밑바닥부터 시작하는 데이터 과학

5회차

- 군집화
- 선형회귀
- 로지스틱 회귀

1. 군집화

군집화 (Clustering)

데이터의 종속 변수 값을 예측하는 것이 아니라 독립 변수의 특성만으로 데이터의 그룹
즉, 클러스터(cluster)를 형성하는 작업.

1. 군집화

K-Means

목적함수 값이 최소화될 때까지 클러스터의 수 K와 각 클러스터의 중심(centroid) u_k 반복해서 찾는 것이다

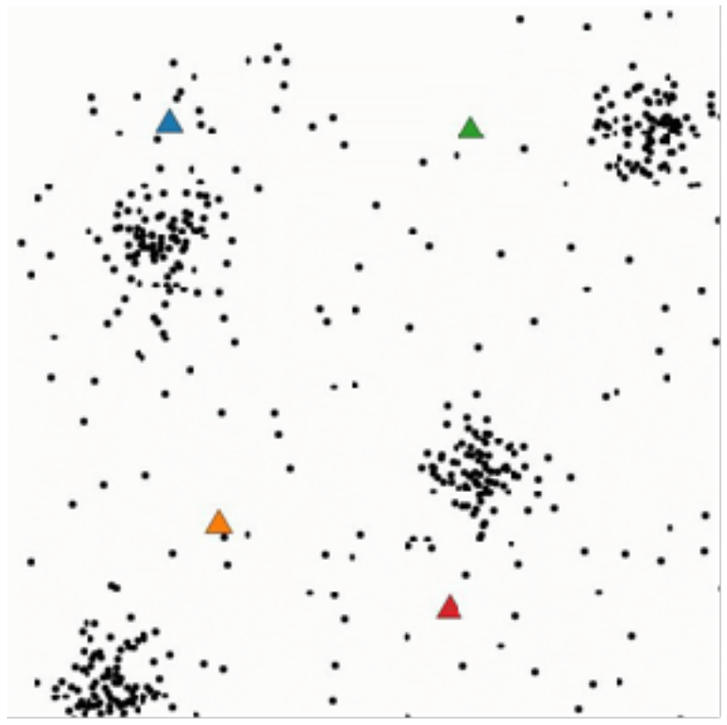
$$J = \sum_{k=1}^{K} \sum_{i \in C_k} d(x_i, \mu_k)$$

- d는 두 데이터의 유사도 함수(Similarity Function) 혹은 거리(Distance)로 정의한다.

$$d(x_i, \mu_k) = \|x_i - \mu_k\|^2$$

연산 절차

1. 임의의 중심값 μ_k 를 고른다. (보통 데이터 샘플 중의 하나를 선택)
2. 중심에서 각 샘플 데이터까지의 거리를 계산
3. 각 데이터 샘플에서 가장 가까운 중심을 선택하여 클러스터 갱신
4. 다시 만들어진 클러스터에 대해 중심을 다시 계산하고 1 ~ 4를 반복한다.



2. 선형회귀

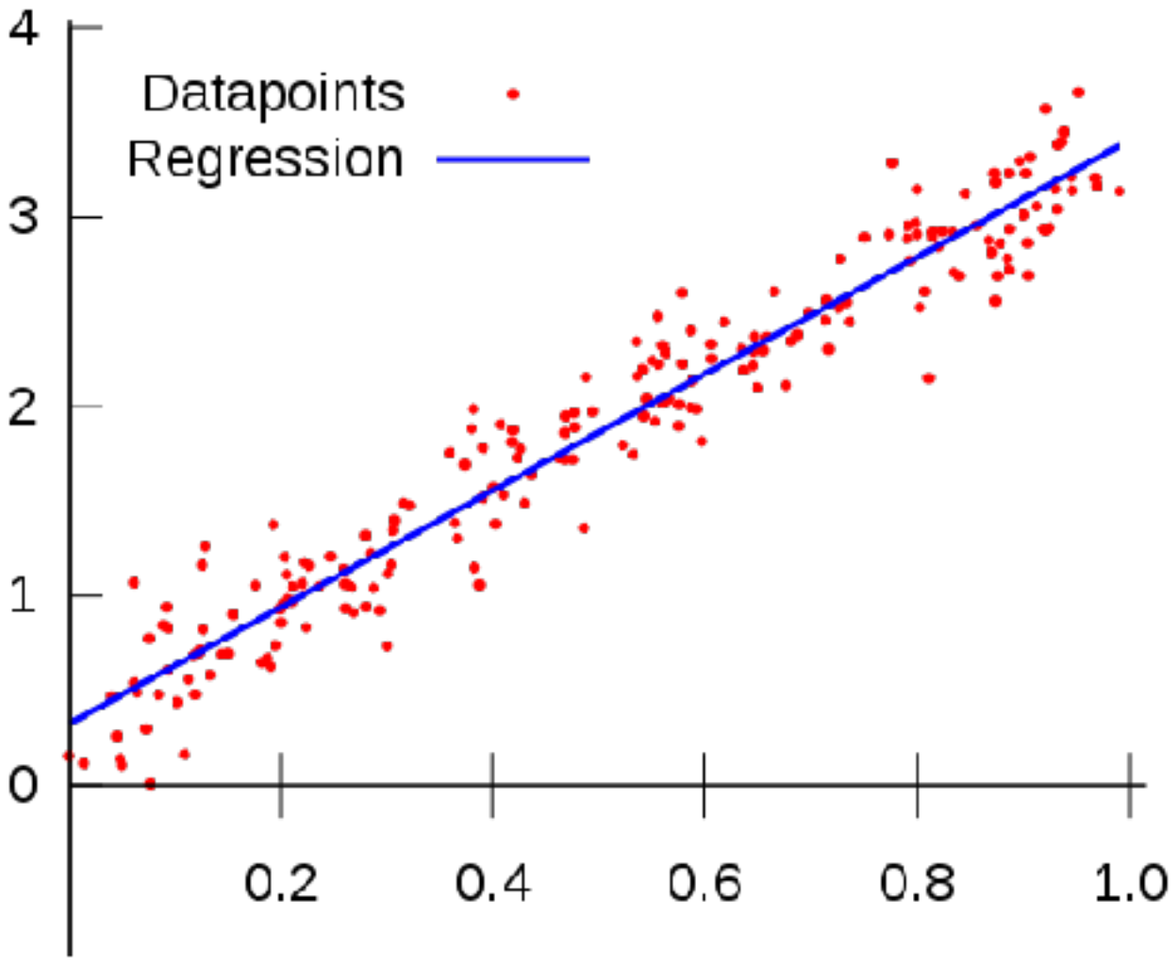
선형회귀 (Linear Regression)

