누구나 할 수 있는!

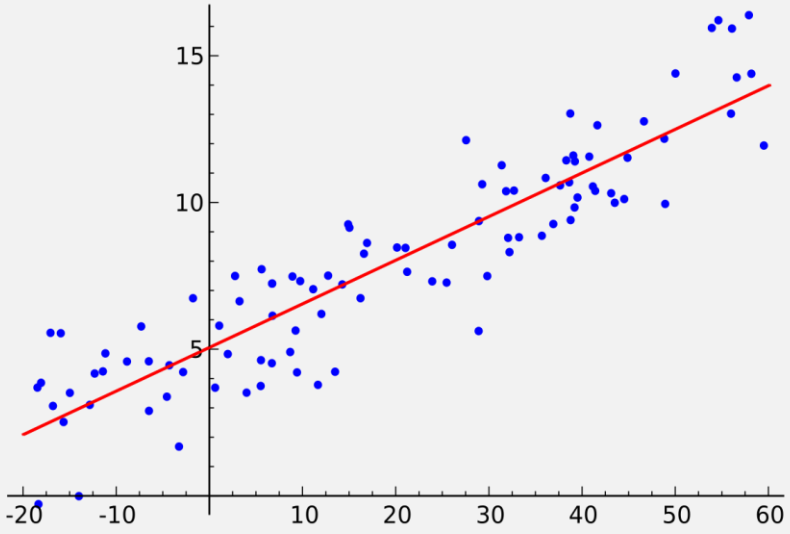
인터페이스 AI 오픈 스터디

3주차. 선형분류

선형회귀 복습

- 선형회귀에서, 우리는 X에 대한 올바른 Y값을 찾는 모델을 만들었음

$$Y = WX + b$$



이미지출처

여러 입력 변수에 대한 가설 이거 지난주에 했어야 했는데.. 죄송합니다 ㅠ

- 입력 X가 하나의 스칼라가 아닌 벡터라면?
  - 해당 날짜의 (온도, 습도)에 따른 아이스크림 판매량!
  - (공부시간, 논 시간)에 따른 성적!
- 행렬의 곱을 사용하면 된다!

$$H(x) = (x_1, x_2, x_3) \cdot \begin{matrix} w_1 \\ w_2 \\ w_3 \end{matrix} + b$$

- X는 (1,3) 행렬, W는 (3,1)행렬

이진분류 (Binary Classification)

- 객체가 A, B 둘 중 무엇인지 분류하는 문제
  - 객체가 고양이인가? 강아지인가? 연필인가? 볼펜인가?
- 로지스틱 회귀라고도 불림
  - 회귀랑 비슷한 가설을 사용!

로지스틱 회귀

- 입력 X에 대해 Y가 0~1 사이의 값을 출력.
  - 값이 0에 가까우면 A로 판단.
  - 값이 1에 가까우면 B로 판단.

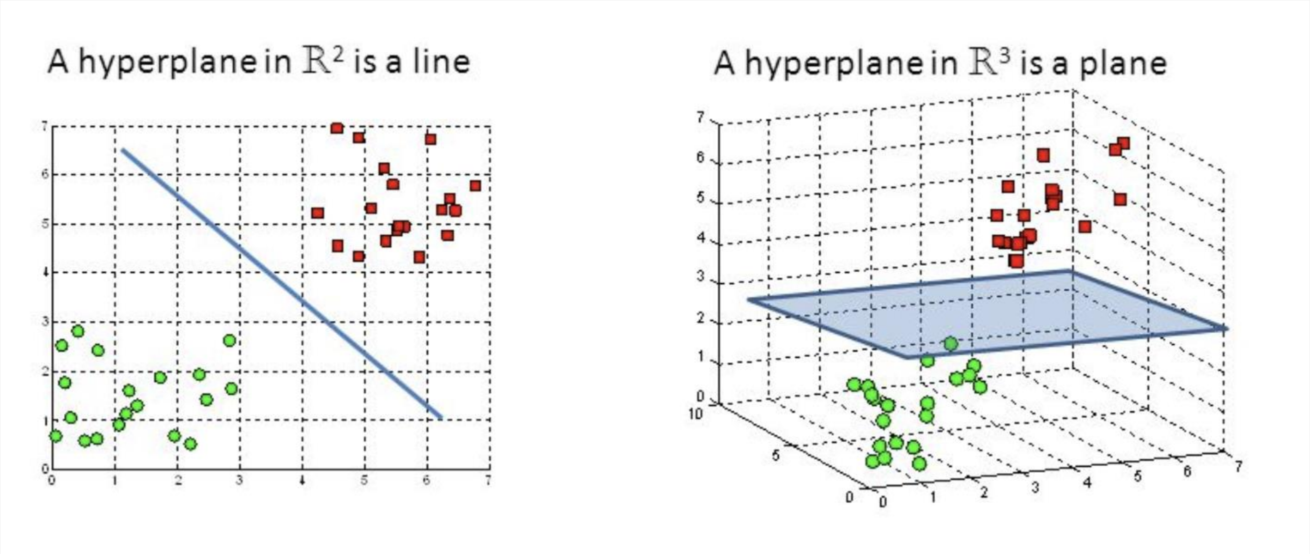
Y=0.21
79% 확률로 고양이!

