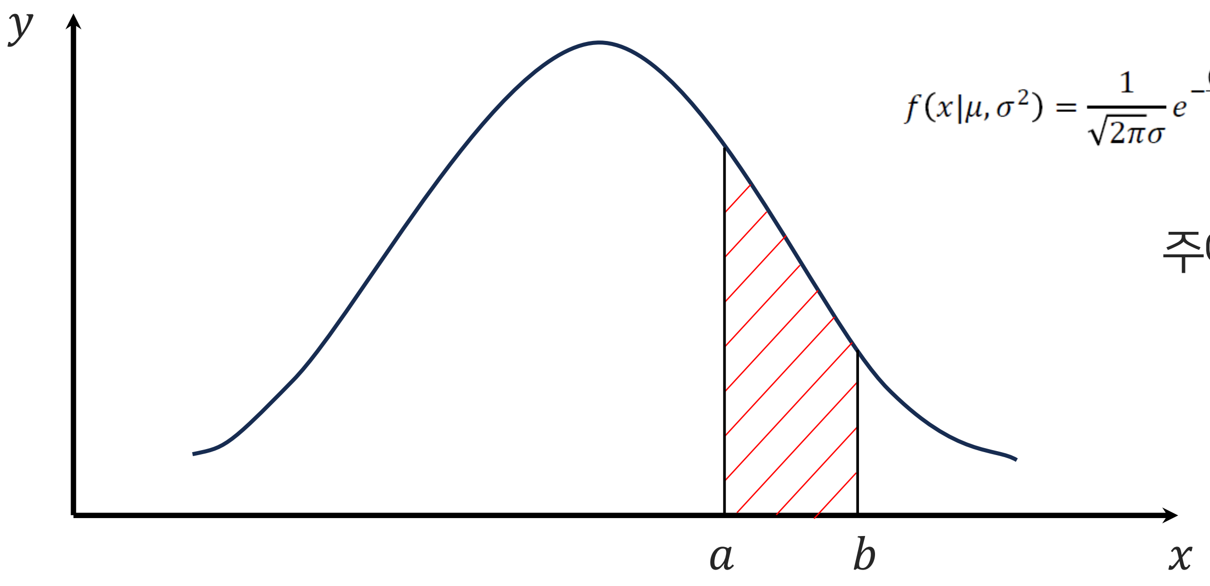
선행지식 - 확률 (Probability)

- '확률'은 주어진 확률분포가 있을 때 관측값 또는 관측 구간이 분포 안에서 어느 정도의 확률로 존재하는지를 나타냄

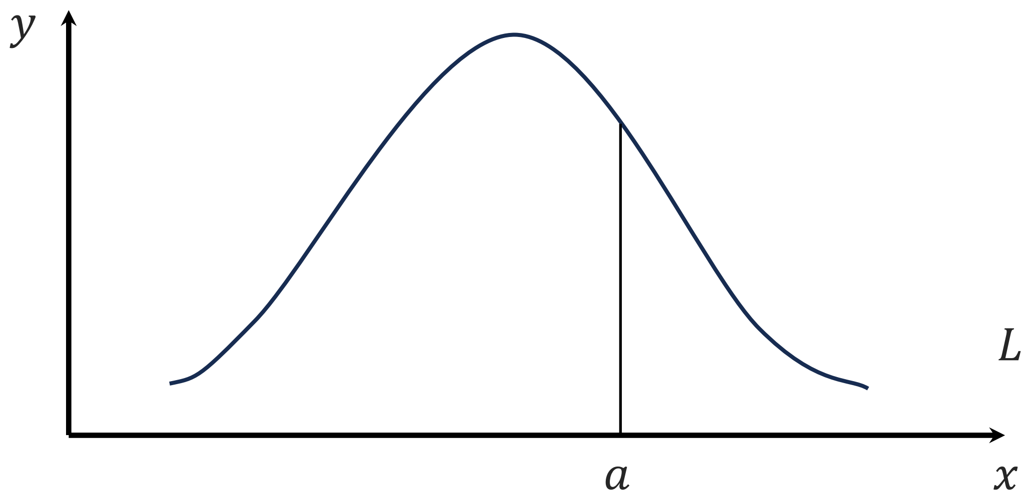


$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

주어진 구간의 넓이는?

