3.1 선형 모델

- 사용 데이타셋
  - make_forge() : 분류용 인위적 데이타셋
  - make_wave() : 회귀용 인위적 데이타셋
  - load_breast_cancer() : 분류용 실제 데이타셋
  - load_boston() : 회귀용 실제 데이타셋

- k-최근접 이웃 알고리즘(k-nearest neighbors algorithm) 적용한 분류(cell 10 ~)
  - 모델 생성 : 훈련데이타를 저장하기만 하면 됨.
  - 예측 : 훈련데이타셋에서 가장 가까운 포인트(들)중 다수로 결정

- k-최근접 이웃 알고리즘(k-nearest neighbors algorithm) 특징
  - 이웃이 적을 수록 모델의 복잡도가 높아지고 많이 선택할수록 모델이 단순해짐(cell 17).
  - 비교적 양호한 예측성능 보임(cell 18 cancer example).
  - 데이터가 많아지면 예측속도가 늦어짐.

3.1 선형 모델

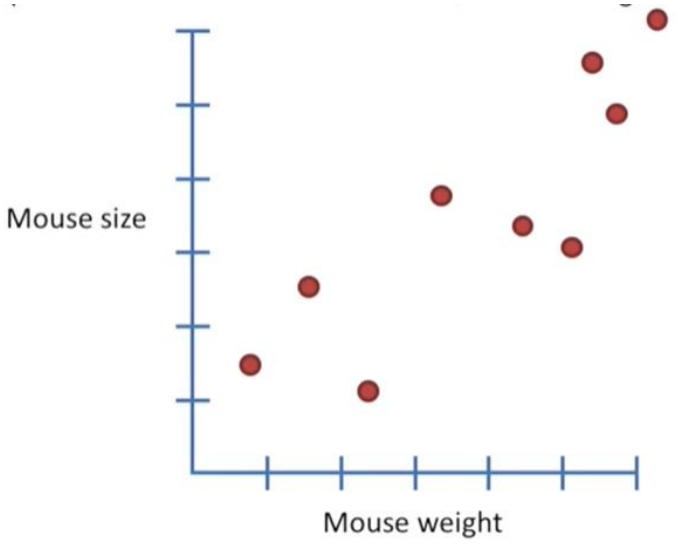
- 입력특성에 대해 선형함수를 만들어 예측을 수행.
- 선형회귀모델은 100여년 전에 개발된 모델임.
- 정답(y) - 예측(y^)의 차이를 MSE(mean Squared Error)를 최소화하는 w,b를 찾음.
- 훈련데이타로부터 모델파라미터 w, b를 학습하여 모델을 완성
  - w : 가중치/계수(weight, coefficient)
  - b : 편향/절편(bias, intercept, offset)

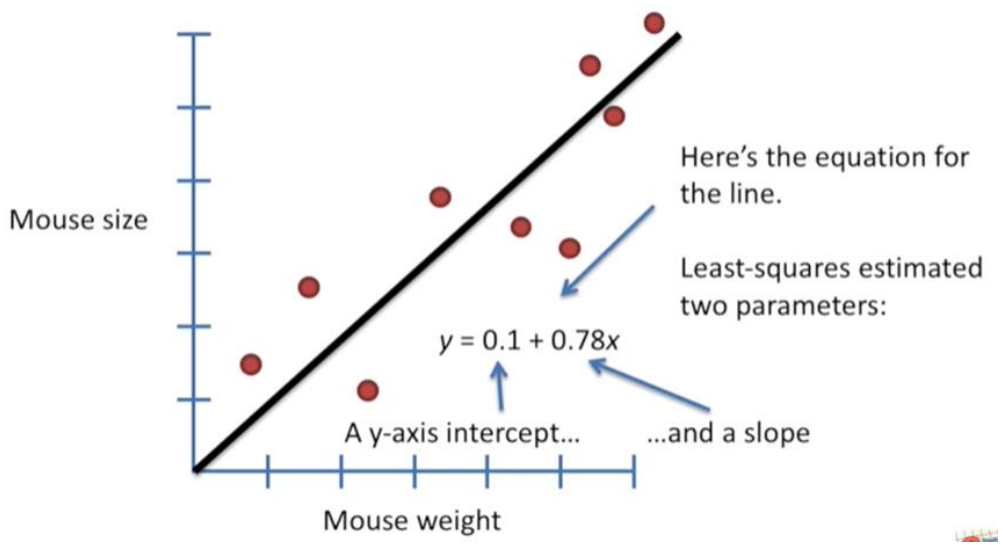
$$y = w_1x_1 + w_2x_2 + w_3x_3 + \ ... + b$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

