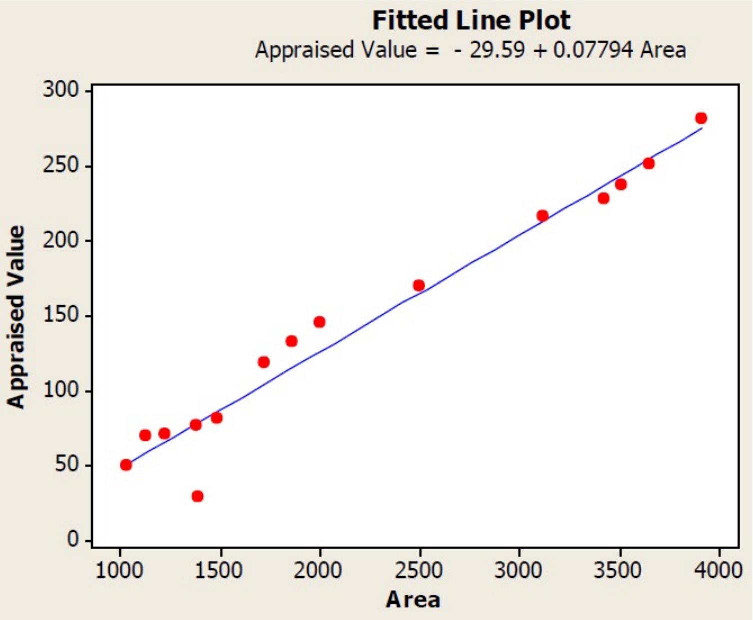
선형 회귀

Multiple Linear Regression

다중선형회귀(Multiple Linear Regression)

- **다중선형회귀 란?**
  - 수치형 설명변수 X와 연속형 숫자로 이뤄진 종속변수 Y간의 관계를 선형으로 가정하고 이를 가장 잘 표현할 수 있는 회귀계수를 데이터로부터 추정하는 모델

  - (예시) 1개의 설명변수(주택 크기:x1)와 연속형 종속변수(주택 가격:y)의 관계를 나타내는 직선을 찾는 문제



다중선형회귀(Multiple Linear Regression)

- **다중선형회귀 모델 방정식**
  - N개의 데이터, K개의 설명변수($x_{11}, \cdots, x_{nk}$), K개의 회귀계수($\beta_0, \cdots, \beta_k$)

$$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \ldots & \ldots & & \ldots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}$$

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}, \quad \vec{\varepsilon} \sim N(E(\vec{\varepsilon}), V(\vec{\varepsilon}))$$

$$E(\vec{\varepsilon}) = \begin{bmatrix} 0 \\ 0 \\ \ldots \\ 0 \end{bmatrix}, \quad V(\vec{\varepsilon}) = \sigma^2 I$$

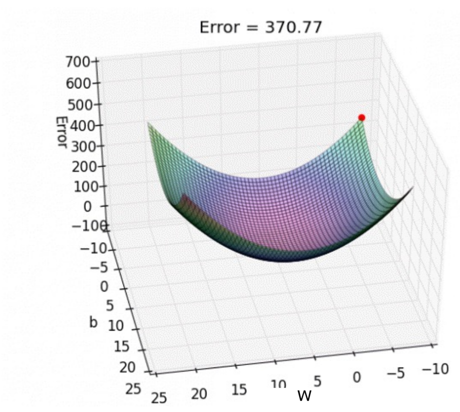
다중선형회귀(Multiple Linear Regression)