2. 선형회귀

선형회귀

수치형 **설명변수 X**와 **연속형 숫자**로 이뤄진 **종속변수 Y**간의 관계를 **선형**으로 가정하고
이를 가장 잘 표현할 수 있는 **회귀계수**를 데이터로부터 추정하는 모델

- **예시 : 집 크기 (x)에 대한 집 값 (y) 예측**



$$Y = Xw$$

2. 선형회귀

선형회귀식의 표기

$$Y=Xw$$

- 설명 변수 x가 D개가 있고, 데이터가 N개가 있을 때, x에 대한 전체 데이터를 오른쪽과 같이 (D X N) 행렬로 표기가 가능하다.

$$x_i=\begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iD} \end{bmatrix} \rightarrow X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

- 전체 수식은 설명 변수 벡터(x)와 가중치 벡터(w)의 내적으로 간단하게 나타낼 수 있다.

$$f(x)=w_1x_1+w_2x_2+\cdots+w_Dx_D=\begin{bmatrix} x_1 & x_2 & \cdots & x_D \end{bmatrix}\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}=x_a^Tw_a=w_a^Tx_a$$

2. 선형회귀

선형회귀식의 적합성

결정 계수 (R square)

- 종속변수 (y)의 총 변화량 중 모델이 잡아낼 수 있는 변화량의 비율

- **결정계수 = 1- RSS (모델이 잡아내지 못한 변화량) / TSS (데이터 전체 변화량)**

  -> 전체 변화량 중 모델이 잡아내지 못하는 변화량의 비율

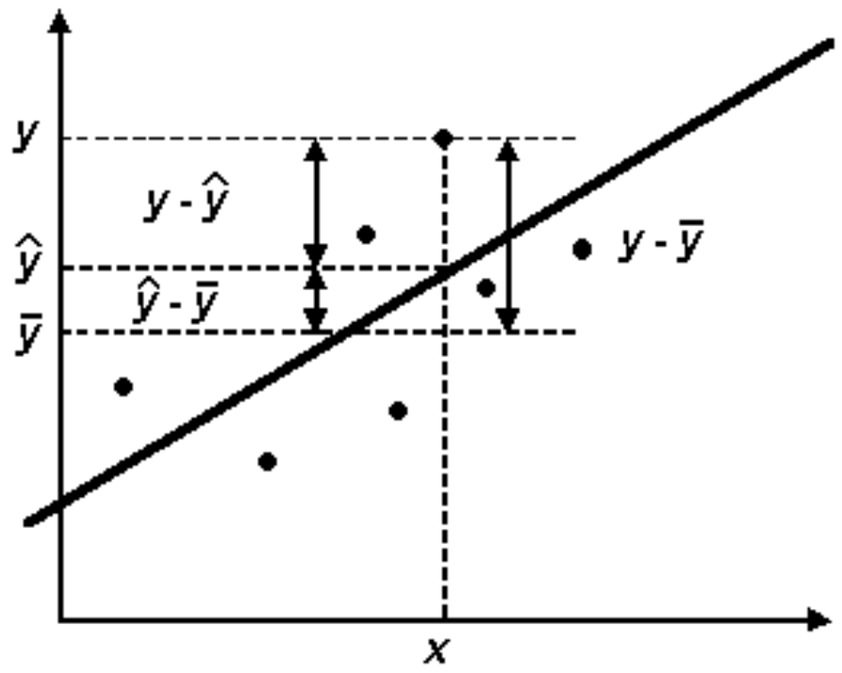
  - 만들어낸 모델이 실제 관측된 값의 평균값 정도만 예측한다고 하면 결정계수는 0이 된다.

$$RSS(Residual\ Sum\ of\ Square) = \sum (y - \hat{y})^2$$ **(실제값 - 예측값)**

$$TSS(Total\ Sum\ of\ Square) = \sum (y - \overline{y})^2$$ **(실제값 - 평균)**

$$ESS(Explained\ Sum\ of\ Square) = \sum (\hat{y} - \overline{y})^2$$ **(예측값 - 평균)**

$$R^2 = 1 - \frac{RSS}{TSS} \quad (0 \leq R^2 \leq 1)$$



2. 선형회귀

해를 구하는 방식 (1)

