4장 로지스틱 회귀

감사의 글

자료를 공개한 저자 오렐리앙 제롱과 강의자료를 지원한 한빛아카데미에게 진심어린 감사를 전합니다.

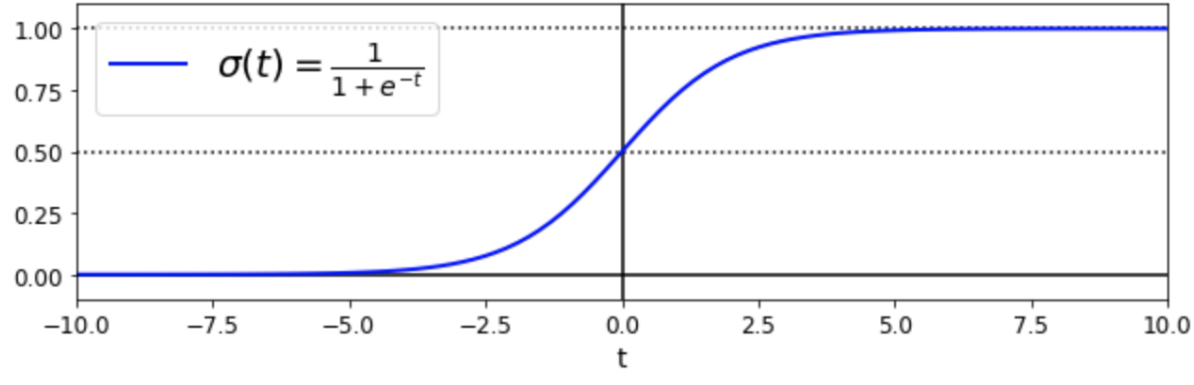
회귀 모델을 분류 모델로 활용할 수 있다.

- 이진 분류: 로지스틱 회귀
- 다중 클래스 분류: 소프트맥스 회귀

확률 추정

- 시그모이드 함수

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$



- 로지스틱 회귀 모델에서 샘플 $\mathbf{x}$가 양성 클래스에 속할 확률

$$\hat{p} = \sigma(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n)$$

예측값

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} \geq 0.5 \end{cases}$$

- 양성 클래스인 경우:

$$\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n \geq 0$$

- 음성 클래스인 경우:

$$\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n < 0$$

- 예측을 잘 하는 파라미터 $\theta_0, \ldots, \theta_n$를 찾아야 함.
    - 다양한 머신러닝 기법 존재
    - 기본은 경사하강법

예제: 붓꽃 데이터셋

- 꽃받침(sepal)과 꽃입(petal)과 관련된 4개의 특성 사용: 꽃받침 길이, 꽃받침 너비, 꽃잎 길이, 꽃잎 너비
- 세 개의 품종 사용: 세토사, 버시컬러, 버지니카
- 타깃:
    - 1: 버지니카 품종 맞음
    - 0: 버지니카 품종 아님
- 양성(1) 클래스인 경우:

$$\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_4 \geq 0$$

- 음성(0) 클래스인 경우:

$$\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_4 < 0$$

훈련과 비용함수

- 비용함수: 로그 손실(log loss) 함수 사용. 단, (i)는 i번째 샘플을 가리킴.

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(\hat{p}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{p}^{(i)})]$$

$$\hat{p}^{(i)} = \sigma(\theta_0 + \theta_1 x_1^{(i)} + \cdots + \theta_n x_n^{(i)})$$

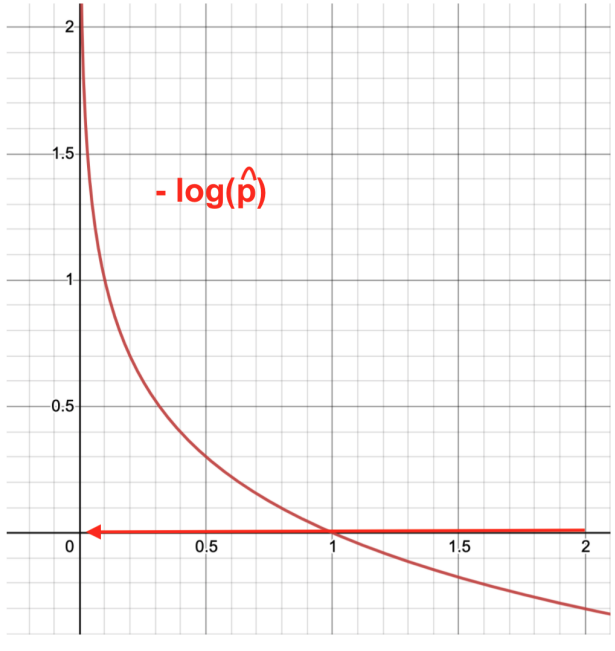
- 모델 훈련
    - 위 비용함수를 최소로 하는 $\theta_0, \ldots, \theta_n$을 찾아야 함.
    - 위 비용함수에 대해 경사 하강법 적용