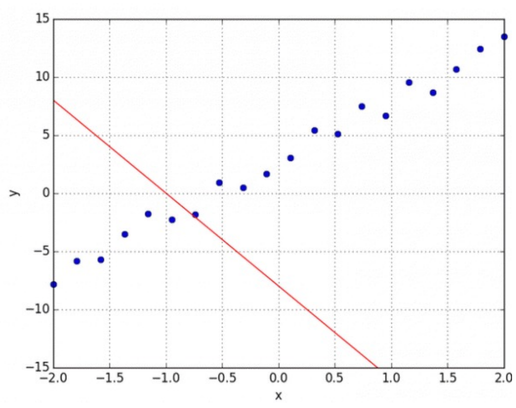
- **회귀 계수 결정법 : Direct Solution**
  - 선형회귀의 계수들은 실제값(Y)과 모델 예측값(Y’)의 차이, 오차제곱합(error sum of squares)을 최소로 하는 값을 회귀 계수로 선정
  - 최적의 계수들은 회귀 계수에 대해 미분한 식을 0으로 놓고 풀면 **명시적인 해**를 구할 수 있음. 즉, X와 Y데이터 만으로 회귀 계수를 구할 수 있음

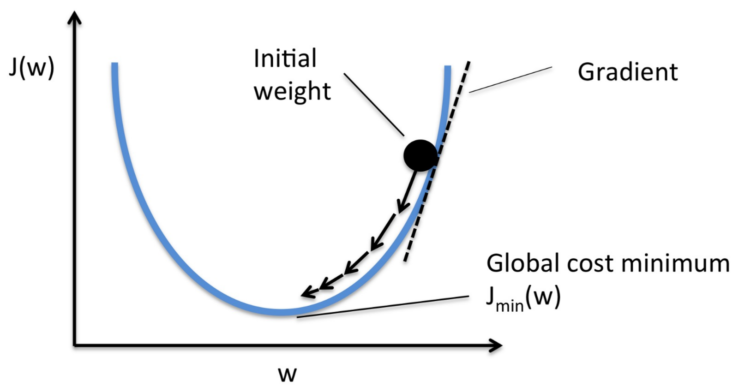
$$\vec{\beta} = (X^T X)^{-1} X^T \vec{y}$$

다중선형회귀(Multiple Linear Regression)

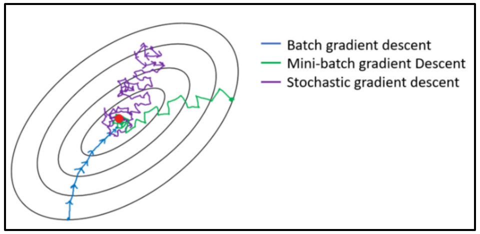
- **회귀 계수 결정법: Numerical Search**
  - 경사하강법(gradient descent) 같은 반복적인 방식으로 선형회귀 계수를 구할 수 있음
  - 경사하강법이란 어떤 함수 값(목적 함수, 비용 함수, 에러 값)을 최소화하기 위해 임의의 시작점을 잡은 후 해당 지점에서의 그래디언트(경사)를 구하고, 그래디언트의 반대 방향으로 조금씩 이동하는 과정을 여러번 반복하는 것







다중선형회귀(Multiple Linear Regression)