Y=0.83
83% 확률로 강아지!

회귀 문제의 가설에서, 출력을 0~1로 제한해주기만 하면 됨!

선형 분류의 의미

- 선형 이진 분류란, 데이터를 두 부류로 나누는 하나의 초평면 (hyperplane)을 찾는 것으로 볼 수 있다.



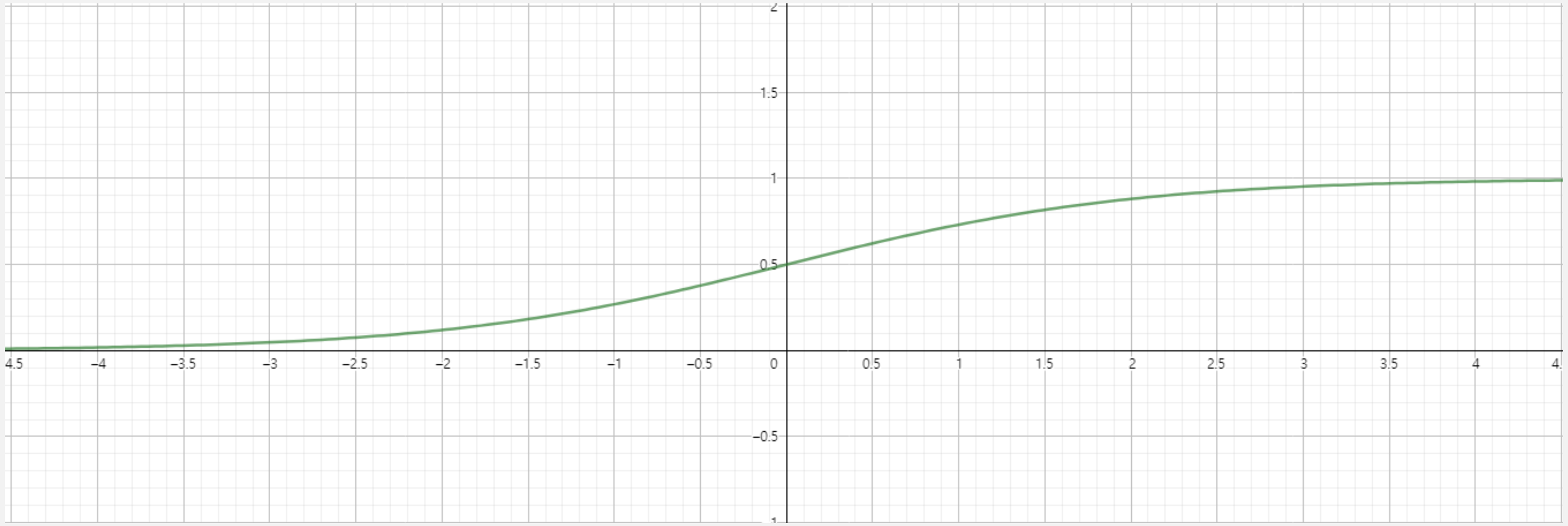
https://deepai.org/machine-learning-glossary-and-terms/hyperplane

Sigmoid 함수

- S자 모양 함수라는 뜻

$$f(x) = \frac{1}{1 + e^{-x}}$$

- (−∞, ∞) 전체 구간에서 0과 1 사이의 결과를 가짐



로지스틱 회귀 가설

- 입력 X에 대한 로지스틱 회귀의 가설은 아래와 같다.

$$Y = Sigmoid(WX + b)$$

- 출력은 0~1 사이의 값을 갖는다.

비용 함수

- 값이 0~1로만 출력되기 때문에, 기존에 쓰던 MSE를 쓰면 문제가 발생
  - 예측한 값에 대한 오차가 가장 크다고 해봤자 Y−T=1 일 때이다.
  - MSE = $(1)^2 = 1$
- 새롭게 교차 엔트로피(Cross Entropy)를 써보자!

