
파이썬 기반의
빅데이터 처리 및
분석 기술
Crayon

빅데이터 회귀 분석

3 빅데이터 회귀 분석 심화

- 최적의 모델 선택
- Scikit Learn 다항식 회귀 모델
- 학습 곡선

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

" 모델의 성과 개선 방법 "

1	더 복잡하거나 더 유연한 모델 사용
2	덜 복잡하거나 덜 유연한 모델 사용
3	더 많은 훈련 표본 수집
4	각 표본에 특징을 추가하기 위해 더 많은 데이터 수집

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

" 모델의 성과 개선 방법 "

예상했던 결과가 나오지 않을 수도 있다.

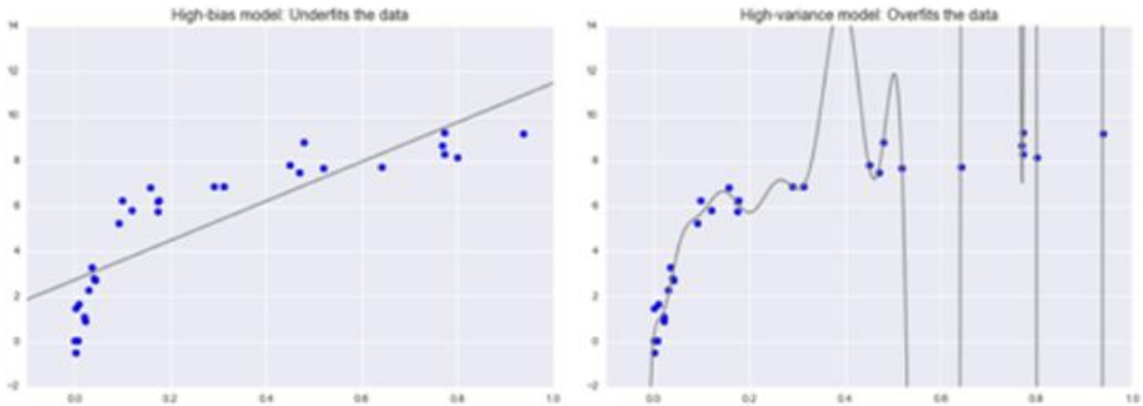
더 복잡하고, 더 많은 데이터를 사용해도 정확하지 않을 수 있다.

최소의 노력으로 최대의 개선을 끌어낼 수 있어야 한다.

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

“편향 – 분산 트레이드 오프”

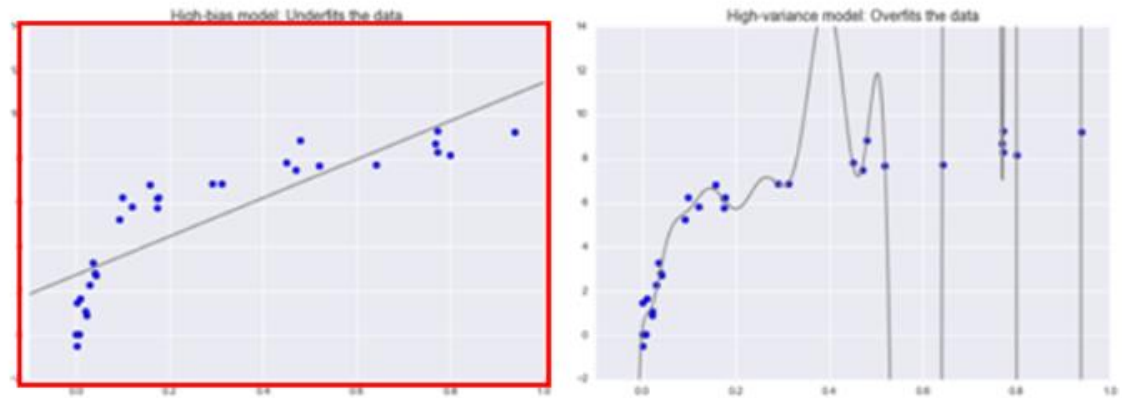


서로 다른 방식으로 인해 결과 실패 초래

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

“편향 – 분산 트레이드 오프”