OLS (Ordinary Least Square)

- 잔차제곱합(RSS: Residual Sum of Squares)를 최소화하는 가중치 벡터를 행렬 미분으로 구하는 방법
  - 잔차제곱합 : 실제값과 예측값 차이의 제곱합
- 선형회귀의 목표가 실제값을 가장 잘 예측할 수 있는 함수 f(x)를 만드는 과정이라고 생각하면 된다. $\hat{y} = f(x) \approx y$

- **OLS의 가정**
- **1) X의 각 열벡터들은 서로 일차 독립해야한다.**
  - 일차 독립 : 어떤 벡터도 다른 벡터와의 선형 결합으로 만들어질 수 없는 상태
  - 이 가정이 성립하지 않는다면 w를 추정할 수 없다.

- **2) X의 모든 열은 오류(잔차, residual)와는 아무런 상관 관계가 없어야한다.**
  - x_1이 오류와 1에 가까운 correlation이 있다고 생각해보자.
  - 오류를 줄이기 최소화 시키는 모델링을 하게되면 w_1 (x_1의 가중치)는 0에 가깝게 될 것이고, 나머지 가중치들은 편향되게 구해질 것이다.

2. 선형회귀

해를 구하는 방식 (1)

$$\hat{y} = Xw$$

$$e = y - \hat{y} = y - Xw$$

$$\begin{aligned} RSS &= e^T e \\ &= (y - Xw)^T (y - Xw) \\ &= y^T y - 2y^T Xw + w^T X^T Xw \end{aligned}$$

- RSS의 최솟값을 구하기 위해 그레디언트 벡터를 구한다. (RSS를 w에 대해 미분한다)

$$\frac{dRSS}{dw} = -2X^T y + 2X^T Xw$$

- RSS가 최소가 되는 최적화 조건은 그레디언트 벡터가 0벡터이어야 하므로 다음 식이 성립한다.

$$\frac{dRSS}{dw} = 0$$

$$X^T Xw^* = X^T y$$

- X^TX의 역행렬이 존재한다면 최적 가중치 벡터 w를 구할 수 있다.

$$w^* = (X^T X)^{-1} X^T y$$

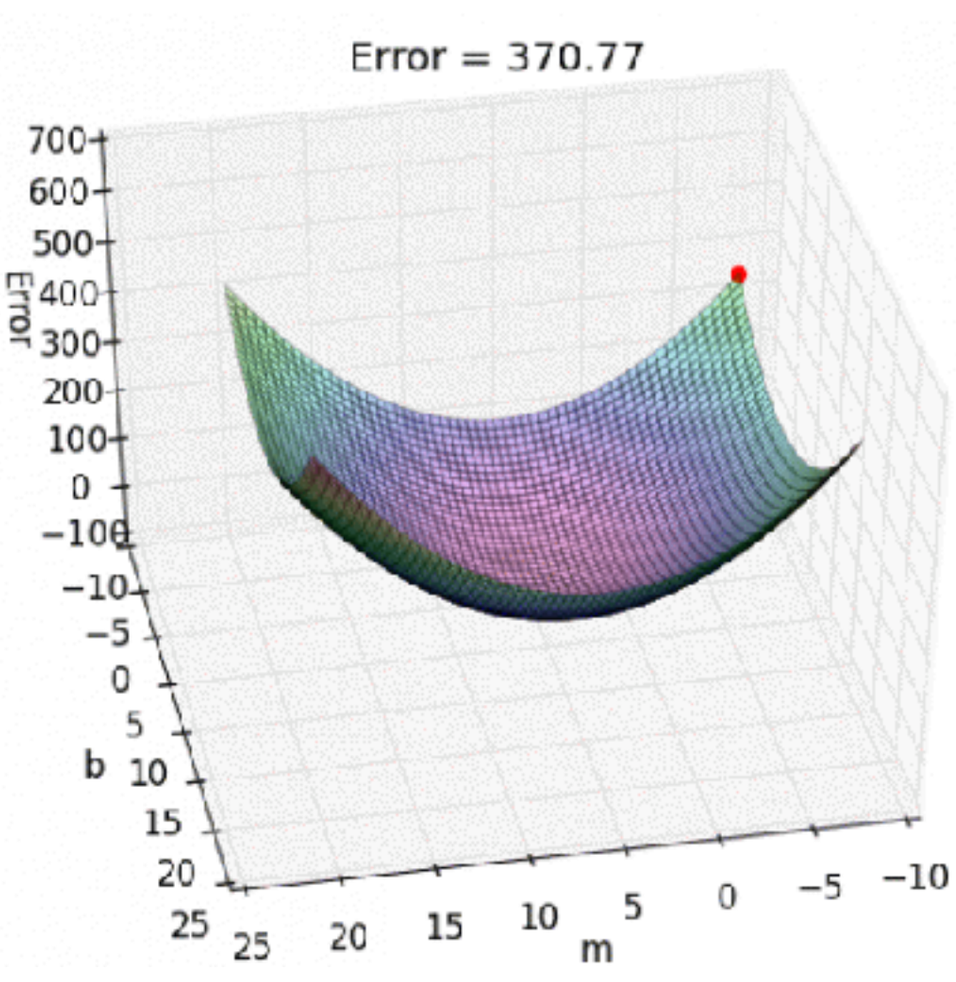
- 이와 같은 방식으로 명시적 해를 단번에 추정할 수 있다.