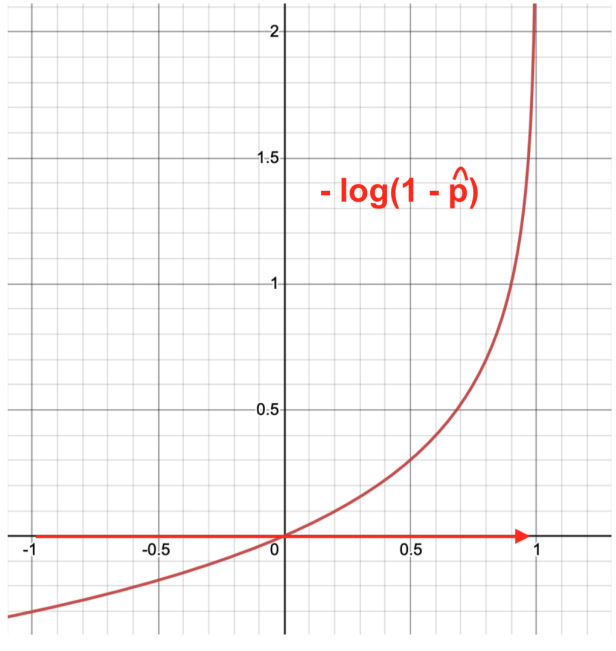
로그 손실 함수 이해

- 틀린 예측을 하면 손실값이 많이 커짐

$$-[y\ \log(\hat{p}\ ) + (1-y)\ \log(1-\hat{p}\ )]$$

y는 1인데 $\hat{p}$는 0에 가까워지는 경우	y는 0인데 $\hat{p}$는 1에 가까워지는 경우

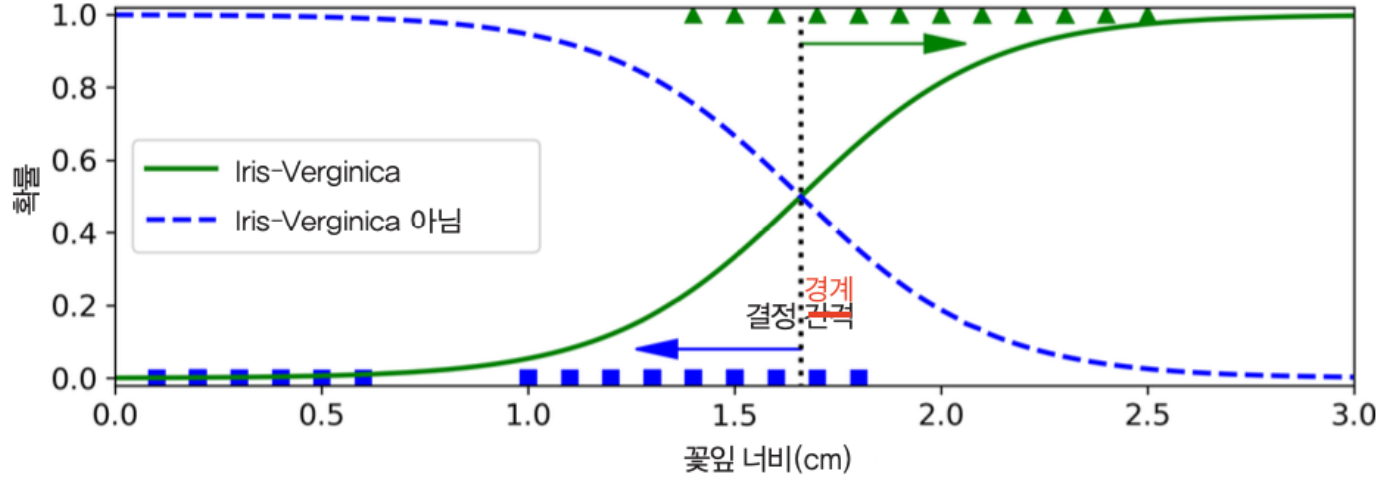




결정 경계

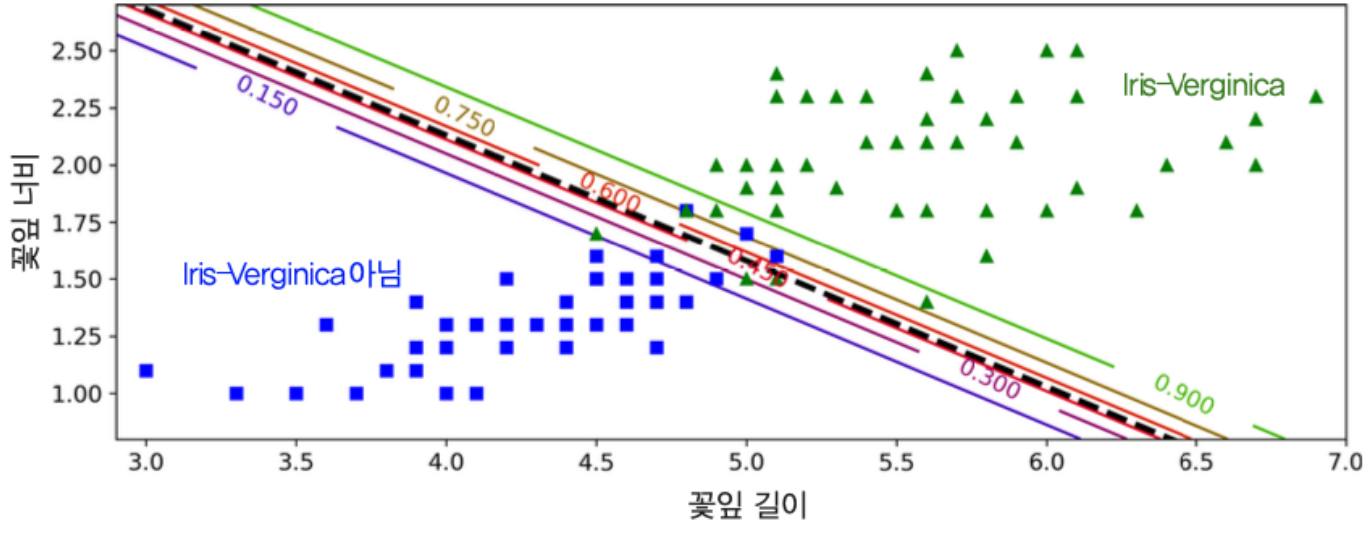
예제: 꽃잎의 너비를 기준으로 Iris-Virginica 여부 판정하기

- 결정경계: 약 1.6cm



꽃잎의 너비와 길이를 기준으로 Iris-Virginica 여부 판정하기

- 결정경계: 검정 점선



로지스틱 회귀 규제하기

- 하이퍼파라미터 `penalty`와 `C` 이용
- `penalty`
    - `l1`, `l2`, `elasticnet` 세 개중에 하나 사용.
    - 기본은 `l2`, 즉, ℓ_2 규제를 사용하는 릿지 규제.
    - `elasticnet`을 선택한 경우 `l1_ration` 옵션 값을 함께 지정.
- `C`
    - 릿지 또는 라쏘 규제 정도를 지정하는 α의 역수에 해당.
    - 따라서 0에 가까울 수록 강한 규제 의미.

소프트맥스(softmax) 회귀

- 로지스틱 회귀 모델을 일반화하여 다중 클래스 분류를 지원하도록 한 회귀 모델
- **다항 로지스틱 회귀** 라고도 불림
- 주의사항: 소프트맥스 회귀는 다중 출력 분류 지원 못함. 예를 들어, 하나의 사진에서 여러 사람의 얼굴 인식 불가능.

소프트맥스 회귀 학습 아이디어