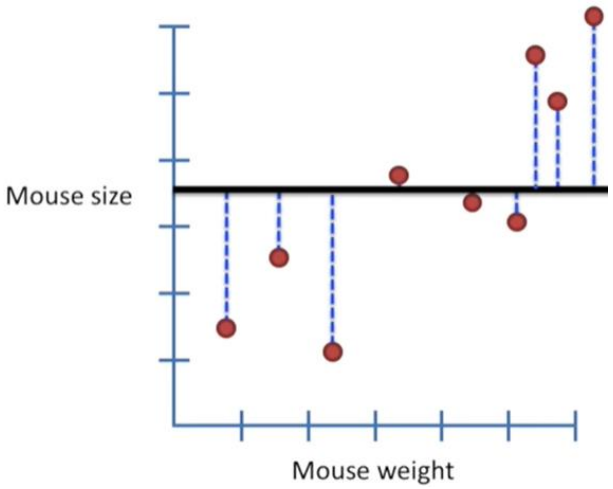
$\hat{y}_i$: Predicted output value

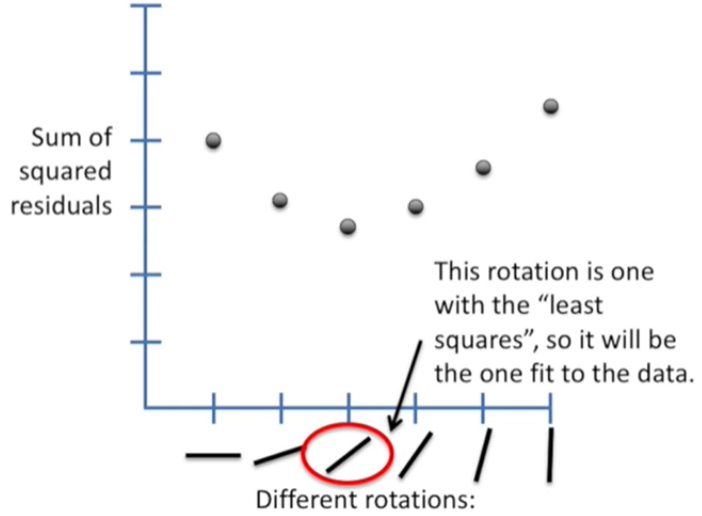
y_i : Actual (expected) output value

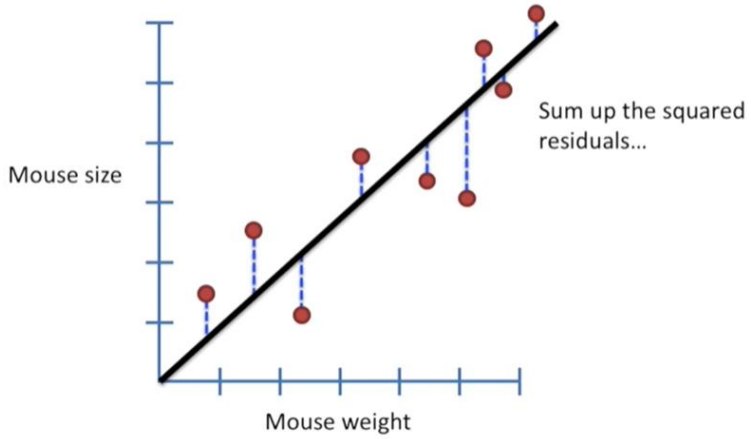


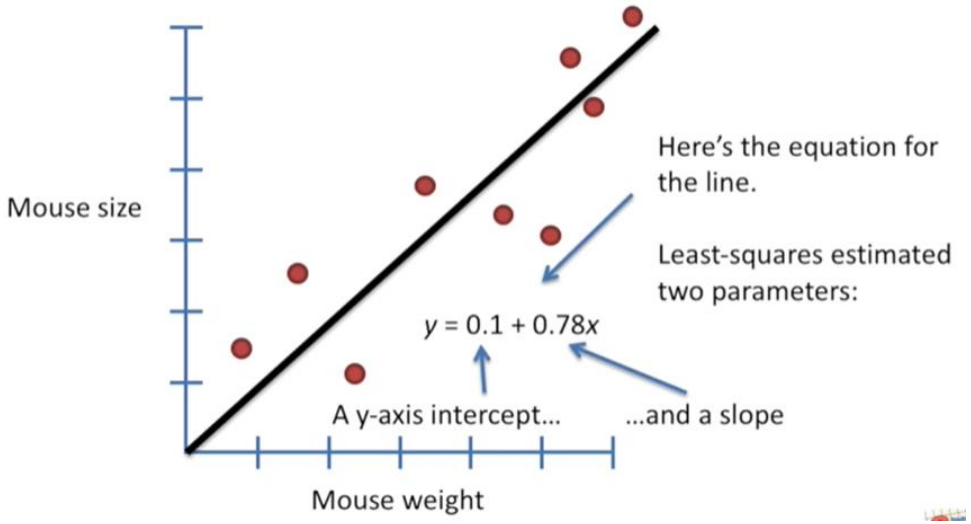


3.1 선형 모델의 원리(Least Square)









https://www.youtube.com/watch?v=nk2CQITm_eo&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=10

3.1.1 선형 모델 실습

- 가공의 데이터를 선형회귀로 예측(cell 26~)
  - 훈련세트 점수(R2) : 0.67
  - 테스트세트 점수 : 0.66
  - 과소적합 상태임(모델이 너무 단순)
  - 그러나 더 이상 개선할 수 있는 수단이 없음.

- 보스턴 주택가격 예측을 선형회귀로 구현(cell 29~)
  - 훈련세트 점수(R2) : 0.95
  - 테스트세트 점수 : 0.61
  - 과대적합 상태임(모델이 너무 복잡)
  - 모델의 복잡도를 제어하는 기법 필요.

3.1.2 과대적합,과소적합

- 일반화(generalization) - 학습을 통하여 새로운 문제를 최적의 성능으로 예측하는 모델
- 과대적합(overfitting) - 학습문제를 맞추는 능력은 우수하나 새로운 문제의 예측성능이 눈에 띄게 떨어지는 상태
- 과소적합((underfitting) - 학습문제를 맞추는 능력과 새로운 문제의 예측성능이 비슷한 상태

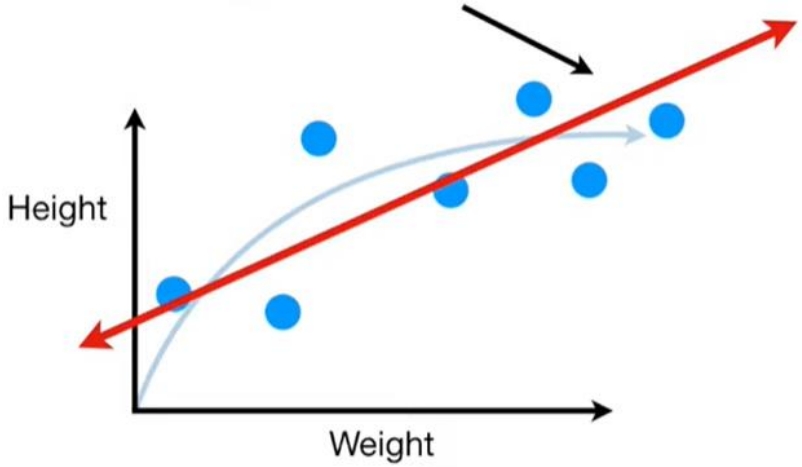
$$
\begin{aligned}
MSE &= E[(y-\hat{y})^2] \\
&= E[(y-E[\hat{y}]+E[\hat{y}]-\hat{y})^2] \\
&= E[(y-E[\hat{y}])^2]+E[(E[\hat{y}]-\hat{y})^2]+E[2(y-E[\hat{y}])(E[\hat{y}]-\hat{y})] \\
&= E[(y-E[\hat{y}])^2]+E[(E[\hat{y}]-\hat{y})^2]+2(y-E[\hat{y}])(E[\hat{y}]-E[\hat{y}]) \\
&= (E[\hat{y}-y])^2+E[\hat{y}^2]-E[\hat{y}]^2 = Bias[\hat{y}]^2+Variance[\hat{y}]
\end{aligned}
$$

- 예측오차(MSE)는 분산 + 편향으로 분해될 수 있음.
  - 편향(Bias) : 추정값- 참값 차이의 기대값.
  - 분산(Variance) : 추정값들이 흩어진 정도. 참값여부는 관계없음.