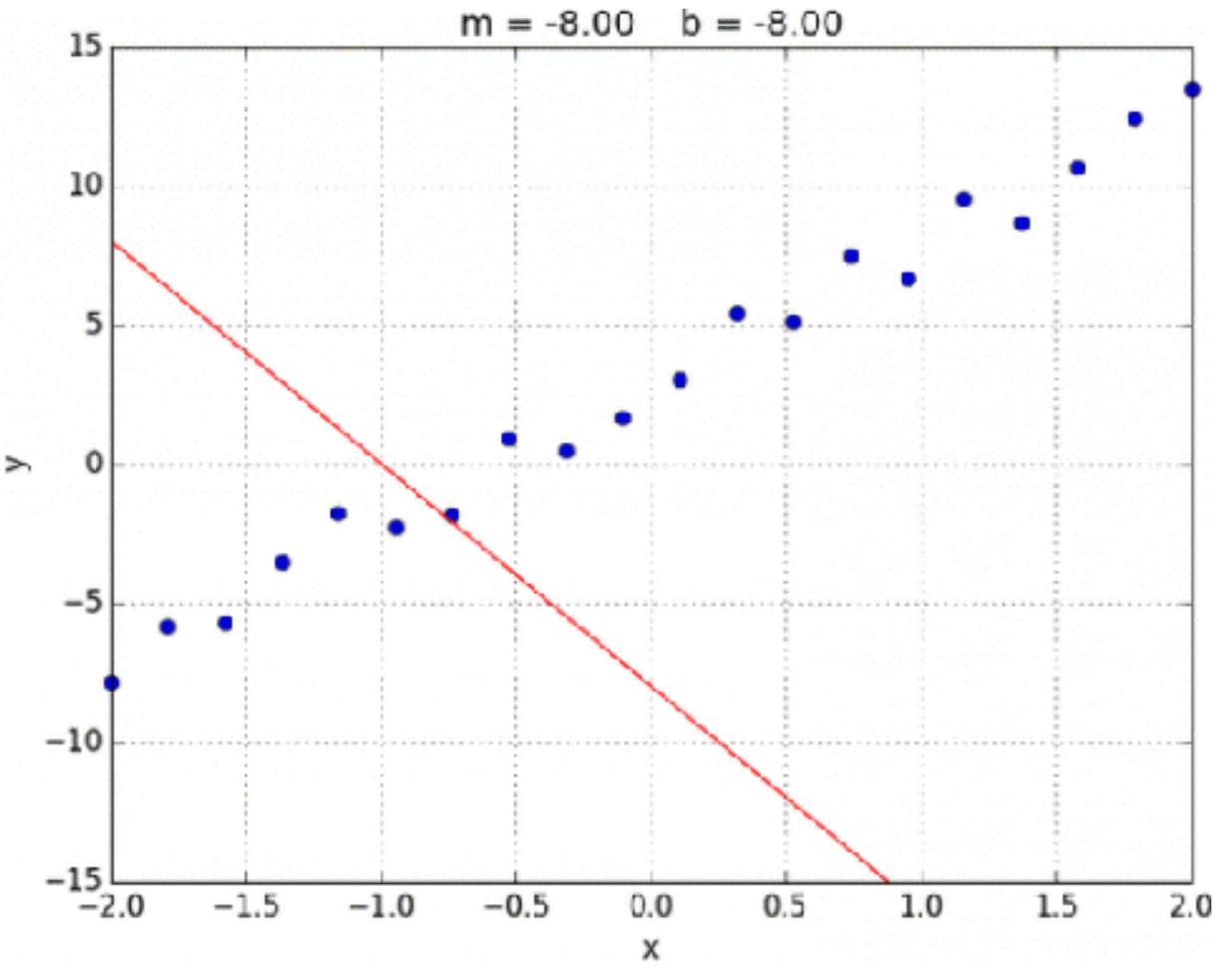
2. 선형회귀

해를 구하는 방식 (2)

수치적 최적화

- 잔차제곱합(RSS: Residual Sum of Squares)를 최소화하는 가중치 벡터를 행렬 미분으로 구하는 방식이 아니라 경사하강법을 이용하여 최적값을 찾을 수 있다.





2. 선형회귀

회귀계수의 신뢰성 문제

부트스트래핑(Bootstrapping)

- OLS(Ordinary Least Square) 방법을 사용하면 데이터에 대한 확률론적인 가정없이도 최적의 가중치를 계산할 수 있다. 그러나 이 경우에는 계산한 가중치가 어느 정도의 신뢰도 또는 안정성을 가지는지 확인할 수 있는 방법이 없다.
- 부트스트래핑(bootstrapping)은 회귀 분석에 사용한 데이터가 달라진다면 회귀 분석의 결과는 어느 정도 영향을 받는지를 알기 위한 방법이다.

절차

- 기존 데이터에서 중복 재추출하여 새로운 데이터를 만들어낸다.
- 새로운 만들어진 데이터에 대해 회귀분석을 실시한다.
- 분석 결과를 저장하고, 통계치를 확인한다.