- 샘플 $\mathbf{x}$가 주어졌을 때 각각의 분류 클래스 k 에 대한 점수 $s_k(\mathbf{x})$ 계산. 즉, `k*(n+1)` 개의 파라미터를 학습시켜야 함.

$$s_k(\mathbf{x}) = \theta_0^{(k)} + \theta_1^{(k)} x_1 + \cdots + \theta_n^{(k)} x_n$$

- **소프트맥스 함수**를 이용하여 각 클래스 k에 속할 확률 $\hat{p}_k$ 계산

$$\hat{p}_k = \frac{\exp(s_k(\mathbf{x}))}{\sum_{j=1}^{K} \exp(s_j(\mathbf{x}))}$$

- 추정 확률이 가장 높은 클래스 선택

$$\hat{y} = \text{argmax}_k s_k(\mathbf{x})$$

소프트맥스 회귀 비용함수

- 각 분류 클래스 k에 대한 적절한 가중치 벡터 θ_k를 학습해 나가야 함.
- 비용함수: 크로스 엔트로피 비용 함수 사용

$$J(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)}\ \log(\hat{p}_k^{(i)})$$

- 위 비용함수에 대해 경사 하강법 적용

- $K = 2$이면 로지스틱 회귀의 로그 손실 함수와 정확하게 일치.
- 주어진 샘플의 타깃 클래스를 제대로 예측할 경우 높은 확률값 계산
- 크로스 엔트로피 개념은 정보 이론에서 유래함. 자세한 설명은 생략.

다중 클래스 분류 예제

- 사이킷런의 `LogisticRegression` 예측기 활용
    - `multi_class=multinomial`로 지정
    - `solver=lbfgs` : 다중 클래스 분류 사용할 때 반드시 지정
- 붓꽃 꽃잎의 너비와 길이를 기준으로 품종 분류
    - 결정경계: 배경색으로 구분
    - 곡선: Iris-Versicolor 클래스에 속할 확률