들고 학습시키기

- model1 = smf.ols(formula= ' Target~CRIM ' , data=boston_data).fit()

  # ols(ordinary least square) 가장 기본적인 회귀방법 : 잔차제곱합(실제값과 예측값의 차이)을 최소화하는 회귀직선 모델

2) 분석 요약 보기

- model1.summary()
- 주요계수

(P-value) 0.05보다 작고 0에 가까울수록 유의미한 변수

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	24.0331	0.409	58.740	0.000	23.229	24.837
CRIM	-0.4152	0.044	-9.460	0.000	-0.501	-0.329

(절편)

(기울기) 회귀계수 범죄율이 1 증가 할 때마다 집값은 0.4152단위 만큼 감소

y값에 대해 x변수의 설명력은 15% 0~1사이의 값

R-squared:	0.151
Adj. R-squared:	0.149
F-statistic:	89.49

2. Target~CRIM 단순선형회귀 분석

3) 회귀계수 보기

- model1.params

4) 학습데이터에 대해 생성한 회귀모델로 예측하기

- pred1=model1.predict(boston_data['CRIM'])

```
# boston_data[ 'CRIM' ]을 x값으로하는 y값 산출
# 일반적으로 괄호에 X데이터 값을 쓰나, 여기서는 X데이터 값을 생략하고, prdict()로만 써도 가능
```


3. 시각화 하기

1) 적합시킨 직선 그리기

```
import matplotlib.pyplot as plt
plt.scatter(boston_data[ 'CRIM' ], boston_data[ 'Target' ],
            label='data' , color='red')  # 학습데이터
plt.plot(boston_data['CRIM'], pred1,  label="result", color='red')  #예측직선
plt.legend()  # 범례출력
plt.show()
```

2) 실제집값과 예측집값 나타내기

```
plt.scatter(boston_data[ 'Target' ], pred1)
plt.xlabel('real_value')
plt.ylabel('pred_value')
plt.show()
```


3. 시각화 하기

3) 잔차(residual) 보기 (실제값과 예측값의 차이)

```
model1.resid
```

4) 잔차그래프 그리기

```
model1.resid.plot()  #잔차(residual)가 균일할 수록 좋음
plt.xlabel("residual_number")
plt.show()
```

5) 잔차의 합보기

```
sum(model1.resid)   #잔차의 합은 0에 수렴, 잔차의 합이 0이 되도록 회귀 직선을 만듦
```


4. 스스로 해보기

1) RM(주택당 방수)에 대해 회귀분석하기

2) LSTAT(인구 중 하위 계층 비율)에 대해 회귀분석하기

3) CRIM(범죄율), RM(주택당 방수), LSTAT(인구 중 하위 계층 비율) 중 변수 설명력이 가장 좋은 변수는 무엇인가?

5. Target~CRIM + RM + LSTAT 다중선형회귀 분석

1) target과 crim 선형회귀모델 만들고 학습시키기

▸ multi_model = smf.ols(formula= ' Target~CRIM + RM + LSTAT ' ,
data=boston_data).fit()

2) 나머지 과정은 단순선형회귀와 동일

다음시간에 배울 내용

주	주제	온라인	오프라인
1	인공지능의 과거 현재와 미래	1. 강의 및 교과목 소개(공통, 핵심만) 2. 인공지능의 과거와 현재 3. 인공지능의 미래와 다양한 시선 4. 인공지능 개발환경 구축과 사용법(Anaconda/Colab)	1. 강의 및 교과목 소개(분반별, 자세히) 2. 다양한 인공지능 기술 경험하기 (자연어처리, 시각, 음성,) 3. 인공지능 챗봇만들기(IBM 왓슨 어시스턴트)
2	공공데이터를 이용한 사회문제 발견과 해결책 모색	1. 빅데이터의 정의와 가치 2. 공공데이터 수집하기 3. 공공데이터로부터 새로운 인사이트 발견하기 - 행정구역별 인구 데이터와 공공의료기관 현황 데이터 분석	1. 서울시 CCTV설치 현황 분석하기 2. 서울시 범죄발생 현황 분석하기
3	인공지능의 개요 및 머신러닝을 이용한 예측	1. 인공지능의 정의와 분류 2. 인공지능 학습방법 이해하기 3. 인공지능 알고리즘 소개	1. 머신러닝을 이용한 이미지 식별(구글 티쳐블 머신) 2. 머신러닝을 이용한 보스톤 집값 예측
4	인공지능과 데이터 윤리	1. 데이터의 불완전성과 결함에 따른 예측 오류와 차별 2. 데이터 왜곡에 따른 분석과 예측 결과 비교	1. 데이터 편향성이 예측에 미치는 영향 (구글 티쳐블 머신) 2. 데이터 왜곡에 따른 예측 결과 비교 - 타이타닉호 생존자 예측
5	인공지능과 알고리즘 윤리	1. 알고리즘 기반 의사결정 시스템의 한계 2. 윤리가 적용된 인공지능 알고리즘	1. 알고리즘에 따른 예측 결과 비교 - 보스톤 집값 예측 - 폐암환자 생존 여부 예측
6	인공지능에 대한 다양한 이슈와 우리의 자세 고찰	1. 인공지능의 윤리적/법적 쟁점 (자율주행자동차, AI로봇, 트랜스 휴먼 등) 2. 인공지능시대 사회, 경제적 불평등 문제 3. 인공지능과 프라이버시 4. 인공지능의 윤리적 대응과 규제	1. 자율주행 자동차의 행동학습 시나리오 경험하기 2. 비윤리적 데이터 생성과 수집(웹 크롤링을 이용한 데이터 수집)
7		기말고사	