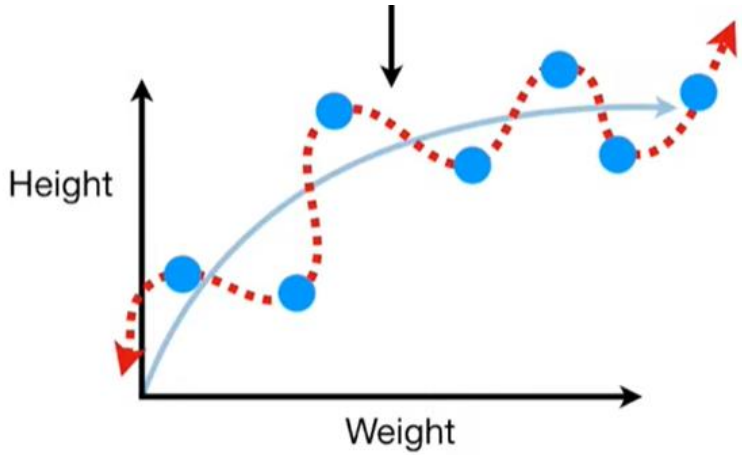
3.1.2 과대적합,과소적합

< 데이터 셋 >

< 선형회귀 모델 >

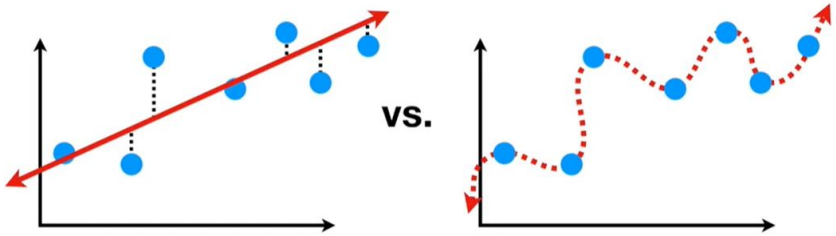


< 과대적합 모델 >



3.1.2 과대적합,과소적합

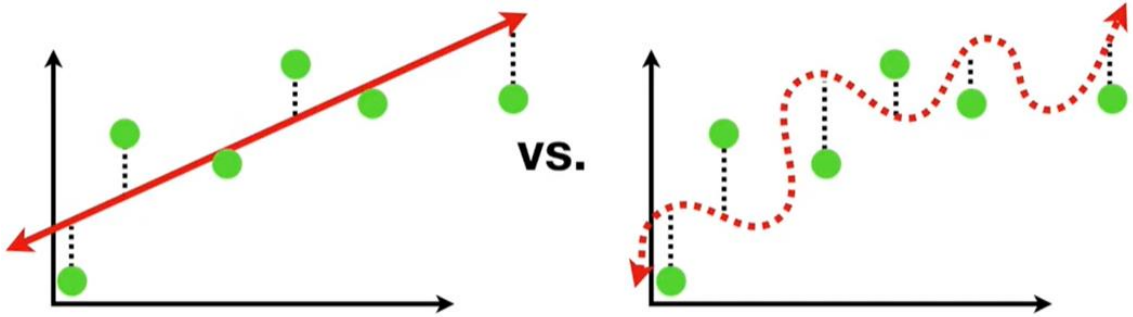
For training set, 곡선의 MSE = 0



High Bias
Low Variance
