머신러닝(Machine Learning)

지도학습 알아보기(logistic)

학습 목표

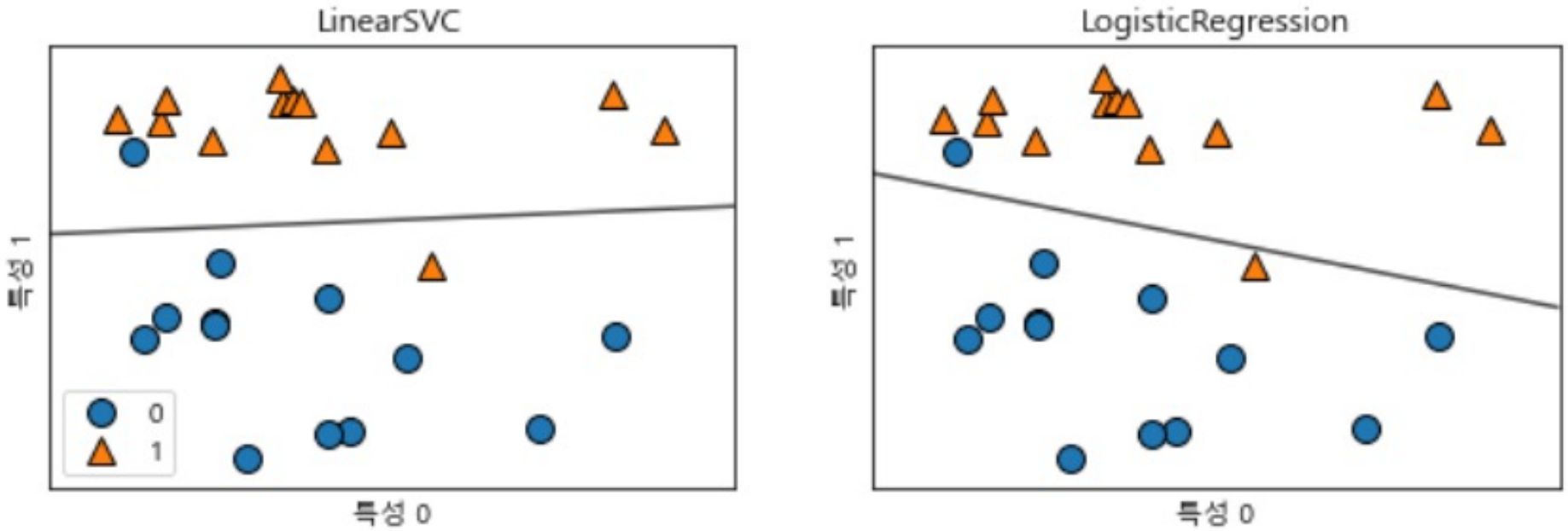
- 로지스틱 회귀에 대해 알아봅니다.
- 분류용 선형 모델에 대해 알아봅니다.

목차

범주형 변수에서 회귀 모델을 사용하려면 어떻게 해야 할까?

01 분류용 선형 모델

어떻게 분류되는가?



- LinearSVC와 LogisticRegression으로 만든 결정 경계가 직선으로 표현됨.

- 직선을 기준으로 위쪽은 클래스 1, 아래쪽은 클래스 0으로 나눠짐.