이진 교차 엔트로피

$$\mathrm{BC}E = -\frac{1}{n}\sum_{i=1}^{n}(y_i\log(t) + (1-y_i)\log(1-t))$$

- 일단 여러 입출력에 대해 구해야 하기때문에 평균을 취한다.
- 가운데 +를 기준으로 좌우를 나눠보자.
- 정답이 1일 때는 좌측만 의미가 있고, 정답이 0일 때는 우측만 의미가 있다.
  - y_i과 $(1-y_i)$의 의미
- 중요한 건 log 부분이다.

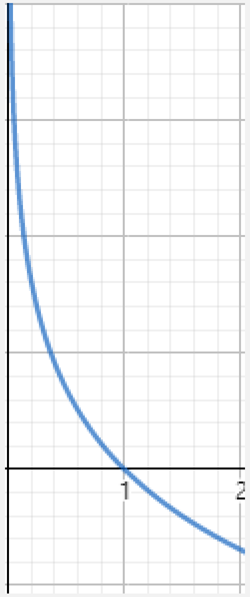
엔트로피: 놀람의 정도

- 엔트로피란 확률적으로 발생하는 사건에 대한 정보량을 의미한다.
- 우리가 확률적인 사건을 겪고 놀라는 정도를 수치로 나타낸다고 생각해보자.
  - 동전을 던져서 앞면이 나오더라도, 확률은 ½ 이므로 그렇게 놀라지 않을 것이다.
  - 반면, 로또 1등이 당첨될 확률은 굉장히 낮으므로 많이 놀라울 것이다.
  - 가장 단순한 방법은 확률의 역수를 취하는 것이다. (동전 앞면의 놀라움: 2, 로또 1등: 100만쯤?)
  - Log를 취하면, 예측값과 실제 사건 사이의 정보량을 구할 수 있다.
- 놀람이란, 우리가 예측한 값(t; 동전이 뒷면일 것이다. 로또 꽝일 것이다.)과 실제 사건(y;동전 앞면, 1등 당첨) 사이의 차이이다.
- t와 y가 많이 다를수록 entropy가 커진다.

다시 교차 엔트로피로 돌아와서…

$$\mathrm{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(t) + (1 - y_i)\log(1 - t))$$

- Y=1, t=0.7인 경우를 생각해보자.
- 우측은 (1−y) = 0 이므로 사라진다.
- $y_i \log(t)$에서 log(0.7)은 −0.15이다.
- t가 좀 더 좋아져서 0.9이면 log(0.9)는 −0.04이다



이진 분류 (Binary Classification)

1. Y를 0과 1로 매핑한다. (0은 A, 1은 B)
2. Sigmoid 함수를 이용한 가설 설정
3. Binary Cross Entropy를 이용한 오차 측정
4. 학습!

Multi-Class Classification

- 이제 여러 개의 클래스를 구분해보자!
  - 고양이, 강아지 + 사람 …
  - 연필, 볼펜 + 샤프 + 지우개 …
- 이진 분류에서 몇 가지만 수정하면 된다!

원-핫 인코딩 (one-hot encoding)

- 데이터가 어느 클래스에 속하는지를 이진수로 매핑하는 방법

고양이	강아지	사람	앵무새
0	0	1	0

고양이	강아지	사람	앵무새
1	0	0	0

고양이	강아지	사람	앵무새
0	1	0	0

가설 수정

- 0~1 사이의 값 n개(클래스 개수)를 출력하도록 수정!
- Sigmoid의 변형인 Softmax 함수를 사용!

$$f(x_t) = \frac{e^{x_t}}{\sum_{i=1}^{n} e^{x_i}}$$

Softmax 함수

- Sigmoid 함수의 일반화
- 모든 출력에 일단 exp를 씌운다.
- 현재 출력(예를 들어, 고양이일 확률)/(전체출력의 합)을 한다.
- 각 클래스에 대한 확률이 0~1사이로 나오고, 모든 출력을 더하면 1이 된다.

- 예시 출력(0.5, 0.2, 0.3) <= 클래스 A일 확률이 50%로 가장 높음!

Cross Entropy

- 우리가 위에서 배운 교차 엔트로피는 Binary CE라 하는 이진 분류 문제용 교차 엔트로피 함수이다.
- 다중 분류를 위한 교차 엔트로피는 아래와 같다.

$$CE = -\sum_{i=0}^{n} y'_i \log(t_i)$$

- 여기서 n은, 클래스의 개수
- 각 클래스별로 오차를 계산해서 합산하는 것이다!

오늘의 키워드

- 이진 분류(Binary Classification)
  - Logistic Regression
  - Sigmoid Function
  - Binary Cross Entropy Error
- 다중 분류(Multi-Class Classification)
  - Softmax Function
  - Cross Entropy Error

감사합니다!