Underfit

Low Bias(=0)
High Variance
Overfit

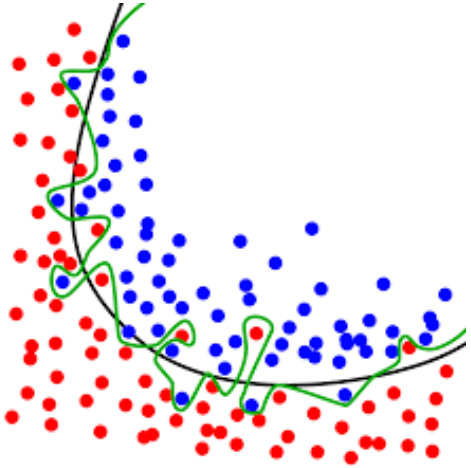
For testing set, 직선의 MSE < 곡선의 MSE

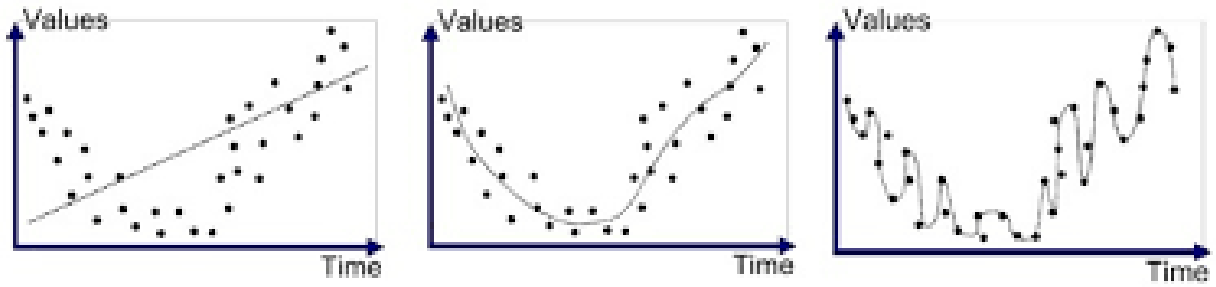


과소 적합

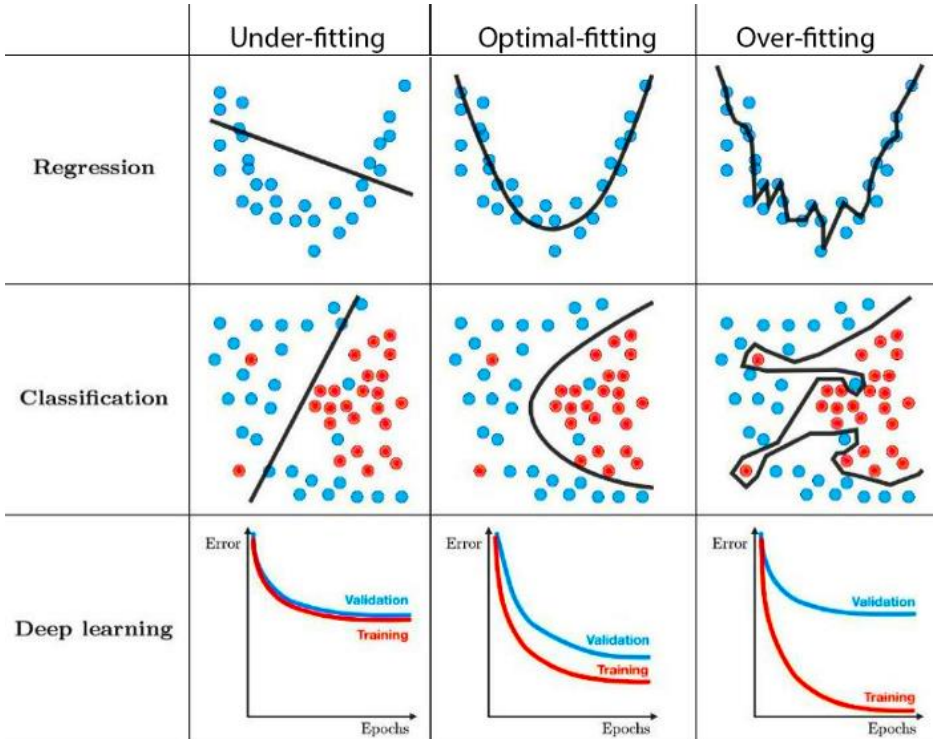
과대적합

3.1.2 과대적합,과소적합