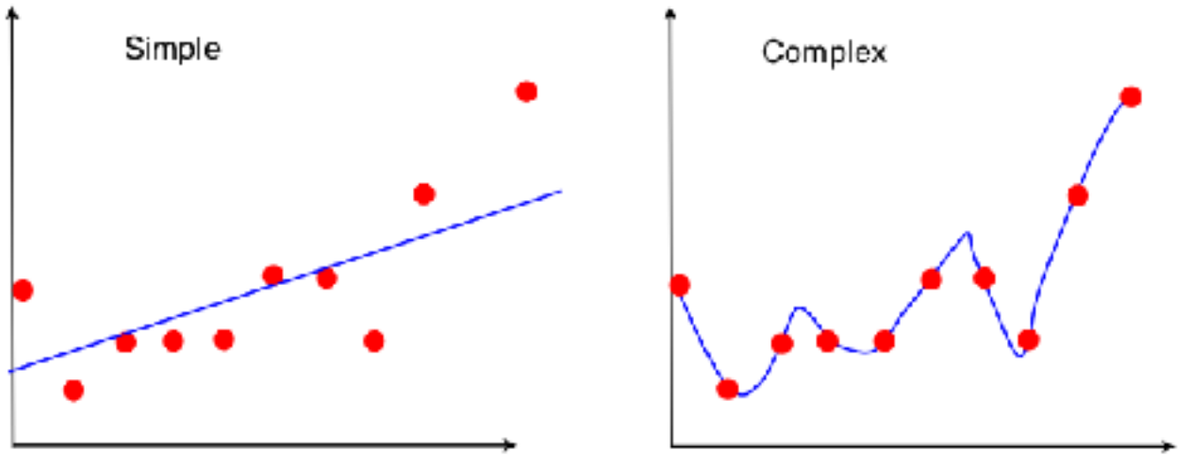
2. 선형회귀

정규화 (Regularization)

변수가 많을 때 발생할 수 있는 문제점
(1) 오버피팅의 가능성
(2) 모델 해석의 어려움
- 회귀계수가 너무 많으면 모델을 해석하기 어려워진다.

- 이 문제를 해결하기 위해, 적은 개수의 변수로 일반화된 모델을 만들기도한다.

정규화 (Regularization)
- 회귀계수가 가질 수 있는 값에 제약조건을 부여하여 일반화 성능을 높이는 기법
- 모델의 설명력은 다소 포기하더라도 안정적인 결과를 내도록 만든다.
  - 안정적인 결과를 만들어내는 변수들을 선택하게된다.



2. 선형회귀

정규화 (Regularization)

Ridge Regression (릿지 회귀)

- 잔차제곱합(RSS)를 최소화하면서 회귀계수 벡터 w에 L2 norm을 제한하는 기법이다.
  - L2 norm : 벡터의 크기 ||w||_2 = wTw
  - 어떤 크기의 람다를 선택하느냐에 따라 제약의 정도가 결정된다.
- 총 계수의 합을 줄여준다.

$$\hat{w}^{ridge} = arg\min_{w}\{(Y-Xw)^T(Y-Xw)+\lambda w^T w\}$$

$$\hat{w}^{ridge*} = (X^T X+\lambda I)^{-1}X^T Y$$

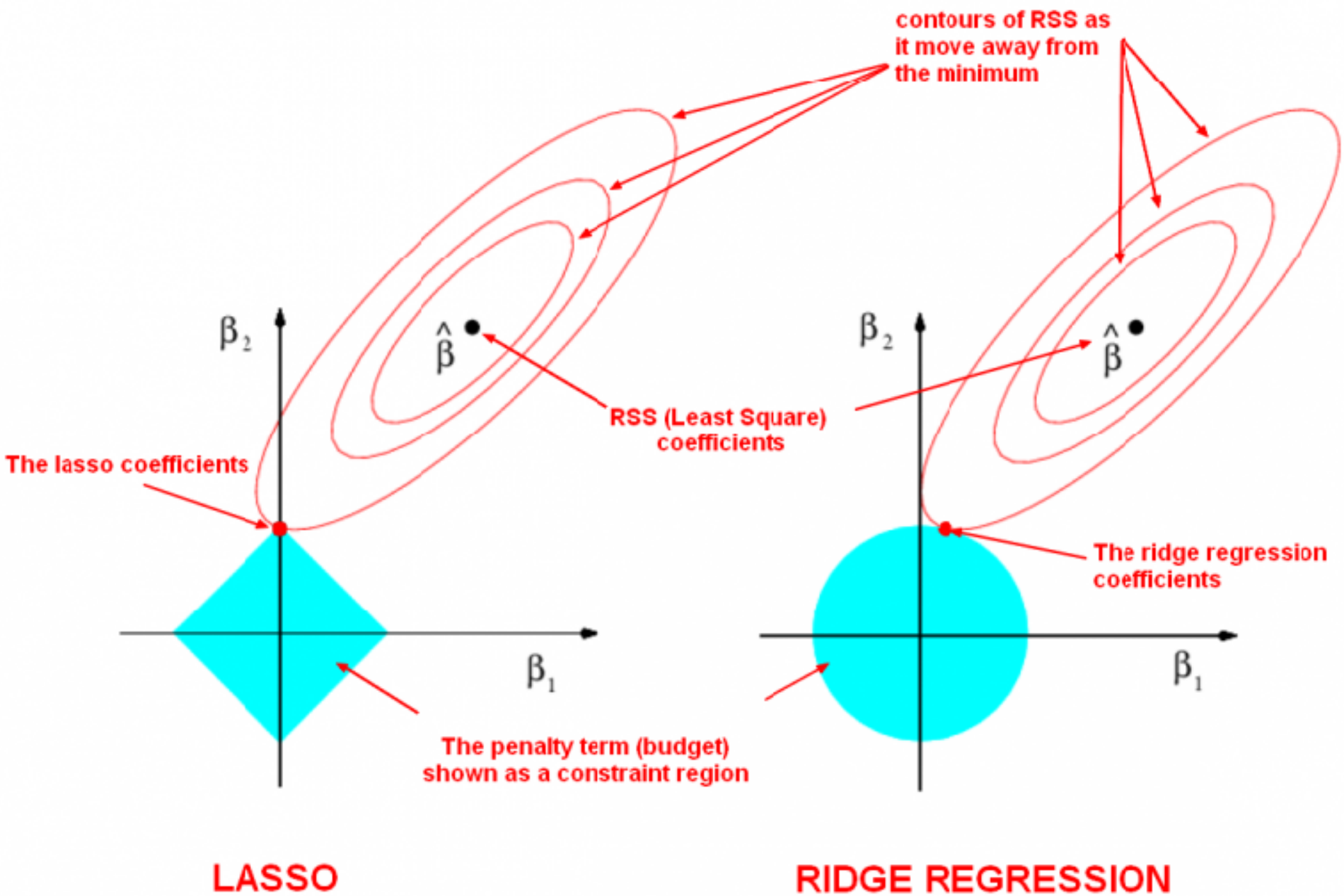
Lasso Regression (릿지 회귀)