데이터가 직선보다 복잡하게 변동하므로

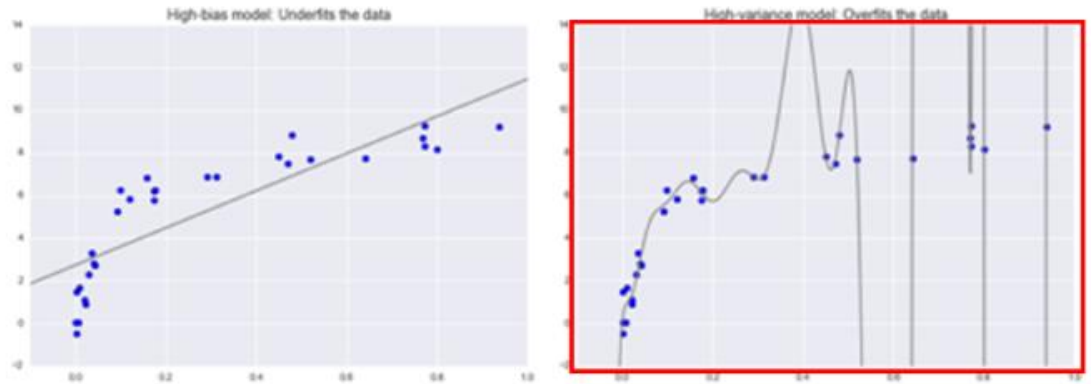
선형 모델로는 데이터 세트 설명 불가

모델이 고편향되었음

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

“편향 – 분산 트레이드 오프”



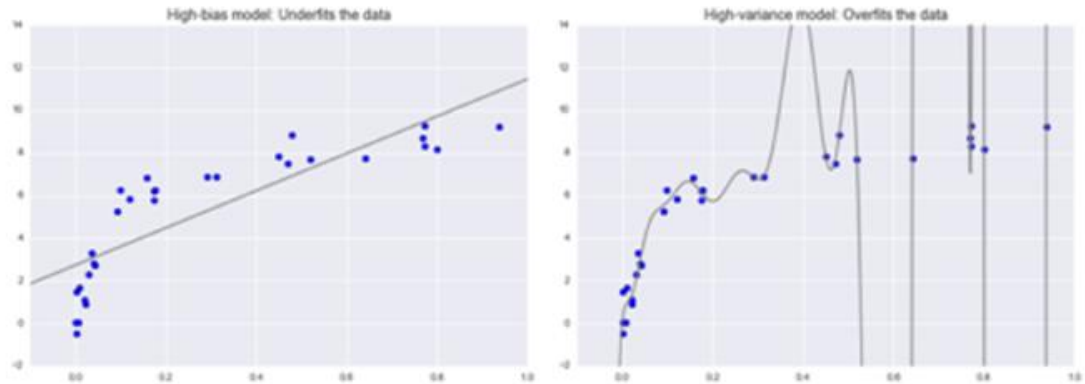
기존 훈련 데이터에 대해서만 적합할 가능성이 높음

모델이 고분산을 가지고 있음

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

" 편향 – 분산 트레이드 오프 "



예상했던 결과가 나오지 않을 수도 있다.

더 복잡하고, 더 많은 데이터를 사용해도 정확하지 않을 수 있다.

최소의 노력으로 최대의 개선을 끌어낼 수 있어야 한다.