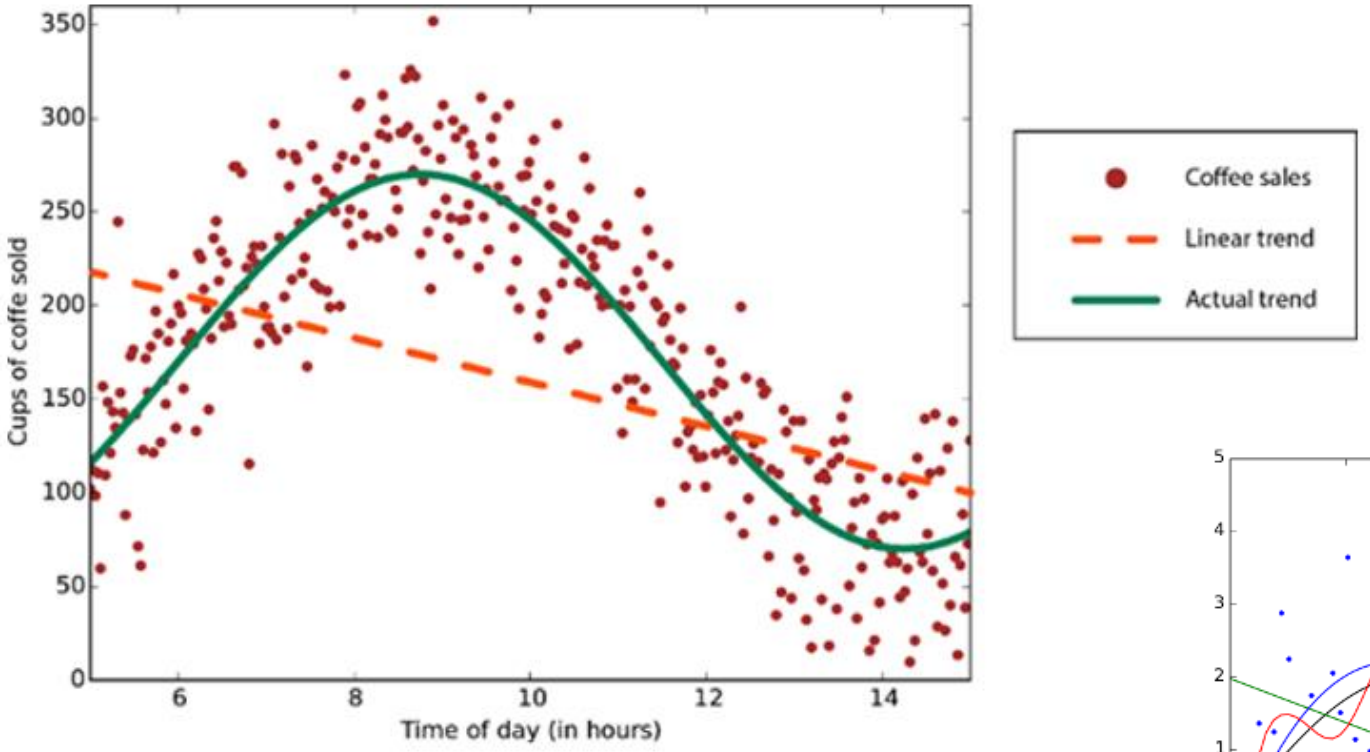


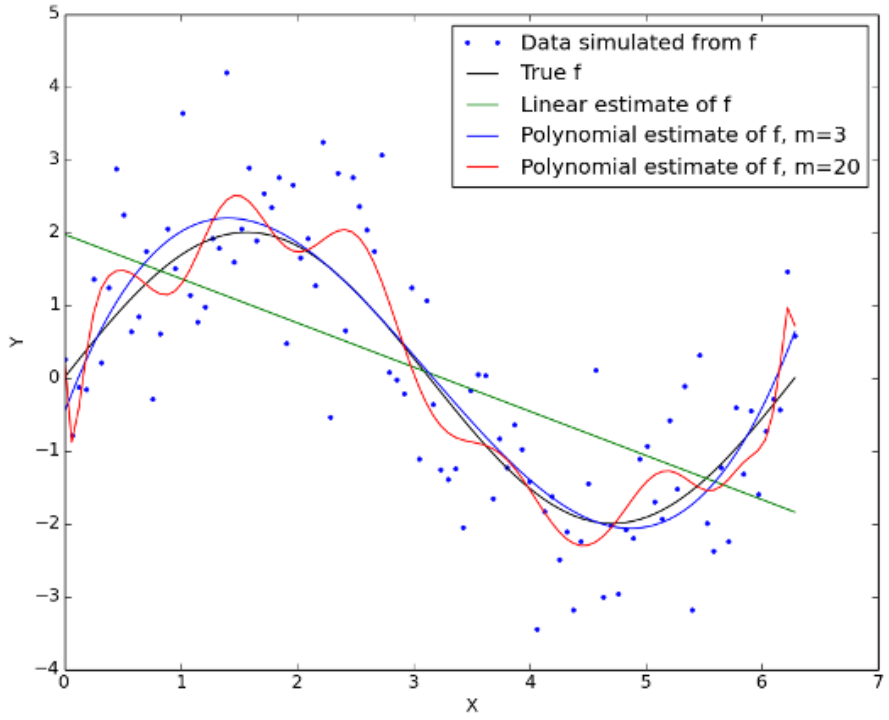


Underfitted　　Good Fit/Robust　　Overfitted



3.1.2 과대적합,과소적합





3.1.2 과대적합,과소적합

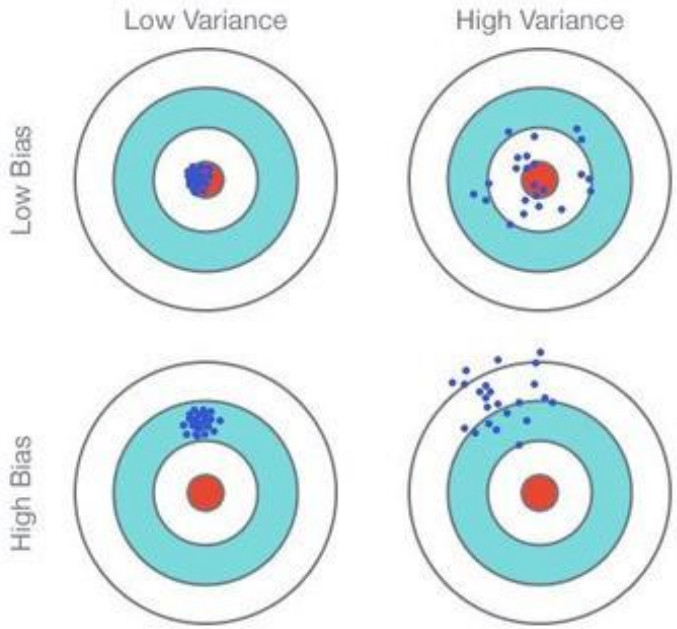
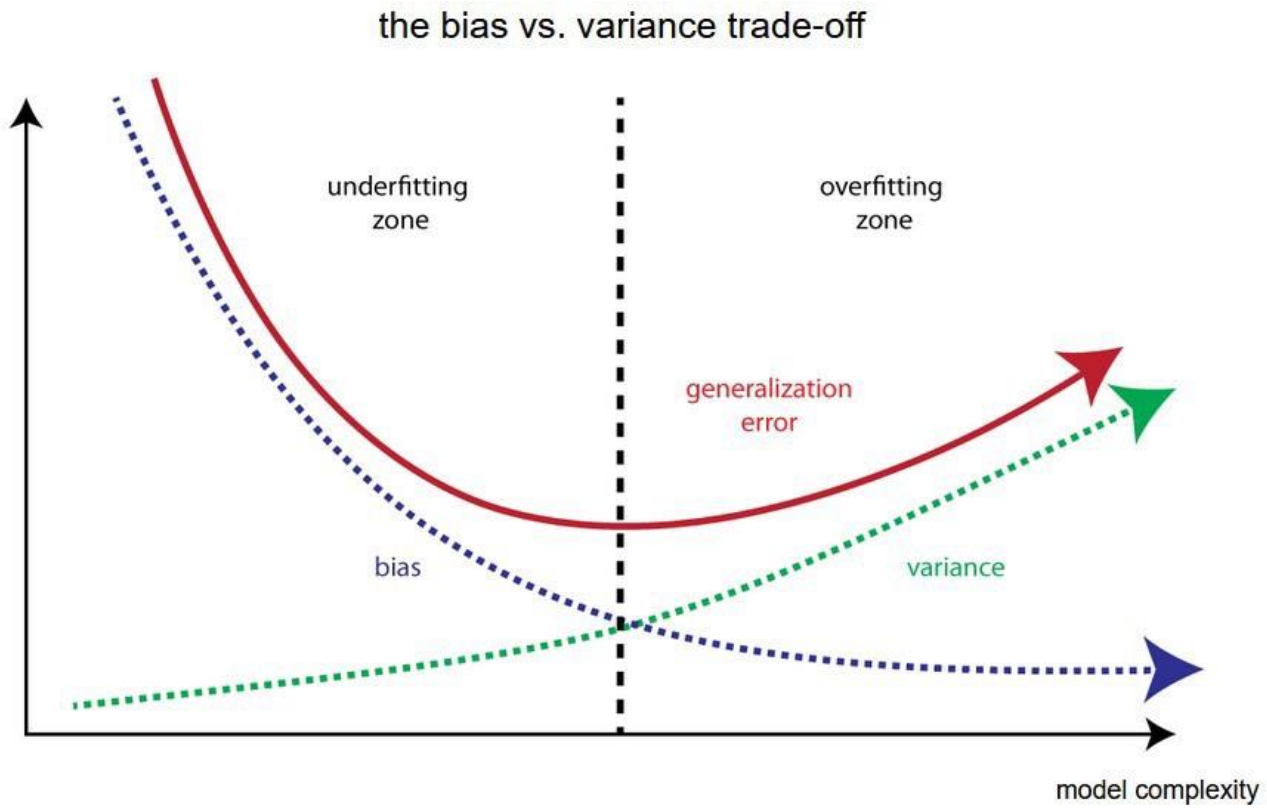