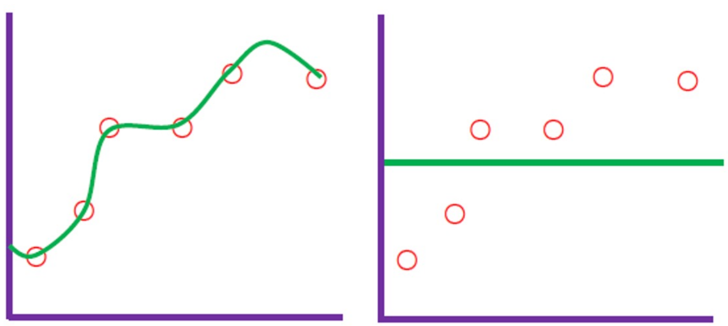


- Numerical Search
  - 경사하강법의 종류
    - Batch Gradient Descent (GD)
      - 파라미터를 업데이트 할 때마다 모든 학습 데이터를 사용하여 cost function 의 gradient 를 계산
      - Vanilla Gradient Descent 라 불림
      - 매우 낮은 학습 효율을 보일 수 있음
    - Stochastic Gradient Descent (SGD)
      - 파라미터를 업데이트 할 때, 무작위로 샘플링된 학습 데이터를 하나씩만 이용하여 cost function의 gradient를 계산
      - 모델을 자주 업데이트 하며, 성능 개선 정도를 빠르게 확인 가능
      - Local minima에 빠질 가능성을 줄일 수 있음
      - 최소 cost에 수렴했는지의 판단이 상대적으로 어려움
    - Mini Batch Gradient Descent
      - 파라미터를 업데이트 할때마다 일정량의 일부 데이터를 무작위로 뽑아 cost function의 gradient 를 계산
      - Batch Gradient Descent와 Stochastic Gradient Descent 개념의 혼합
      - SGD의 노이즈를 줄이면서, GD의 전체 배치보다 효율적
      - 널리 사용되는 기법

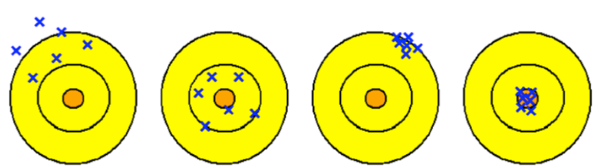
정규화

- **정규화(regularization)란?**
  - 회귀계수가 가질 수 있는 값에 제약조건을 부여하여 <u>미래 데이터에 대한 오차 기대</u>
  - 미래데이터에 대한 오차의 기대 값은 모델의 **Bias**와 **variance**로 분해 가능.
  - 정규화는 variance를 감소시켜 일반화 성능을 높이는 기법
    - 단, 이 과정에서 bias가 증가할 수 있음

  - 왼쪽 그림은 학습데이터를 정말 잘 맞추고 있지만, 미래 데이터가 조금만 바뀌어도 예측 값이 들쭉날쭉할 수 있음
  - 우측 그림은 가장 강한 수준의 정규화를 수행한 결과로 학습데이터에 대한 설명력을 다소 포기하는 대신 미래 데이터 변화에 상대적으로 안정적인 결과를 나타냄



정규화의 결과를 직관적으로 나타낸 그림

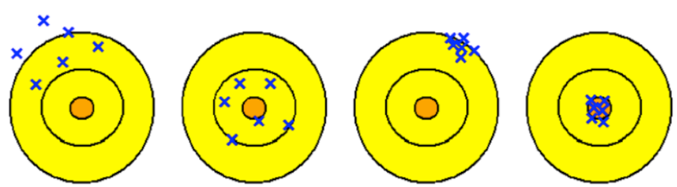


Bias	High	Low	High	Low
Variance	High	High	Low	Low

Bias-Variance Decomposition

- **Bias-Variance Decomposition란?**
  - **일반화(generalization)** 성능을 높이는 정규화(Regularization), 앙상블(ensemble) 기법의 이론적 배경
  - 학습에 쓰지 않은 미래데이터에 대한 오차의 기대값을 모델의 Bias와 Variance로 분해하자는 내용

- **Bias-Variance의 직관적인 이해**
  - 첫번째 그림을 보면 예측값(파란색 엑스표)의 평균이 과녁(Truth)과 멀리 떨어져 있어 Bias가 크고, 예측값들이 서로 멀리 떨어져 있어 Variance 또한 큼
  - 네번째 그림의 경우 Bias, Variance 모두 작음. 제일 이상적임
  - 부스팅(Boosting)은 Bias 를 줄여 성능을 높이고, 라쏘회귀(Lasso regression)는 Variance를 줄여 성능을 높이는 기법임



Bias	High	Low	High	Low
Variance	High	High	Low	Low
		NN SVM KNN (small K) 튜닝만 잘하면 과녁을 제대로 맞출 모델	로지스틱 회귀 LDA KNN (Large K) 데이터 노이즈에 강인한 모델	