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

" 편향 – 분산 트레이드 오프 "

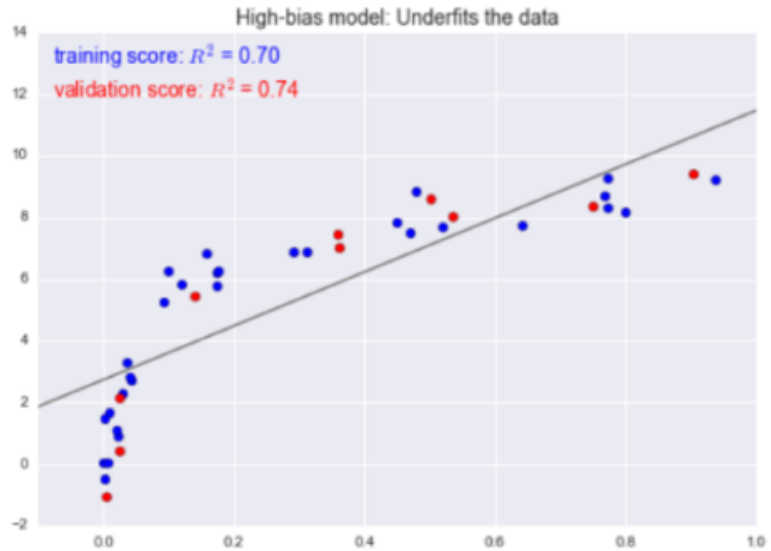
R^2 값 : 결정 계수, 회귀 모델의 정확도 표현

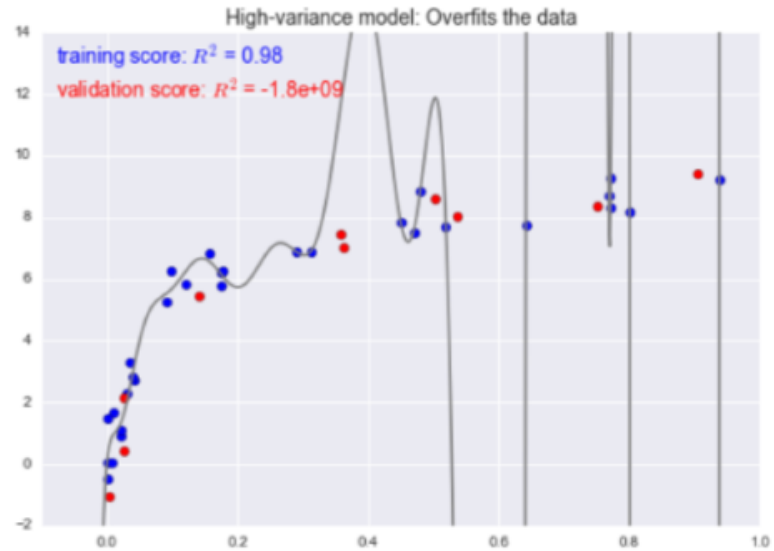
결정 계수 1 : 완벽 일치, 결정 계수 0 : 단순 평균을 구하는 수준

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

" 편향 – 분산 트레이드 오프 "