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off



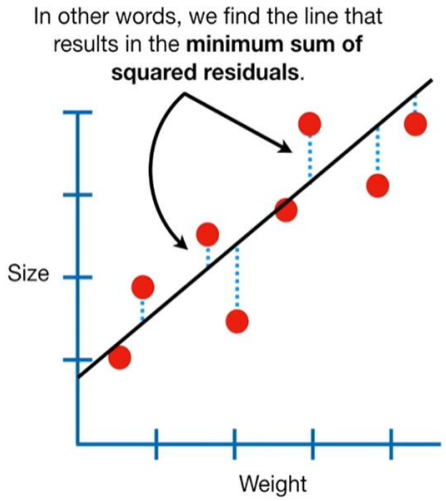
3.1.3 선형 모델 - Ridge 회귀

- L2 규제(regularization) 적용 : 가중치의 영향을 줄여서 모델이 과대적합되는 것을 억제.
- 평균제곱오차식에 가중치의 제곱항을 더하여 가중치 값이 커질수록 오차값이 커짐.
- 선형회귀에서 산출된 가중치 대비 리지 가중치(기울기)가 작을수록 X에 대한 y 값의 변동이 적으므로 편향을 희생하는 대신 분산을 줄여서 더 일반화된 모델을 생성함.
- L2 값 조정변수(람다)를 두어 최적 가중치를 찾음.
- 데이터의 양에 따라 성능의 차이가 남.

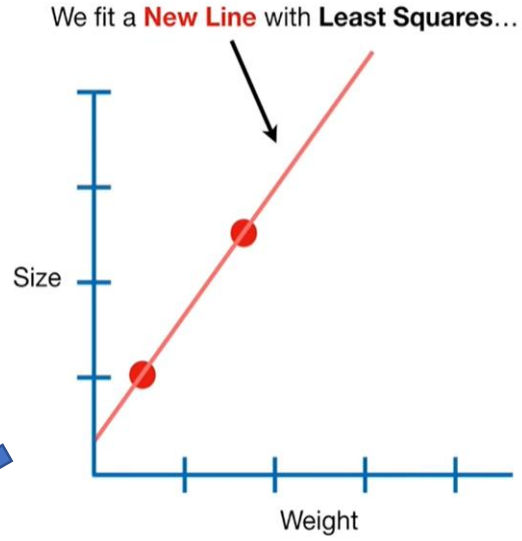
$$\text{Cost} = \underbrace{\sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

3.1.3 선형 모델 - Ridge 회귀

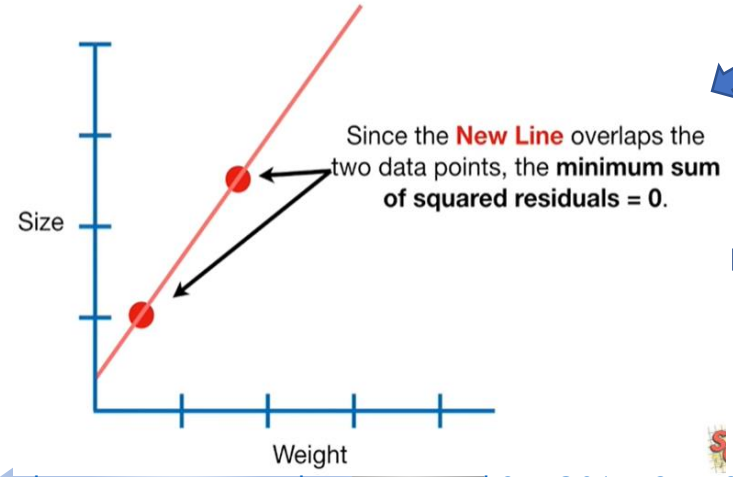
기존 선형회귀 방법



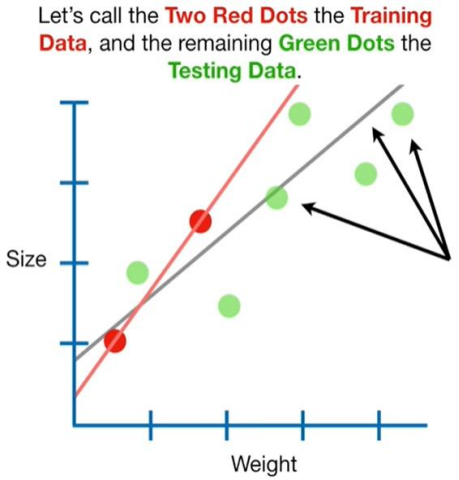
훈련데이타가 2점이라면...