- 잔차제곱합(RSS)를 최소화하면서 회귀계수 벡터 w에 L1 norm을 제한하는 기법이다.
  - L1 norm : wTw

$$\hat{w}^{lasso} = arg\min_{w}\{(Y-Xw)^T(Y-Xw)+\lambda||w||_1\}$$

2. 선형회귀

정규화 (Regularization)



3. 로지스틱 회귀

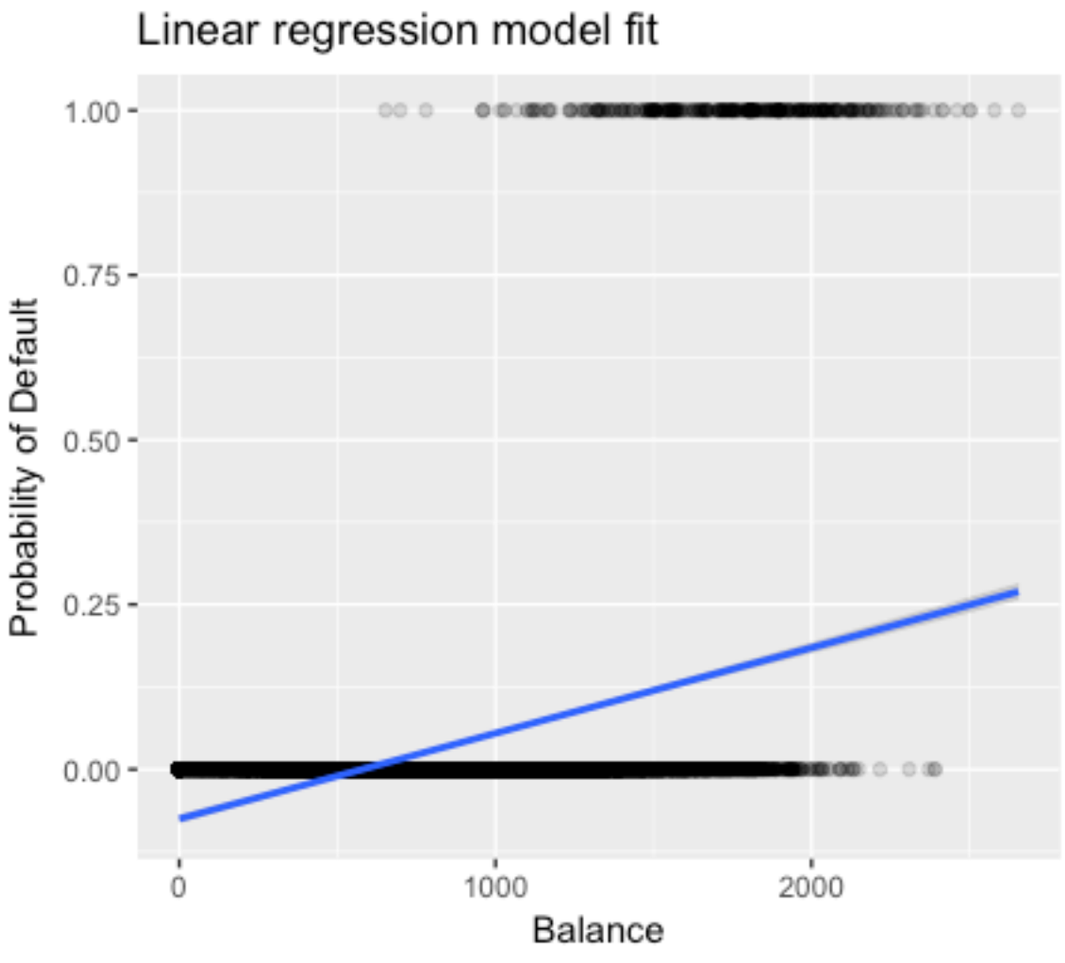
로지스틱회귀 (Logistic Regression)