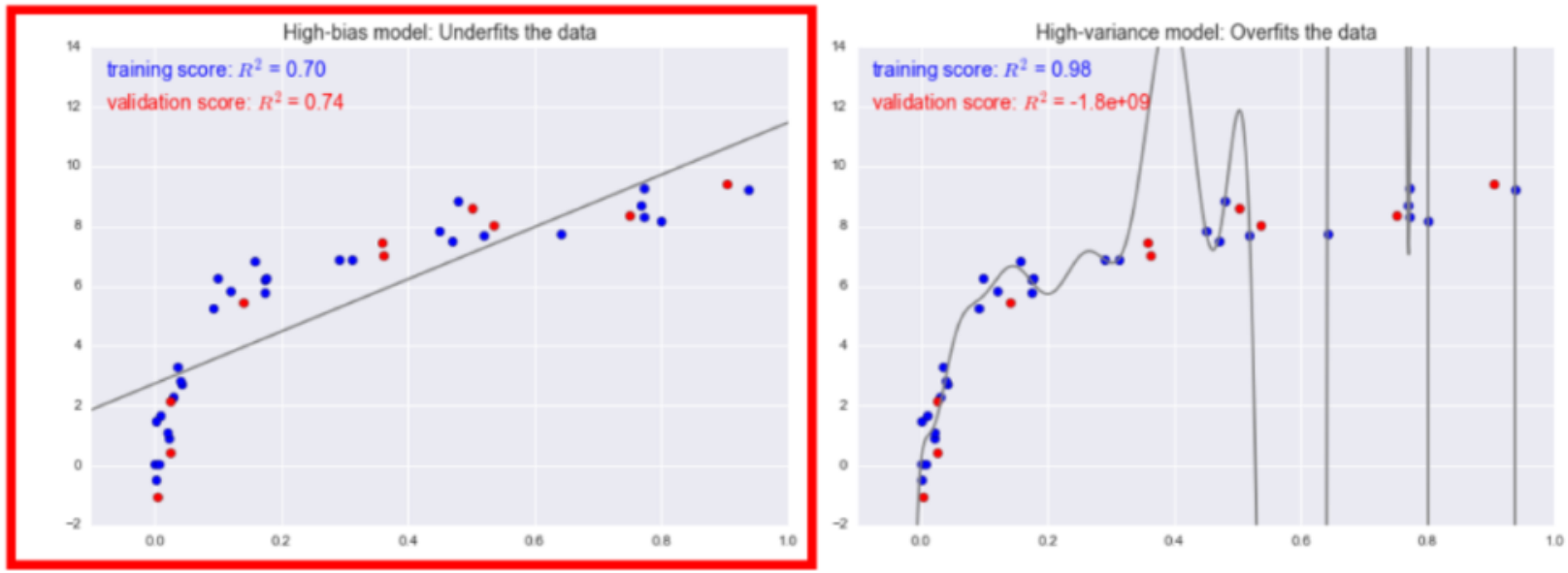


검정 : 데이터에 대한 모델의 정확도를 측정하는 과정

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

" 편향 – 분산 트레이드 오프 "