선행지식 - 우도 (Likelihood)

- '우도(가능도)'는 관측값들이 주어져 있을 때 이들이 해당 확률 분포에서 왔을지에 대한 확률
- 연속확률밀도함수 pdf의 y값으로 측정



$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

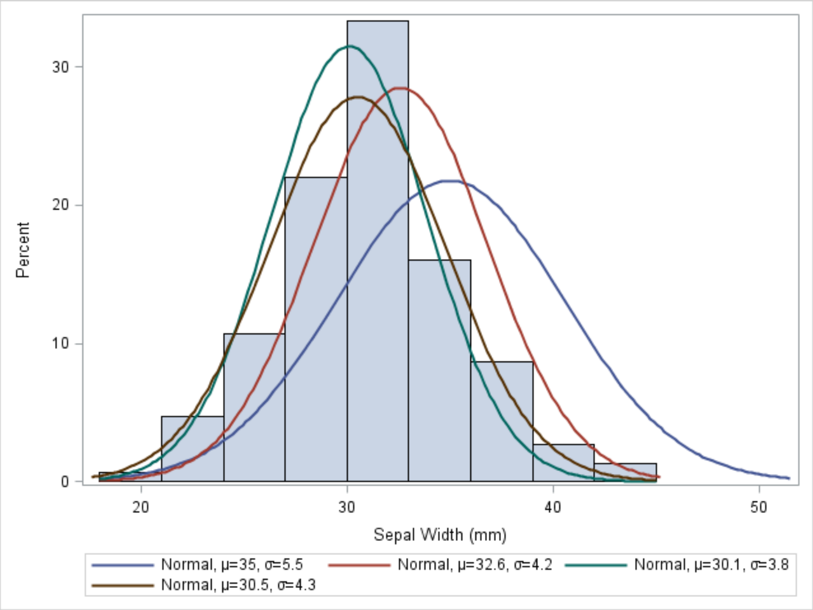
$\mu = 3, \sigma = 0.01$이라 할 때,

a가 해당 확률분포에서 나왔을 확률?

$$L(\theta = k|X) = \Pr(\theta = k|X \sim f(x,\theta), X = x_1, x_2, \dots, x_n)$$

선행지식 - 최대우도추정법 (Maximum Likelihood Estimation)

- 관측값들이 존재할 때, 그렇게 관측될 가능성이 가장 큰 확률분포를 찾아내는 방법
- '우도(가능도)'를 최대화하는 확률분포의 모수를 추정하는 방법



최대우도추정법의 계산

- 가정 : 관측치는 서로 i.i.d (independent identically distributed)하다.
- 위 가정 하에, 전체 데이터의 우도는 각 관측치들의 우도의 곱으로 표현 가능
- 계산을 쉽게 하기 위해 우도에 Log를 취해 줄 때도 있음
- 우도의 모수로 편미분하였을 때 그 값이 0이 되는 모수가 가장 적합한 모수로 선정됨

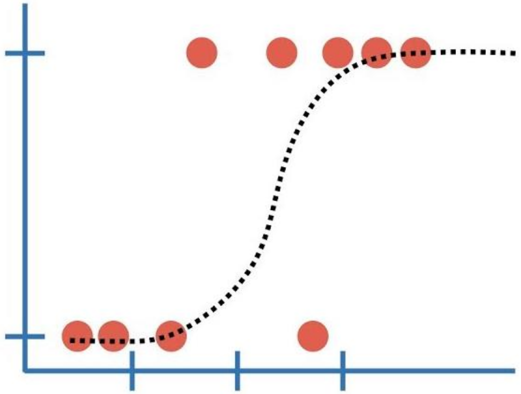
최대우도추정법의 계산 - 예시

[문제]
주머니 안에 검은 구슬과 흰 구슬을 합쳐 9개가 있다.
3개의 구슬을 뽑았는데, 검은 구슬이 2개, 흰 구슬이 1개 있었다.
이 때 주머니 안에 검은 구슬은 몇 개 정도가 있었을까?