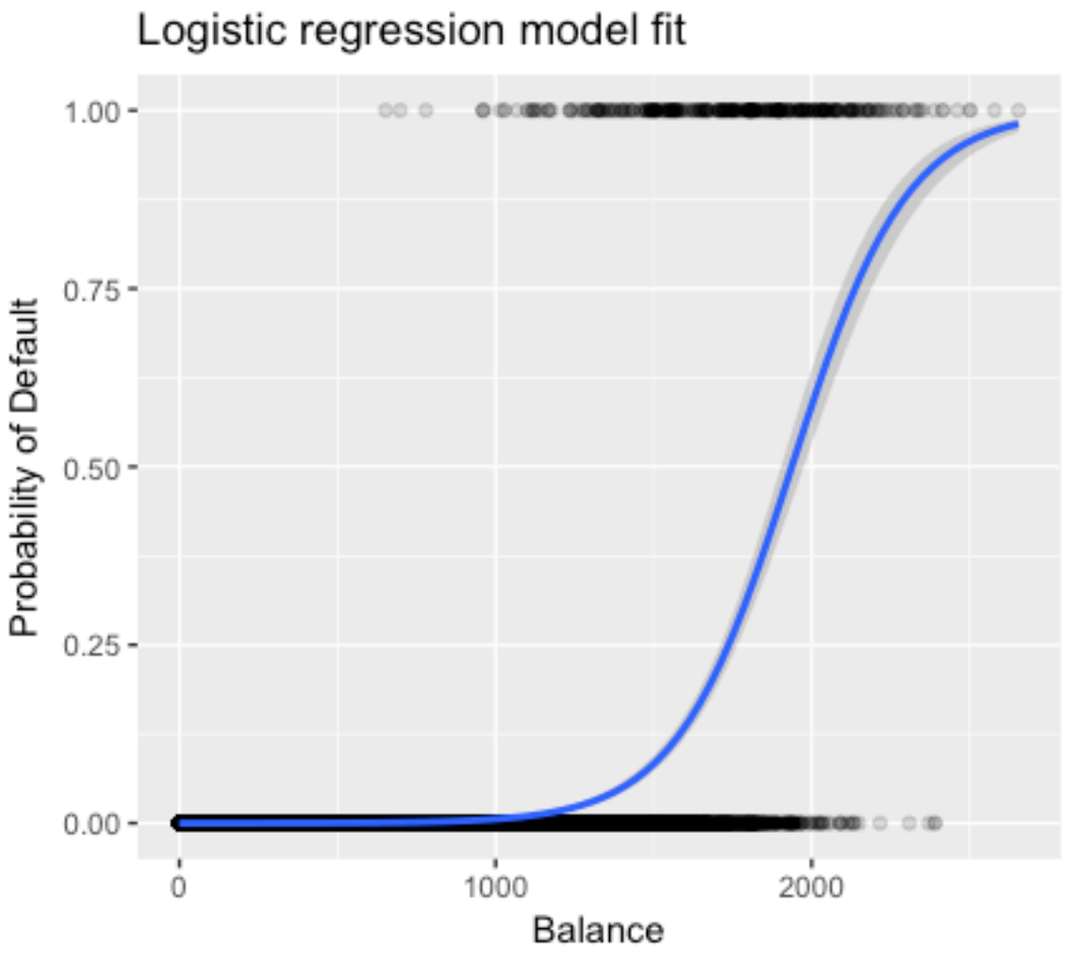
3. 로지스틱 회귀

로지스틱 회귀

선형회귀와는 다르게 종속 변수(y)가 **범주형 데이터**일 경우에도 사용이 가능하다.

- **예시 : 잔고 (x)에 대한 채무불이행 (y, (0, 1)) 예측**





3. 로지스틱 회귀

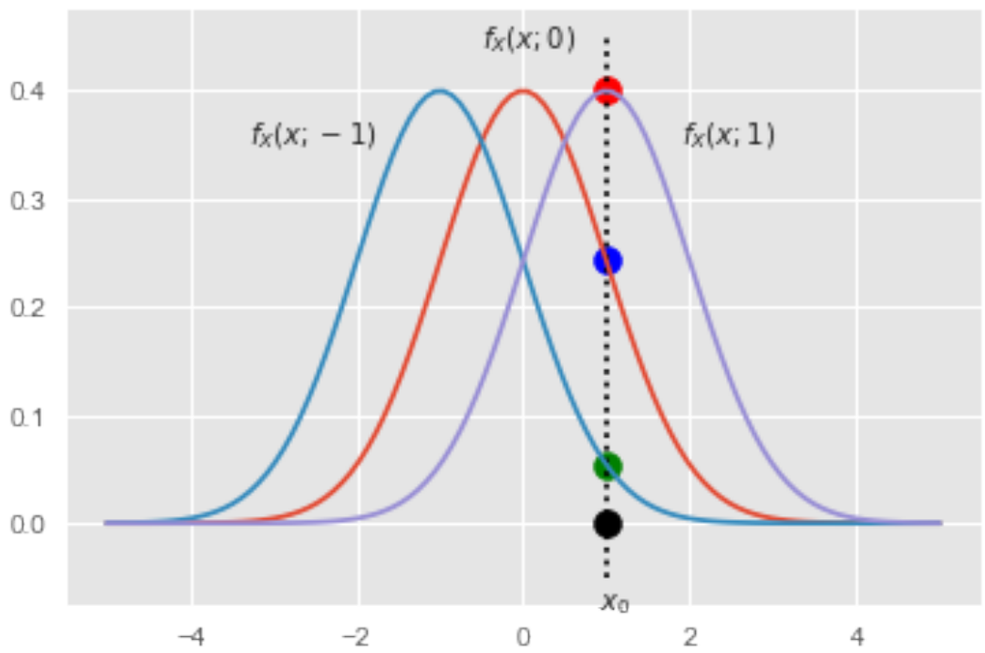
로지스틱 회귀

최대우도추정법 (Maximum Likelihood Estimation)