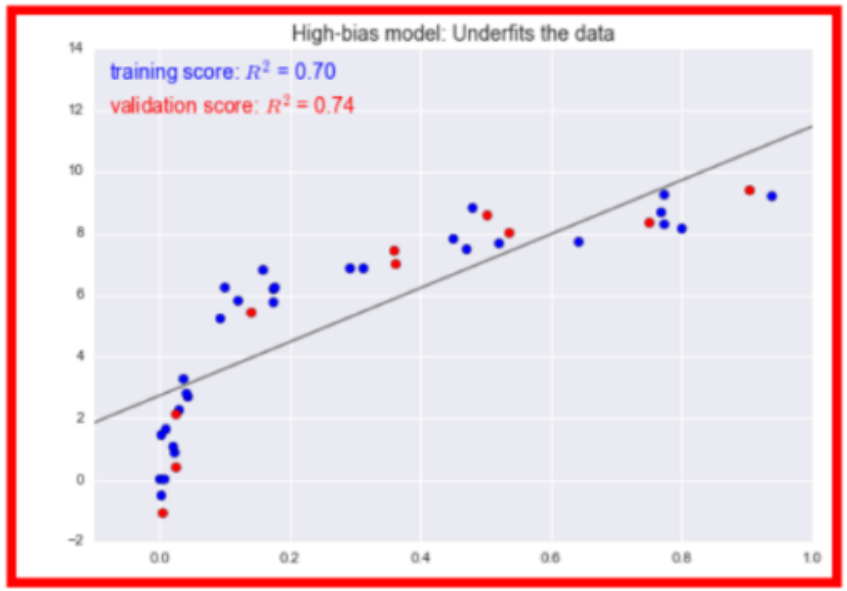


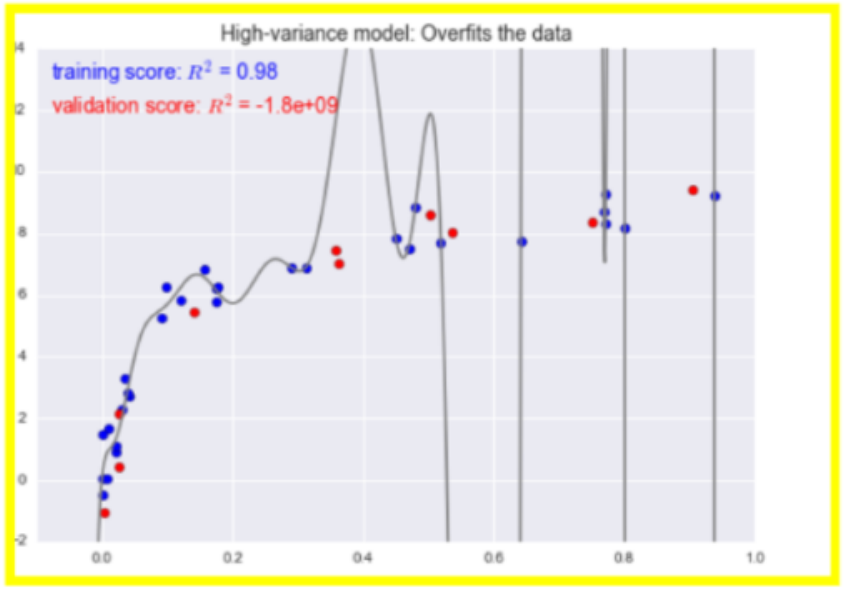
고편향 모델 : 검정 표본에서의 모델 성능이
훈련 표본에서의 성능과 유사

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

" 편향 – 분산 트레이드 오프 "





고편향 모델 : 검정 표본에서의 모델 성능이
훈련 표본에서의 성능과 유사

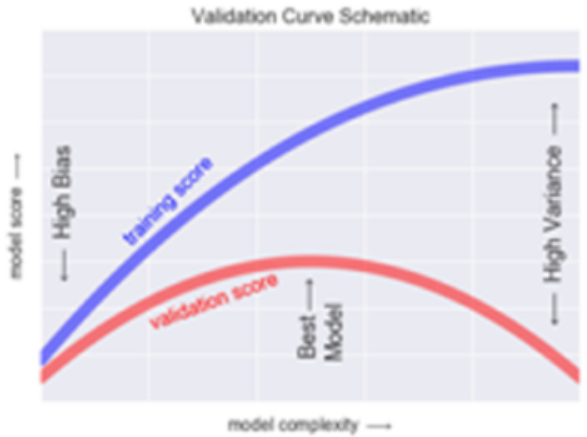
고분산 모델 : 훈련 표본에서는 모델 성능이
우수하지만 검정 표본에서는 성능 저하

1. 빅데이터 회귀 분석 심화

1) 최적의 모델 선택하기

“편향 – 분산 트레이드 오프”

- 모델의 복잡도를 조정할 수 있을 때



고편향과 고분산이 절충되어 검정 점수가 가장 높은 정도의 모델 복잡도가 나오는 것이 바람직함

1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

1차 선형 회귀 모델

$$y = ax + b$$

3차 다항식 회귀 모델

$$y = ax^3 + bx^2 + cx + d$$

다항식의 차수가 높아질수록 모델의 복잡도가 높아짐

1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

“ 파이프 라인 ”

- 전처리 프로그램과 선형 회귀 모델을 묶어주는 프로그램

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.linear_model import LinearRegression
from sklearn.pipeline import make_pipeline

def PolynomialRegression(degree=2, **kwargs):
  return make_pipeline(PolynomialFeatures(degree), LinearRegression(**kwargs))
```

1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

" 데이터 생성 "

```
import numpy as np

def make_data(N, err=1.0, rseed=1) :
  rng = np.random.RandomState(rseed)
  X = rng.rand(N) ** 2
  y = 10 - 1. / (X + 0.1)
  if err > 0:
    y += err * rng.randn(N)
    return X, y