[풀이]
$L(\theta = k|X) = \frac{k}{9}$ (상수함수) -> 검은 구슬을 뽑을 확률을 식으로 표현
편의상 $L(\theta = k|X) = p$라 하자.
문제 경우의 전체 우도 $L = 3C2p^2(1-p) = 3p^2(1-p)$
$\ln L = \ln 3 + 2\ln p + \ln(1-p)$
$(\ln L)' = \frac{2}{p} - \frac{1}{1-p}$ <- 이 식을 0으로 만드는 p의 값=2/3
따라서 k=6을 얻을 수 있음.

로지스틱 회귀 (Logistic Regression)

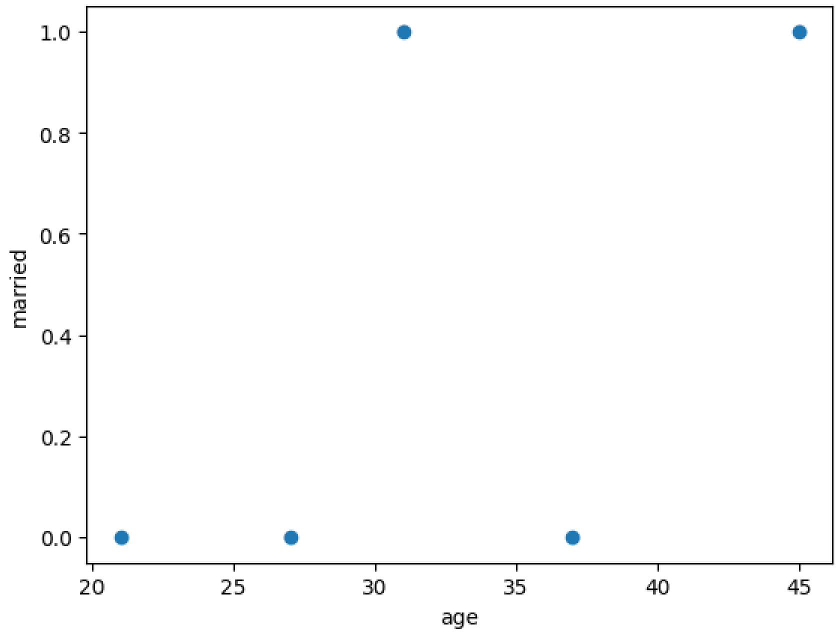
- 종속변수 (Y값)이 범주형 변수일 때 사용하는 분류 (Classification) 기법
- 이진 분류를 푸는 대표적인 알고리즘 중 하나
- 종속 변수 값이 항상 범위 [0,1] 사이에 있도록 함
  -> odds ratio (오즈비, 승산) 과 logit 변환으로 얻음



문제 상황

다음과 같은 데이터가 있다고 가정하자.

