

# 2022학년도 1학기 컴퓨터언어학

## 제3강 로지스틱 회귀분석 (1)

박수지

서울대학교 인문대학 언어학과

2022년 3월 14일 월요일

## 오늘 배울 것

- 1 확률적 분류기
- 2 로지스틱 분류 함수
- 3 교차엔트로피 오차 함수

## 필요한 것 (가르쳐 줌)

- 벡터의 내적과 행렬의 곱셈
- 로그함수와 지수함수의 특성
- 편미분 및 편도함수

# 잠깐!

## 이메일 잘 쓰는 법

제목에 [컴퓨터언어학(001)] 혹은 [컴퓨터언어학(002)]로 과목명과 분반을 명시한다.

# 형식화에 익숙해지기

## 예측 문제

입력  $x$ 의 값이 주어졌을 때 출력될  $y$ 의 값을 예측하는 함수  $f$

$$f(x) = y$$

## 분류 문제

예측 문제에서 입력이 관측(observation)이고 출력이 부류(class)인 경우

자연어처리의 과제들은 대부분 분류 문제를 푸는 것이다!

## 분류 예시: 단어 예측 $y \in \text{Vocabulary}$

- $f(\text{“행렬을 배운 적이”}) = \text{“없다”}$
- $f(\text{“집에 갈 수 ”}) = \text{“있다”}$

## 분류 예시: 감정분석 $y \in \{+, -\}$

- $f(\text{“컴퓨터언어학 재밌어요!!!”}) = +$
- $f(\text{“컴퓨터언어학 힘들어요...”}) = -$

## 분류 예시: 자연어추론 $y \in \{\text{entailment, contradiction, neutral}\}$

- $f(\text{“접시를 깨뜨렸다. 접시가 부서졌다.”}) = \text{entailment}$
- $f(\text{“컵을 씻었다. 컵이 더러워졌다.”}) = \text{contradiction}$

그런데 분류 함수  $f$ 를 어떻게 얻을 수 있는가?

## 규칙기반 분류기

주어진 입력  $x$ 가 특정 규칙(들)을 만족하면 특정 범주  $y$ 로 분류한다.

### 규칙기반 분류기 예시: 스팸메일 감지

- 규칙1: 제목에 “[광고]”가 포함되어 있으면 스팸으로 분류한다.

## 확률적 분류기

주어진 입력  $x$ 가 범주  $y$ 로 분류될 조건부확률  $P(y|x)$ 를 계산한다.

### 통계적 분류기 예시: 감정분석

- $P(+| \text{“컴퓨터언어학 재밌어요!!!”}) = 0.89 \Rightarrow +(\text{긍정})$ 으로 분류
- $P(+| \text{“컴퓨터언어학 힘들어요...”}) = 0.24 \Rightarrow -(\text{부정})$ 으로 분류

## 기계학습 분류기의 네 가지 요소

1 특성 표현: 관측된 데이터를 벡터로 표현한다.

예 “컴퓨터언어학 재밌어요!!!”  $\mapsto \vec{x} = [3.1, -0.7, 2.5]$

2 분류 함수: 관측된 데이터가 속할 부류  $y$ 의 추정치  $\hat{y}$ 를 계산한다.

3 목적 함수: 훈련 집합에서 오차를 최소화한다.

- 훈련 집합: 관측(의 특성 표현)  $\vec{x}^{(i)}$ 와 정답(실제 부류)  $y^{(i)}$ 로 이루어진 집합

4 최적화 알고리즘: 목적 함수를 최적화한다.

## 로지스틱 회귀분석이란?

분류 함수가 로지스틱 함수인 분류기



# 회귀분석 벵락치기

선형 회귀분석 (분류 문제 아님!)

$$y = \vec{w} \cdot \vec{x} + b$$

예시 (Pagel et al. 2007)

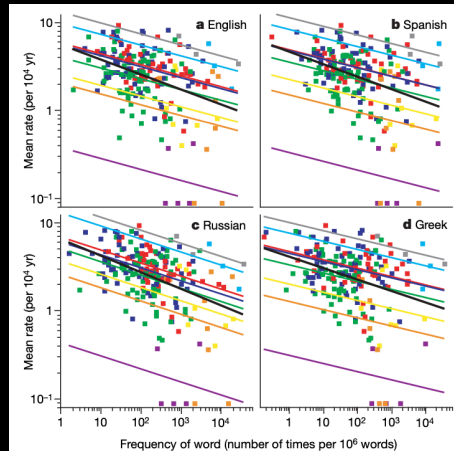
<https://www.doi.org/10.1038/nature06176>

자주 쓰이는 단어일수록 형태가 보존된다.