X, y = make_data(40)
```

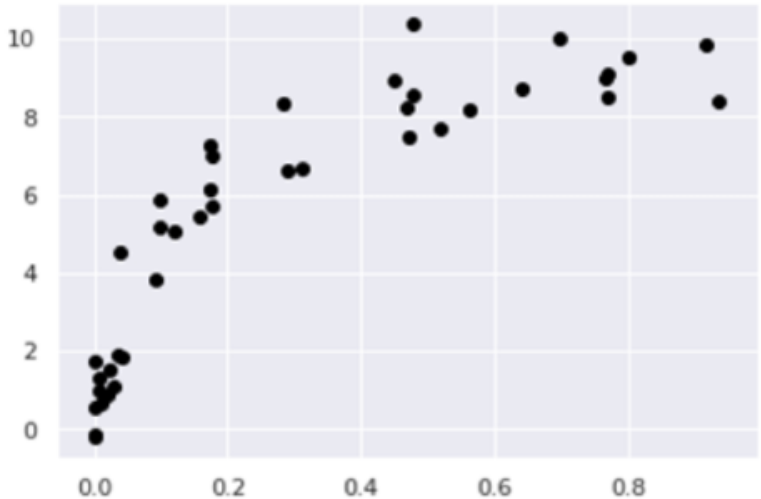
1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

“데이터 시각화”

- %matplotlib inline
  import matplotlib.pyplot as plt
  import seaborn; seaborn.set()

  X_test = np.linspace(-0.1, 1.1, 500)
  plt.scatter(X, y, color='black’) axis = plt.axis()



1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

"데이터 적합"

```
plt.scatter(X, y, color='black')

for degree in [1, 2, 3]:
    model = PolynomialRegression(degree)
    model.fit(X[:, np.newaxis], y)
    y_test = model.predict(X_test[:,np.newaxis])
    plt.plot(X_test, y_test, label=f'Degree={degree}')

plt.axis([-0.1, 1.0, -2, 12])
plt.legend(loc='best');
```

1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

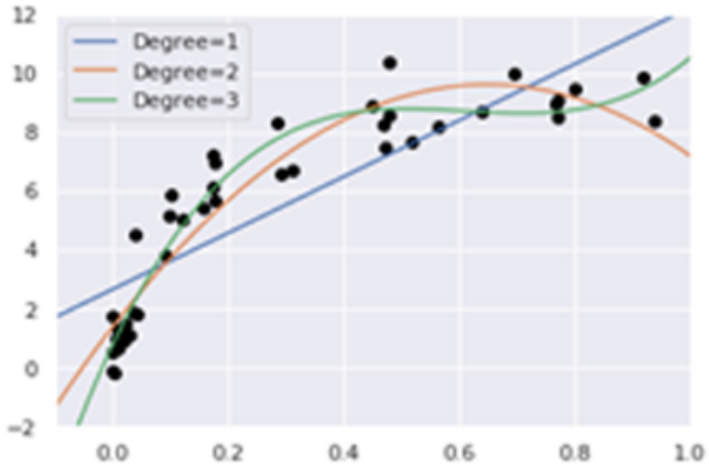
"데이터 적합"

```
plt.scatter(X, y, color='black')

for degree in [1, 2, 3]:
    model = PolynomialRegression(degree)
    model.fit(X[:, np.newaxis], y)
    y_test = model.predict(X_test[:,np.newaxis])
    plt.plot(X_test, y_test, label=f'Degree={degree}')