01 분류용 선형 모델

- 선형 분류기는 선, 평면, 초평면을 사용해서 두 개의 클래스를 분류하는 분류기

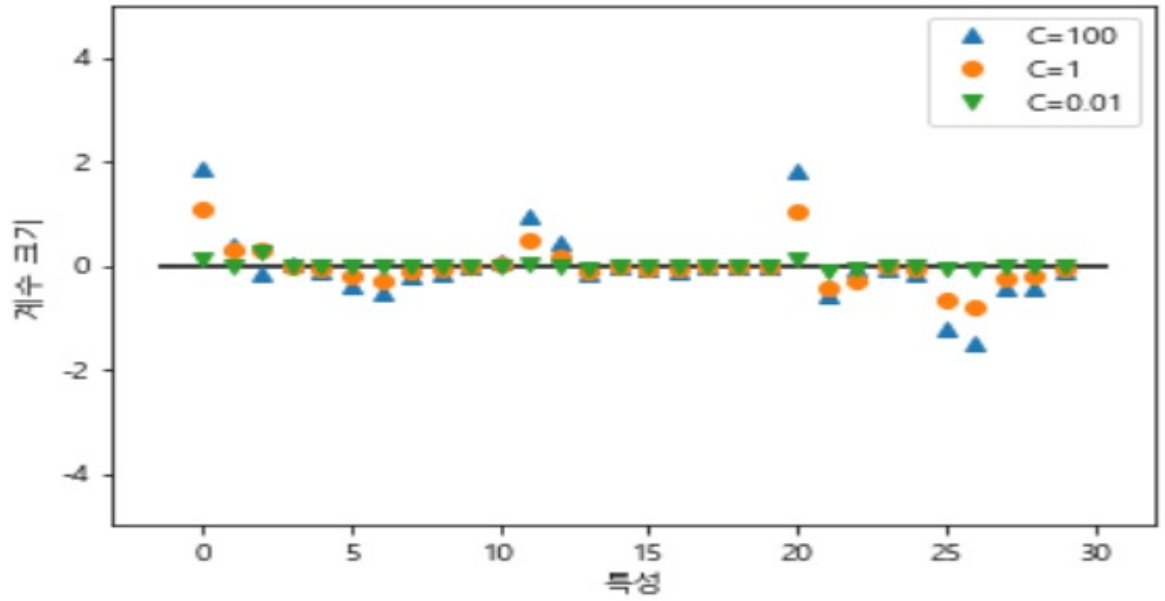
- 두 개의 선형 분류 알고리즘

  (1) 로지스틱 회귀(Logistic Regression)

  (2) 서포트 벡터 머신(support vector machine)

01 분류용 선형 모델 - 일반화 시키기

규제를 결정하는 매개변수 C



- C의 값이 작을수록 규제가 크고, C값이 커지면 규제가 줄어든다. 반비례

01 분류용 선형 모델 – 종류

- **로지스틱 회귀(logistic regression)**
- **서포트 벡터 머신(support vector machine)**

다중 클래스 분류용 선형 모델

A. 로지스틱 회귀만을 제외하고 많은 선형 분류 모델은 태생적으로 이진 분류만을 지원

B. 이진 분류 알고리즘은 다중 클래스 분류 알고리즘으로 확장하는 보편적 기법은 일대다 방법이다.

C. 일대다 방식은 각 클래스를 다른 모든 클래스와 구분하도록 이진 분류 모델을 학습

D. 클래스의 수만큼 이진 분류 모델이 만들어진다.

E. 예측을 할 때, 이렇게 만들어진 모든 이진 분류기가 작동하여 가장 높은 점수를 내는 분류기의 클래스를 예측값으로 선택

01 분류용 선형 모델 - 장단점

▶️ 장점

A. 선형 모델은 학습 속도가 빠르고 예측도 빠르다.

B. 매우 큰 데이터셋과 희소한 데이터셋에도 잘 작동한다.

C. 샘플에 비해 특성이 많을 때 잘 동작.

▶️ 단점

A. 저 차원 데이터셋에서는 다른 모델들의 일반화 성능이 좋음.

장단점과 매개변수

A. 선형 모델의 주요 매개변수는 회귀 모델은 alpha이다.

B. LinearSVC와 LogisticRegression에서는 C가 매개변수이다.

- C와 alpha는 로그 스케일(0.01, 0.1, 1, 10)으로 최적치를 결정.

C. L1규제를 사용할지, L2규제를 사용할지 결정해야함.

D. LogisticRegression과 Ridge에 대용량 데이터 셋이라면 solver='sag'
옵션을 줄 수 있다. 또는 다른 대안으로 SGDClassifier와 SGDRegressor를 사용 가능.

02 로지스틱 회귀(logistic regression)

▶ 로지스틱 회귀(logistic regression)

예측하고자 하는 Y가 범주형(categorical) 일때는 다중 선형 회귀 모델을 그대로 적용할 수 없다.