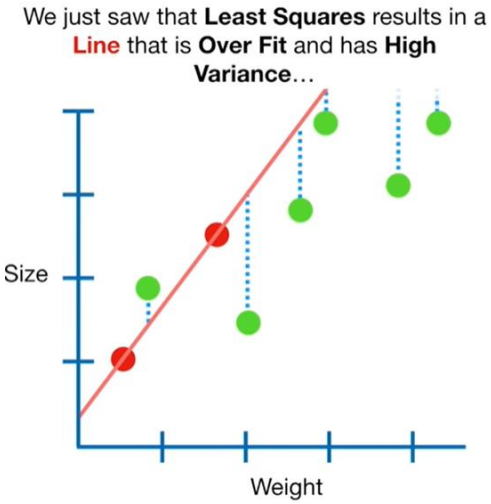
MSE = 0



나머지 : test data

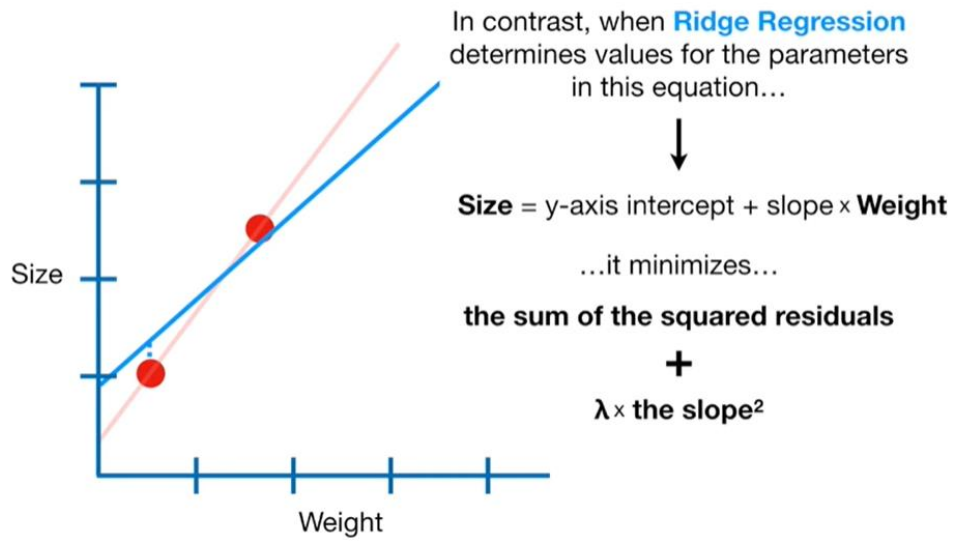


분산이 큰 모델 생성 → Overfit

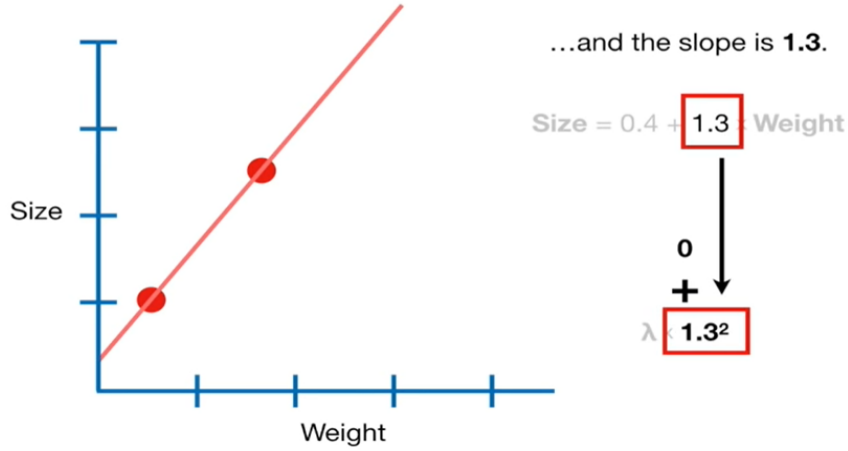


https://www.youtube.com/watch?v=Q81RR3yKn30&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=25

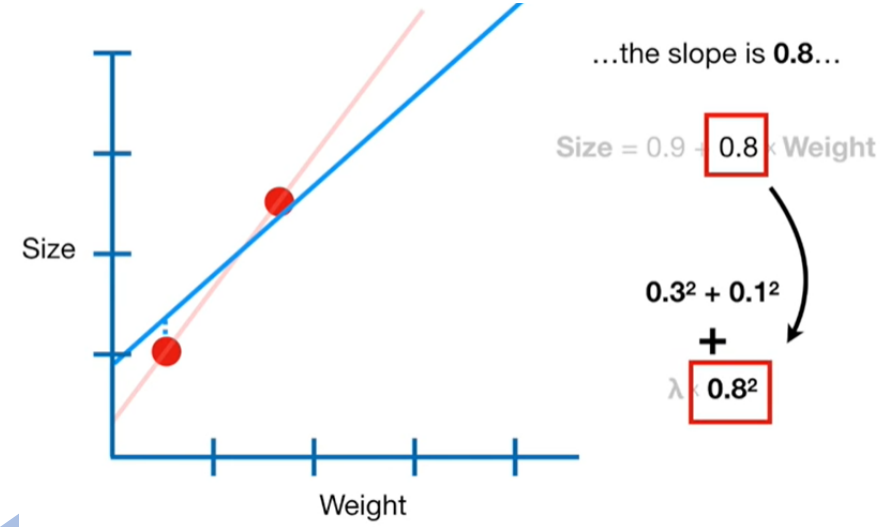
3.1.3 선형 모델 - Ridge 회귀



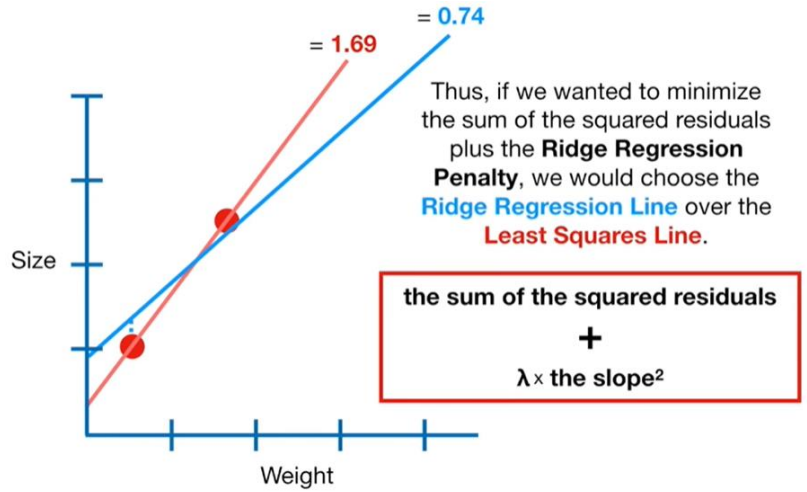
기울기 : 1.3 → MSE = 1.69



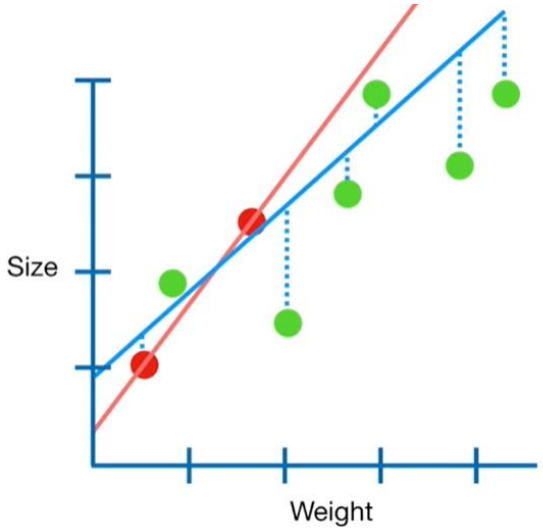
기울기 : 0.8 로 설정 → MSE = 0.74



0.8 기울기의 MSE 가 더 낮음.