plt.axis([-0.1, 1.0, -2, 12])
plt.legend(loc='best');
```

1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

“ 데이터 적합 결과 시각화 ”



과소적합과 과적합 사이에 적절한 트레이드 오프를 제공하는 것은
몇 차 다항식인지 찾는 것

1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

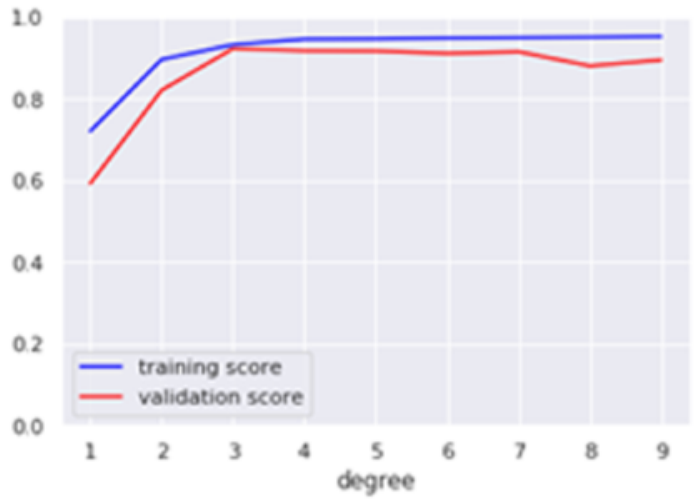
“ 데이터 검증 곡선 ”

```
from sklearn.model_selection import validation_curve
degree = np.arange(1, 10)

train_score, val_score = validation_curve(PolynomialRegression(),
  X[:, np.newaxis], y, 'polynomialfeatures__degree', degree, cv=7)

plt.plot(degree, np.median(train_score, 1), color='blue', label='training score')
plt.plot(degree, np.median(val_score, 1), color='red', label='validation score')
plt.legend(loc='best')
plt.ylim(0, 1)
plt.xlabel('degree');
```

1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

“최적 모델”



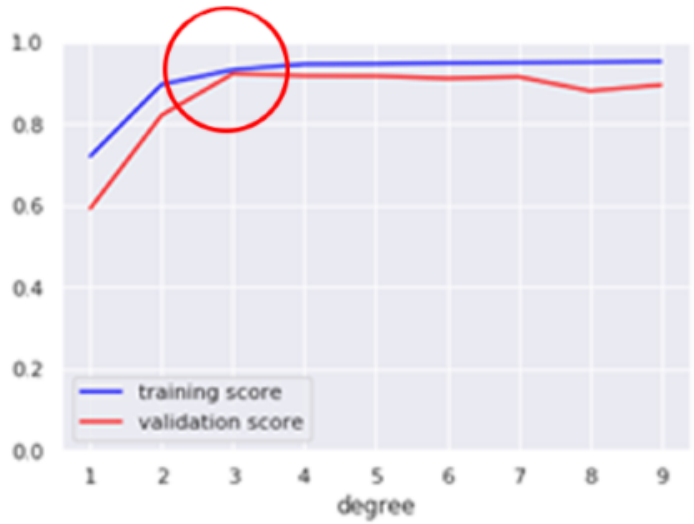
1~3차식 : 훈련 점수와 검정 점수가 함께 상승

4차식 이상 : 검정 점수 하락, 과적합 발생

1. 빅데이터 회귀 분석 심화

2) 다항식 회귀 모델

"최적 모델"



1~3차식 : 훈련 점수와 검정 점수가 함께 상승

4차식 이상 : 검정 점수 하락, 과적합 발생

이번 시간에는

3 빅데이터 회귀 분석 심화

- 최적의 모델 선택
- Scikit Learn 다항식 회귀 모델
- 학습 곡선

이번 시간에는

실습 참고 자료

- Colab 노트북 파일
- Matplotlib 공식 사이트
  → https://matplotlib.org/tutorials/index.html

이번 시간에는

과제 안내

- 과　　제 : 퀴즈
- 제출 방법 : 과제 게시판 제출 방법 안내 참조

질의 응답 게시판

- 학습 내용, 퀴즈, 과제 등에 대한 질의응답
  게시판을 통한 질의응답

다음 시간에는

7 빅데이터 분류 분석

- 빅데이터 분류 분석의 절차
- 초모수와 모델 검증 방법
- 분류 심화 : 나이브 베이즈 기법