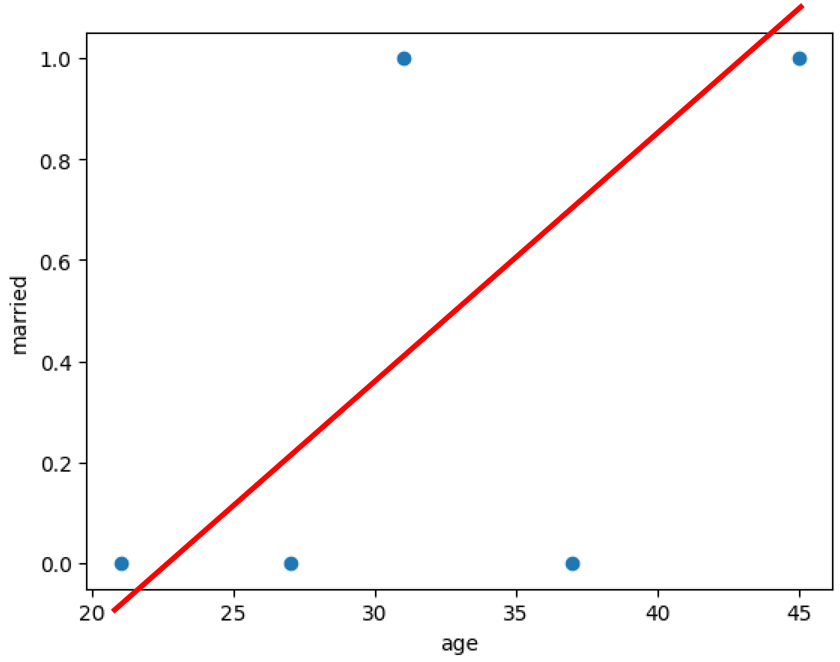
나이	결혼유무
21	0
27	0
31	1
37	0
45	1



문제 상황

선형회귀로는 적합할 수 없는 종류의 데이터

나이	결혼유무
21	0
27	0
31	1
37	0
45	1



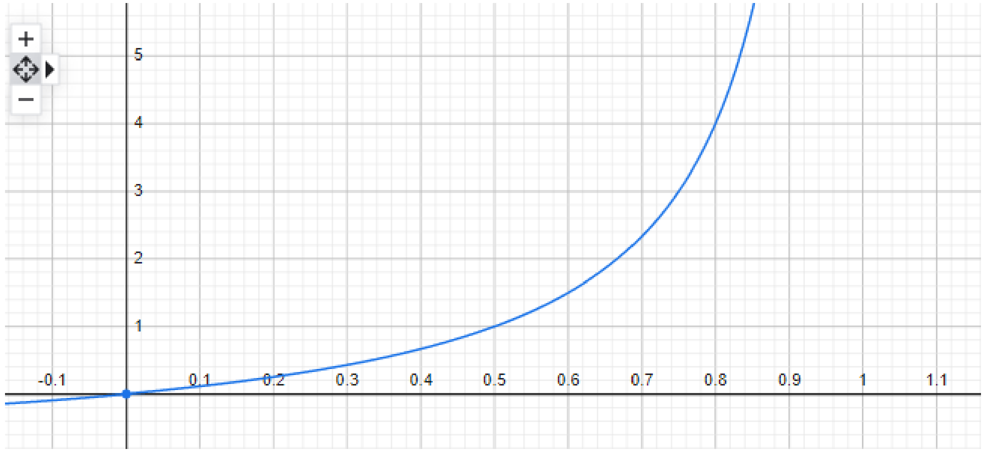
목표 : 회귀식에 나이를 대입하여 나온 값이 0.5를 넘으면 1, 넘지 않으면 0이라고 판단하기

Odds (오즈비, 승산)

본래 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ + \cdots + \beta_i x_i \ + \varepsilon$ 를 알아야 하지만, 이 경우에는 y가 범주형 변수이기 때문에 odds라는 새로운 개념을 도입함.

$odds \ = \frac{p(y=1|x)}{p(y=0|x)} = \frac{p(y=1|x)}{1-p(y=1|x)}$ -> y가 1일 확률과 y가 0일 확률의 비율

간단히, $odds\ = \frac{p}{1-p}$ 라고 지칭하면, 그래프는 아래와 같음



odds의 한계 :
1) $0 < odds < \infty$
2) 비대칭성 (Asymmetric)

logit 변환

logit 변환 : ±∞의 범위에서 어떤 클래스에 속할 확률을 결정하는 함수로 변환하는 것
쉽게 말해, odd에 log를 씌워 주는 것

$\log(odds) = \log\frac{p}{1-p}$

기존의 회귀식인 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon$ 를 $\log(odds)$ 로 치환하고자 함.