=> 로지스틱 회귀 모델 제안

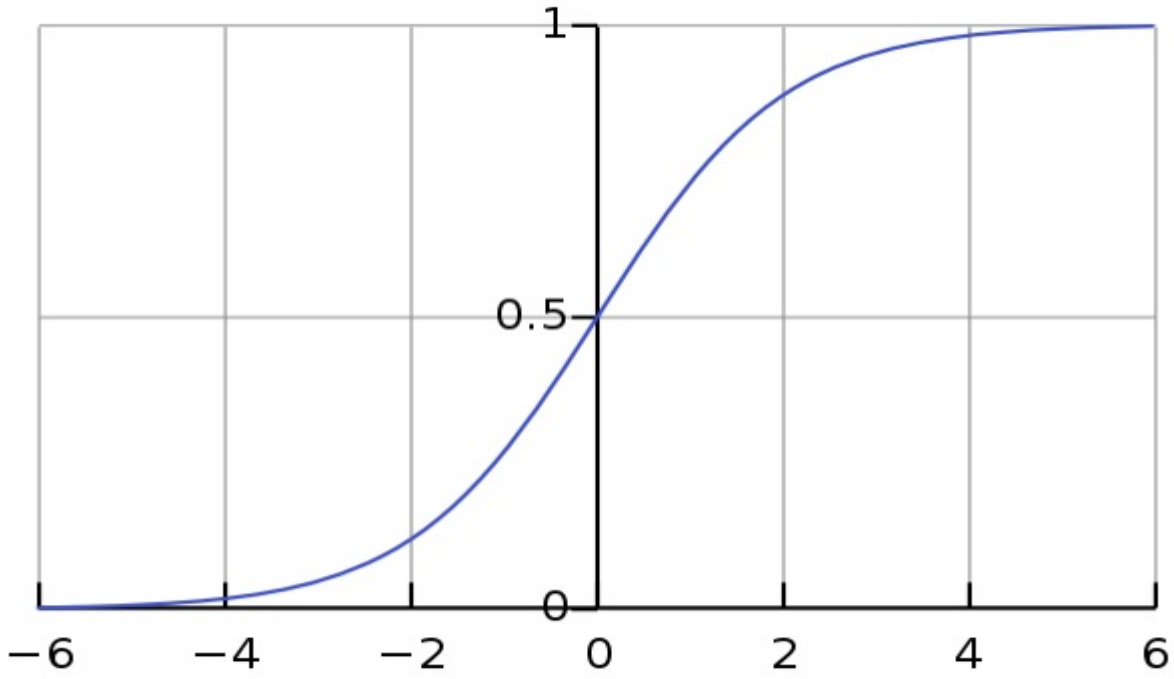
=> 영국의 통계학자인 D.R.Cox가 1958년에 제안한 확률 모델

02 로지스틱 회귀(logistic regression)

▶ 로지스틱 회귀(logistic regression)

회귀를 사용하여 데이터가 어떤 범주에 속할 확률을 **0과 1사이의 값으로 예측**한다.

로짓 변환을 통해 직선형 Regression을 곡선형으로 적합시킨다.(Fitting)



위키 백과 참조

02 로지스틱 회귀(logistic regression)

▶ 선형 회귀

$\hat{y}_i$ = w[0] * x[0] + w[1] * x[1] + ... + w[p] * x[p] + b

▶ 로지스틱 회귀(logistic regression)

$P(Y = 1)$ = w[0] * x[0] + w[1] * x[1] + ... + w[p] * x[p] + b

A. P(Y=1)의 범위는 0~1인데, w[0] * x[0] .. 의 범위는 -∞ ~ ∞ 로 맞지 않음.

=> (해결) P(Y=1)의 범위를 -∞ ~ ∞로 바꾸기

=> (해결) P(Y=1)에 Logit를 적용시키기

02 로지스틱 회귀(logistic regression)

▶ 로짓이란 무엇인가?

Logit = Log Odds

▶ Odds란 무엇인가?

$$\text{Odds} = \frac{\text{성공할 확률}}{\text{실패할 확률}} = \frac{P}{1-P}$$

$$\text{Logit(Odds)} = \log\left(\frac{P}{1-P}\right)$$

Odds ∈ (0, ∞)

P ∈ (0, 1)

Log Odds ∈ (-∞, ∞)

02 로지스틱 회귀(logistic regression)

▶ P(Y=1)에 로짓 적용하기

$$\text{Logit}(P(Y=1)) = ax + b = \log\left(\frac{P}{1-P}\right)$$

$$z = \text{Logit}(P) = \log\left(\frac{P}{1-P}\right) = ax + b$$

02 로지스틱 회귀(logistic regression)

▶ 로지스틱 함수는 Logit 함수의 역함수이다.

$$\log\left(\frac{P}{1-P}\right) = ax + b$$

$$\Rightarrow \frac{P}{1-P} = e^{ax+b}$$

$$\Rightarrow P = (e^{ax+b}) * (1 - P)$$

$$\Rightarrow P(1 + e^{ax+b}) = e^{ax+b}$$

$$\Rightarrow P = \frac{e^{ax+b}}{1+e^{ax+b}} = \frac{e^{z}}{1+e^{z}} = \frac{1}{1+e^{-z}}$$