Linear Regression 연속적인 데이터 $(-\infty, \infty)$

$$y = h_\theta(x) = W_0 + W_1 x_1 + W_2 x_2 + \cdots$$

수학에서의 함수와 다르게 여기서는 $W_0, W_1, W_2, \cdots$ 이 미지수 역할을 하고 데이터 (x, y)를 넣어 가장 적합한 W_i를 구하는 방식

(가장 적합한 = 손실 함수를 최소화)

손실 함수

$$L(w) = \sum \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

$y^{(i)}$: 실제값, $\hat{y}^{(i)}$: 예측값

Logistic Regression 이분적인 데이터 (0 or 1)

Linear Regression의 식과 동일해지려고 하면

$$\ln\left(\frac{P}{1-P}\right) = W_0 + W_1 x_1 + \cdots$$

이후 나올 결과값 P를 알아보기 쉽게 하면

$$P = \frac{1}{1 + e^{-(W_0 + W_1 x_1 + \cdots)}}$$

P: 1이 나올 확률 $1-P$: 0이 나올 확률.

손실 함수

$$L(w) = \sum \left(-y^{(i)} \log\left(\hat{y}^{(i)}\right) - \left(1-y^{(i)}\right) \log\left(1-\hat{y}^{(i)}\right)\right)$$

학습 성능

K-fold 교차검증

데이터셋을 k개로 나눠서 1개는 테스트, 나머지는 훈련용으로 k번 학습을 진행 후 정확도 비교

S_1 S_2 S_3 S_4 S_5
S_1 S_2 S_3 S_4 S_5
S_1 S_2 S_3 S_4 S_5
S_1 S_2 S_3 S_4 S_5
S_1 S_2 S_3 S_4 S_5

과적합

해결책

prepruning 기준점보다 낮은 향상을 보일 경우에는 더 진행하지 않는다

postpruning 가지치는 것이 더 낮은 오류율로 이끌 경우, 가지 친다

Instance-Based Learning

본인 데이터와 비슷한 데이터들을 비교 후 결과 내린다