$\log(odds) = \log\frac{p}{1-p} = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon}_{\text{편의상 } a\text{로 지칭}}$

$\log\frac{p}{1-p} = a,$

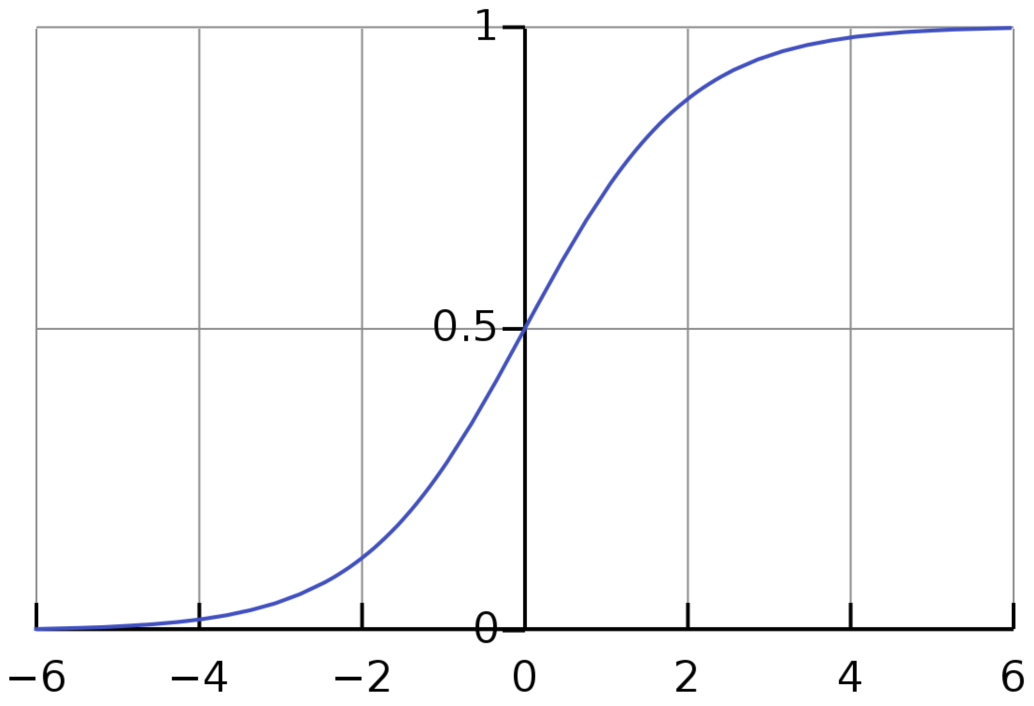
$\frac{p}{1-p} = e^a,$

p에 대해 정리하면

$p = \frac{1}{1+e^{-a}}$

Logistic function (Sigmoid function)

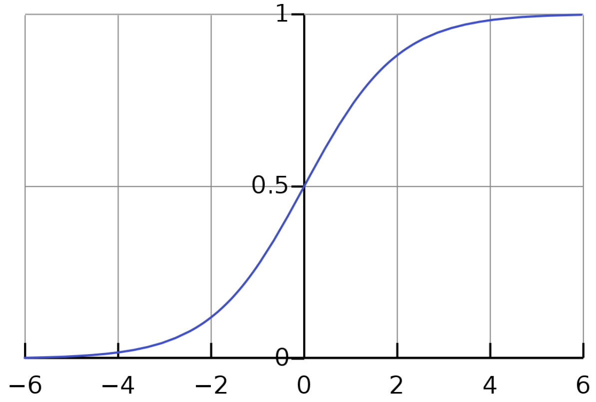
$$y = \frac{1}{1+e^{-x}}$$



Logistic function의 결정경계

$P(Y=1|X) > P(Y=0|X),$
$P(Y=1|X) > 1-P(Y=1|X),$
$\frac{P(Y=1|X)}{1-P(Y=1|X)} > 1,$
$\log\frac{P(Y=1|X)}{1-P(Y=1|X)} > 0,$
여기에서,
$\log(odds) = \log\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon$
결국 $\log(odds)$ 로 표현된 회귀식의 값이 0을 기준으로 결정경계가 생성됨.
그런데,
$P(Y=1|X) > 1-P(Y=1|X)$
이 식의 우변값을 좌변으로 넘겨 정리하면,
$P(Y=1|X) > 0.5$
다시 말해,
Logistic function의 값이 0.5 이상이면 1, 0.5 미만이면 0으로 분류한다.

$$y = \frac{1}{1+e^{-x}}$$



Logistic function의 모수 추정

로지스틱 회귀는 베르누이 시행(Bernoulli trial)을 전제로 하는 모델.
베르누이 확률변수 Y의 분포는

Y	0	1
$P(Y=y_i)$	$1-p$	p

수식으로 정리하면,

$$P(Y=y_i) = p^{y_i}(1-p)^{1-y_i} \quad (y_i = 0, 1)$$

베르누이 확률변수의 우도함수는 다음과 같이 표현 가능함.