- 주어진 샘플 x에 대해 likelihood를 가장 크게 해주는 모수 theta를 찾는 방법.
- 데이터가 어떤 확률 분포 f(x;theta)에서 나왔고, x에 대한 확률 값을 알아야되는 상황을 가정해보자.
  - theta : 확률 모형의 모수(parameter)집합
  - theta가 주어진 경우에는 x에 대한 확률값을 바로 계산할 수 있다.
    - (함수의 관점에서는 theta는 주어져있고, x만을 변수로 가정한다.)
  - 하지만 현실세계에서는 주어진 데이터에 대한 theta를 알고있기 힘들다.
  - 그렇다면 주어진 데이터에서 가장 그럴듯한 theta는 무엇인지 알면, 반대로 x에 대한 확률 값을 계산할 수 있지 않을까?
    - **likelihood 함수 : L(theta;x)**
    - **주어진 x에 대한 theta의 상대적 가능성을 계산할 수 있다.**

예시) 정규분포를 가지는 확률 변수의 **분산**은 알고 있으나 **평균**은 모르는 상태

- x_0으로 주어졌을 때, 평균=1일때 likelihood가 가장 크다.



3. 로지스틱 회귀

로지스틱 회귀

전제조건

- Y를 확률 변수로 생각해보자.
- 분류(예측)하고자 하는 값이 0과 1로만 이루어져있다면, Y는 **베르누이 확률 변수**가 된다.

$$P(y=y_i)=P^{y_i}(1-P)^{1-y_i}$$

$$L=\prod_i P^{y_i}(1-P)^{1-y_i}$$

로지스틱회귀의 likelihood함수

- 관측치가 i개 있고. v의 결과는 0과1만 있는 이항로지스틱 모델의 파라미터 w가 주어졌다고 가정해보자.
- y_i 는 $\theta(w^T x_i)$ 의 확률로 1, $1-\theta(w^T x_i)$ 의 확률로 0이 된다.
  - 여기서 theta는 로지스틱 함수이다.
- 이 경우 likelihood 함수는 다음과 같이 쓸 수 있다.

$$L=\prod \theta(w^T x_i)^{y_i}\{1-\theta(w^T x_i)\}^{1-y_i}$$

로그 변환

$$lnL=\sum y_i\, ln\{\theta(w^T x_i)\}+\sum (1-y_i) ln\{1-\theta(w^T x_i)\}$$

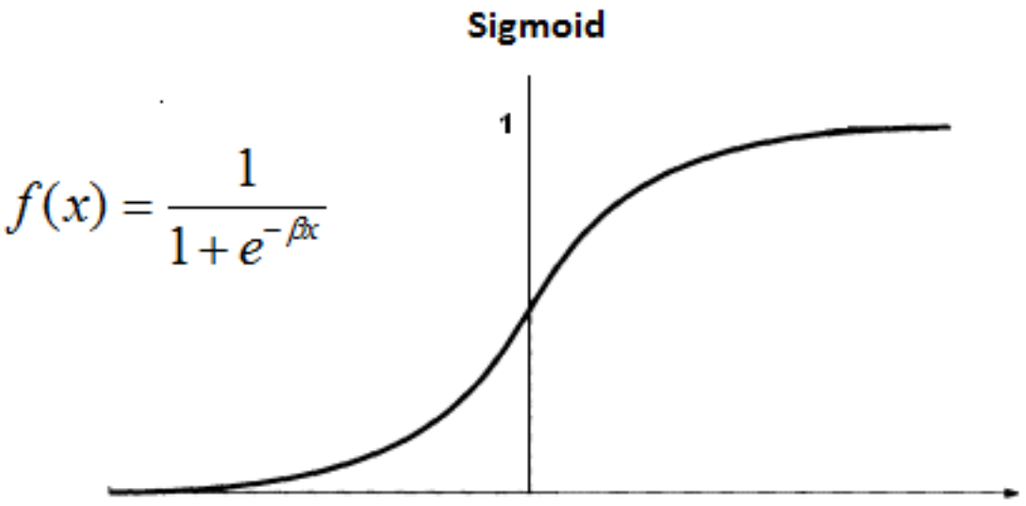
3. 로지스틱 회귀

로지스틱 회귀

로지스틱 함수

- 결과 값이 항상 0과 1사이의 값만 가지도록 변형한다.
  - 이런 형태로 최대값과 최소값에 정해져 있는 함수들을 시그모이드 함수라고 한다.

$$\theta(x;w)=\frac{1}{1+exp(-w^{T}x)}$$



다음에
또!
같이!
만나요!

learnit Flip your study