3.1.3 선형 모델 - Ridge 회귀



In contrast, the **Ridge Regression Line**, which has the small amount of **Bias** due to the penalty, has less **Variance**.

- 규제강도
  - 규제강도 기본값(a = 0.1)으로 훈련했을 때 0.89. 0.75 달성. ← 선형회귀보다 우수한 성능 보임(cell 31)
  - 규제강도를 변화시켜서 학습하고 최적의 규제강도를 찾을 수 있음(cell 32 ~)
  - 규제강도를 높이면 훈련성능은 낮아지고 테스트 성능은 좋아지는 경향이 있음.
  - 규제강도를 낮추면 선형모델에 접근하게 됨.

- 데이터 양에 따른 효과(cell 35)
  - 작은 데이타셋에서는 훈련성능과 테스트 성능의 차이가 큼.
  - 큰 데이타셋일 수록 두 성능이 근접하게 되고 선형회귀도 리지 회귀에 근접하게 됨.

3.1 선형 모델 - Lasso Regresison

- L1 규제(regularization) 적용 : 가중치의 영향을 줄여서 모델이 과대적합되는 것을 억제.
- 평균제곱오차식에 가중치의 절대값항을 더함.
- 무관한 특성은 가중치가 0이 됨. → 모델에서 제외됨.

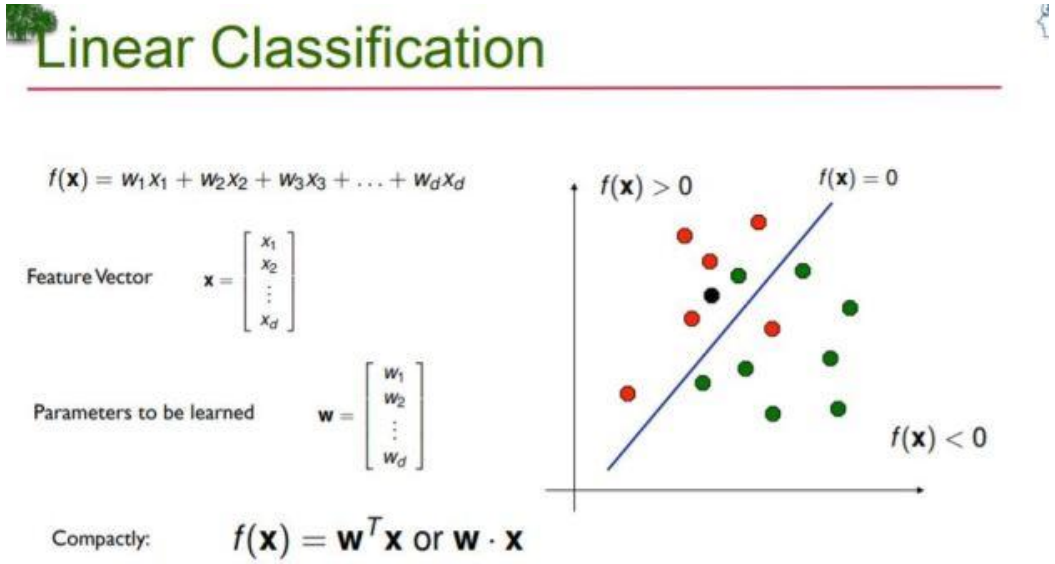
$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

- 보스턴 주택가격 예측 데이터에 적용해보면(cell 36~),
- 기본 설정(a = 0.1)일 때 20% 대 성능으로 좋지 않고 사용 특성은 104개 중 4개임.
- a = 0.01 을 사용하면(더 복잡한 모델) 0.90, 0.77 달성, 사용특성은 33 개이므로 리지에 비해 결과 해석이 좀 더 쉬움.

- 특징
  - 학습 및 예측 속도가 빠름
  - 큰 데이타셋과 희소한 데이타셋에서도 잘 작동함.
  - 샘플에 비해 특성 수가 많을 때 잘 작동
  - 특성수가 적을 때는 다른 모델의 성능이 더 우수함

3.1 선형 모델 – 분류용 선형모델

- 공식은 선형회귀와 동일
- 예측한 값을 임계치(보통 0)와 비교.
- 선, 평면, 초평면으로 두개의 클래스를 구분.
- 널리 알려진 선형분류 알고리즘
  - 로지스틱 회귀(Logistic Regression)
  - LinearSVC(Support Vector Classifier)



- make_forge() 데이타셋에 대해 로지스틱회귀, 선형 SVC 분류(cell 40~)
  - 두개의 포인트를 잘못 분류함.
- L2 규제를 적용하여 분류 성능 향상(cell)
  - C 값 높이면 → 규제 약화 → 모든 데이터에 대해 맞추려고 시도.
  - C 값 낮추면 → 규제 강화 → 많은 데이터에 대해 맞추려고 노력

- 암 데이타셋에 적용(cell 42 ~5)
  - 기본값(C = 1)에서 0.95, 0.95 성능을 보임.
  - 우수한 편이나 과소적합 상태. → 규제 약화 필요

3.1 선형 모델 – 다중 분류

- 이진 분류기법을 이용하고 1:다 기법으로 다중분류기법으로 확장(cell 47~).
  - 하나의 클래스를 나머지 모든 클래스와 구분하는 모델 생성
  - 모든 클래스에 대하여 모델을 생성하고 종합하여 예측 수행