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

Logistic function의 모수 추정

i개의 학습 데이터가 있고 정답이 0 또는 1인 이항 logistic model의 파라미터
가 β로 주어질 때, i번째 관측치에 대해서

$$P(x_i, y_i \mid \beta) = \begin{Bmatrix} \sigma(\mathrm{x}, \beta) & \text{if } y = 1 \\ 1 - \sigma(\mathrm{x}, \beta) & \text{if } y = 0 \end{Bmatrix}$$

우도함수는 다음과 같이 표현 가능함.

$$L(\mathrm{X}, y, \beta) = \prod_{i=1}^{R} \sigma(\mathrm{x_i}, \beta)^{y_i} \left(1 - \sigma(\mathrm{x_i}, \beta)\right)^{1-y_i}$$

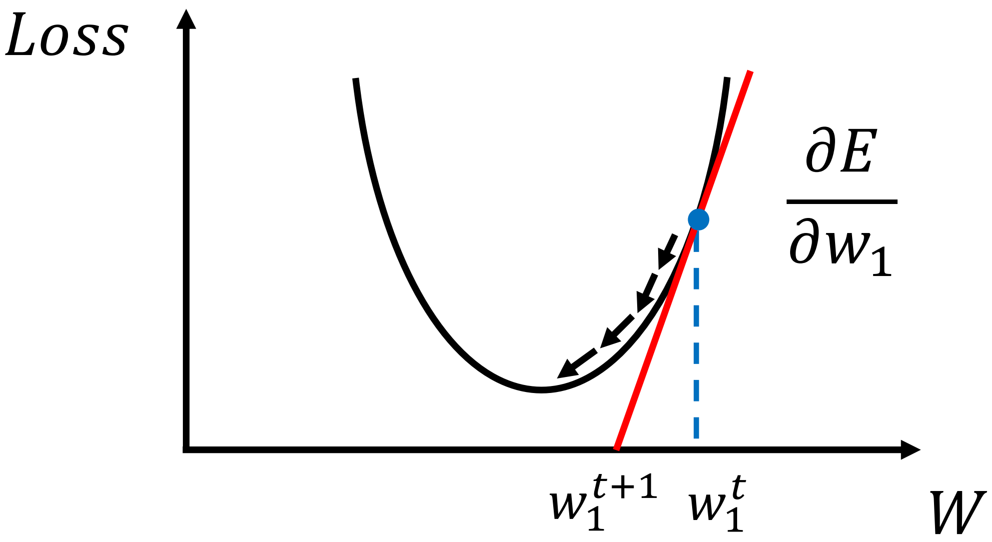
양변에 로그를 취하면,

$$\ln L(\mathrm{X}, y, \beta) = \sum_{i=1}^{R} y_i \ln\left(\sigma(\mathrm{x_i}, \beta)\right) + (1 - y_i) \ln\left(1 - \sigma(\mathrm{x_i}, \beta)\right)$$

이 식을 최대우도추정법을 이용해 풀어야 하는데, 위 식은 회귀계수 β에 대해 비선형적인 식임.
따라서 명시적인 해를 바로 구할 수 없기 때문에 주로 쓰는 방법은...

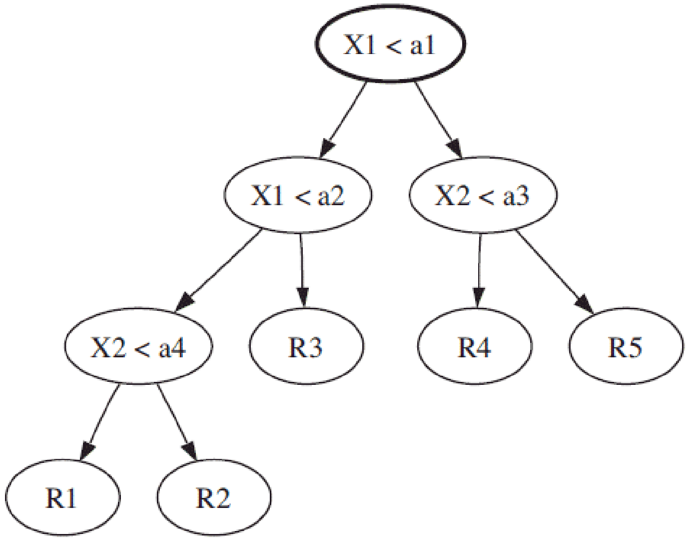
Gradient Descent (경사 하강법)

임의의 한 지점으로부터 시작해, loss가 줄어드는 방향으로
w (weights=parameters)들을 갱신하는 방법



Decision Tree (의사결정나무)

- 일련의 분류 규칙을 통해 데이터를 분류, 회귀하는 지도 학습 모델 중 하나
- 분류 규칙(decision rule)을 나무 구조로 도표화한 방법
- 설명력이 뛰어나지만 과적합 문제가 자주 발생

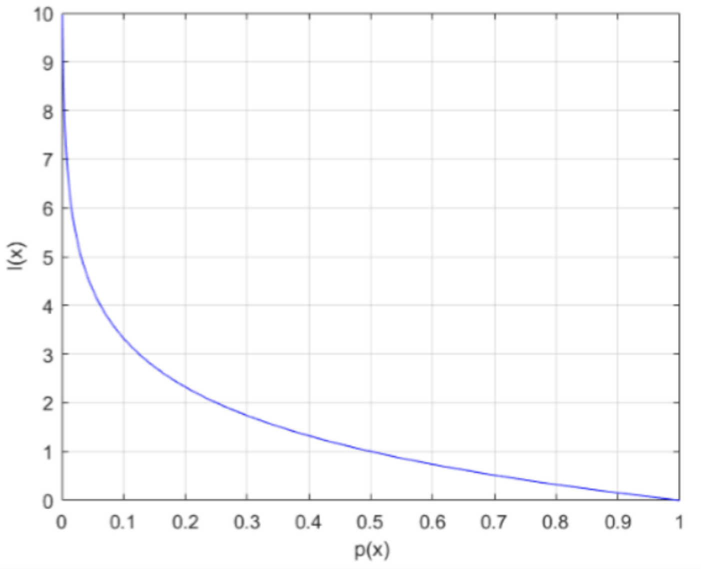


정보량 (Information content)

- 어떤 사건이 가지고 있는 정보의 양
- 정보량이 높을수록 사건이 발생할 확률은 희박해짐

$$I(x) = \log_2 \frac{1}{p(x)} = -\log_2 p(x) \,,$$

$p(x)$: 사건 x의 발생확률



엔트로피(Entropy)

- 불순도 (Impurity)를 측정하는 지표, 정보량의 기댓값
- 섀넌 엔트로피 (Shannon entropy) 라고도 불림

$$Entropy(S) = \sum_{i=1}^{c} p_i \times I(x_i),$$

S: 사건, c: 전체 사건의 갯수

지니 계수 (Gini Index)

- 불순도 (Impurity)를 측정하는 또다른 지표
- 데이터의 통계적 분산 정도를 정량화해서 표현한 값

$$G(S) = 1 - \sum_{i=1}^{c} p_i^2 \, ,$$

S: 사건, c: 전체 사건의 갯수

Decision Tree의 분류 방법

1. 목표 속성의 불순도를 구한다.
2. 각 데이터 종류별로 불순도를 구한다.
3. 각 데이터 종류별로 Information gain을 계산한다.
4. 가장 Information gain이 높은 분류 기준으로 데이터를 분할한다.
5. 1~4를 반복한다.

Decision Tree - 예시

Smoker	Sex	Age	HDattack (HD)
1	0	22	1
1	1	22	0
0	0	50	1
1	0	50	1
0	0	68	0

- 5개의 데이터

  X: Smoker, Sex, Age

  y: HDattack

Decision Tree - 예시

Smoker	Sex	Age	HDattack (HD)
1	0	22	1
1	1	22	0
0	0	50	1
1	0	50	1
0	0	68	0

1. 목표 속성의 불순도를 구한다.

$$E(HD) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.9710$$

Decision Tree - 예시

Smoker	Sex	Age	HDattack (HD)
1	0	22	1
1	1	22	0
0	0	50	1
1	0	50	1
0	0	68	0

2. 각 속성에 대해 분할 후 불순도 계산

2-1) Smoker

$$E(HD|smoker)$$

$$= P(smoker = 1) \times E(HD|smoker = 1)$$

$$+ P(smoker = 0) \times E(HD|smoker = 0)$$

$$= \frac{3}{5}\left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right)$$

$$+ \frac{2}{5}\left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) = 0.9509$$

Decision Tree - 예시

Smoker	Sex	Age	HDattack (HD)
1	0	22	1
1	1	22	0
0	0	50	1
1	0	50	1
0	0	68	0

2. 각 속성에 대해 분할 후 불순도 계산

2-2) Sex

$$E(HD|sex)$$
$$= P(sex = 1) \times E(HD|sex = 1) + P(sex = 0)$$
$$\times E(HD|sex = 0) = 0.6490$$

Decision Tree - 예시

	Smoker	Sex	Age	HDattack (HD)
	1	0	22	1
(1)	1	1	22	0
	0	0	50	1
(2)	1	0	50	1
	0	0	68	0

$(1)\ mean(22{,}50) = 36$

$(2)\ mean(50{,}68) = 59$

2. 각 속성에 대해 분할 후 불순도 계산

* 연속형 변수의 경우 오름차순으로 정렬 후 label이 바뀌는 점을 기준점으로 하여 계산한다.

2-3) Age - (1)의 경우

$E(HD|Age = 36)$

$= P(Age < 36) \times E(HD|Age < 36)$

$+ P(Age > 36) \times E(HD|Age > 36) = 0.9509$

Age - (2)의 경우

$E(HD|Age = 59) = 0.6490$ -> 0.6490 선택

Decision Tree - 예시

Smoker	Sex	Age	HDattack (HD)
1	0	22	1
1	1	22	0
0	0	50	1
1	0	50	1
0	0	68	0

3. Information gain계산

2-1) $E(HD) - E(HD|smoker) = 0.9709 - 0.9509 = 0.02$

2-2) $E(HD) - E(HD|sex) = 0.9709 - 0.6490 = 0.3219$

2-3) $E(HD) - E(HD|age = 59) = 0.9709 - 0.6490 = 0.3219$

Decision Tree - 예시

Smoker	Sex	Age	HDattack (HD)
1	0	22	1
1	1	22	0
0	0	50	1
1	0	50	1
0	0	68	0

4. 가장 Information gain이 높은 분류 기준으로 데이터 분할

2-1) $E(HD) - E(HD|smoker) = 0.9709 - 0.9509 = 0.02$

2-2) $E(HD) - E(HD|sex) = 0.9709 - 0.6490 = 0.3219$

2-3) $\boldsymbol{E(HD) - E(HD|age = 59) = 0.9709 - 0.6490 = 0.3219}$

Decision Tree - 예시

Smoker	Sex	Age	HDattack (HD)
1	0	22	1
1	1	22	0
0	0	50	1
1	0	50	1
0	0	68	0

반복횟수:

1) 모든 노드의 불순도가 0이 될 때까지

2) 최대 깊이(max depth) 지정

3) 최대 가지 노드 (max leaf nodes) 지정

4) 최소 노드 수 (min sample leaf) 지정 등

plotting

Smoker	Sex	Age	HDattack (HD)
1	0	22	1
1	1	22	0
0	0	50	1
1	0	50	1
0	0	68	0