예시(en-fr)    ■ 저빈도 tail-queue (교체 O)  
                   ■ 고빈도 two-deux (교체 X)

x 단어 빈도의 로그 값

y 어휘 교체가 일어난 비율의 로그 값



**Figure 3 | Frequency of meaning-use plotted against estimated rate of lexical evolution for 200 basic meanings in four Indo-European languages.**

선형 회귀분석 (회귀 문제  $y \in \mathbb{R}$ )

$$\begin{aligned}\hat{y} &= \vec{w} \cdot \vec{x} + b \\ &= w_1x_1 + w_2x_2 + \cdots + w_fx_f + b\end{aligned}$$

관측의 특성값  $\vec{x} = [x_1, x_2, \cdots x_f]$

가중치  $\vec{w} = [w_1, w_2, \cdots w_f]$

편향(절편)  $b$

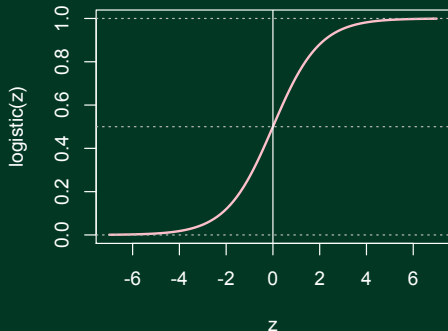
## 로지스틱 회귀분석 (분류 문제임!)

$$\begin{aligned}z &= \vec{w} \cdot \vec{x} + b \\ \hat{y} &= P(y = 1 | \vec{x}) = \sigma(z) = \sigma(\vec{w} \cdot \vec{x} + b)\end{aligned}$$

실제 부류  $y \in \{0, 1\}$

## 로지스틱 함수(시그모이드)

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$



# exp 함수를 처음 보는 사람들을 위해

$e^a = b$ 가 성립할 때(자연상수  $e \approx 2.71828$ )

## 로그함수 $a = \log(b)$

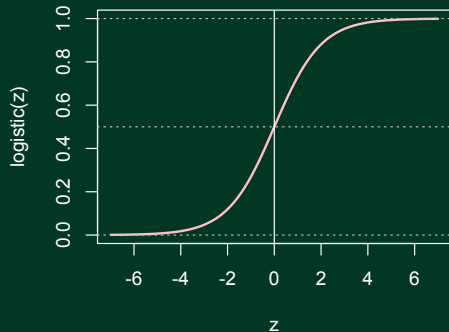
- $\log(xy) = \log(x) + \log(y)$
- $\log(\exp(x)) = x$
- $\log(1) = 0$
- 양의 실수  $x > 0$ 에 대해서만  $\log(x)$  정의 가능
- $\lim_{x \rightarrow \infty} \log(x) = +\infty$
- $\lim_{x \rightarrow 0} \log(x) = -\infty$

## 지수함수 $b = \exp(a)$

- $\exp(x + y) = \exp(x) \exp(y)$
- $\exp(\log(x)) = x$
- $\exp(0) = 1$
- 모든 실수  $x \in \mathbb{R}$ 에 대하여  $\exp(x) > 0$  성립
- $\lim_{x \rightarrow +\infty} \exp(x) = +\infty$
- $\lim_{x \rightarrow -\infty} \exp(x) = 0$

## 로지스틱 함수

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$



- $\sigma(0) = \frac{1}{1 + \exp(0)} = \frac{1}{1 + 1} = 0.5$
- $\lim_{z \rightarrow +\infty} \sigma(z) = 1$
- $\lim_{z \rightarrow -\infty} \sigma(z) = 0$
- 모든  $z \in \mathbb{R}$ 에 대해  $0 < \sigma(z) < 1$  성립

로지스틱 함수를 사용하는 목적

확률의 조건을 충족하는 값을 얻기 위해

## 로지스틱 회귀분석

문서의 표현: 특성값의 벡터  $\vec{x} = [x_1, x_2, \dots, x_f] \in \mathbb{R}^f$

긍정적일 확률  $P(y = 1|\vec{x}) = \sigma(\vec{w} \cdot \vec{x} + b) = \frac{1}{1 + \exp(-(\vec{w} \cdot \vec{x} + b))}$

부정적일 확률  $P(y = 0|\vec{x}) = 1 - P(y = 1|\vec{x}) = \dots = \frac{\exp(-(\vec{w} \cdot \vec{x} + b))}{1 + \exp(-(\vec{w} \cdot \vec{x} + b))}$

주의  $y \in \{0, 1\}$  이므로  $P(y = 1|\vec{x}) + P(y = 0|\vec{x}) = 1$  이 성립해야 한다(이항 분류).

각 부류의 확률을 구하고 나면 더 높은 확률을 가지는 부류를 예측값으로 선택하면 된다.

## 결정 경계

$$\text{decision}(\vec{x}) = \begin{cases} 1 & \text{if } P(y = 1|\vec{x}) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

## 로지스틱 회귀분석 분류 예시: 영화평 감정 분류

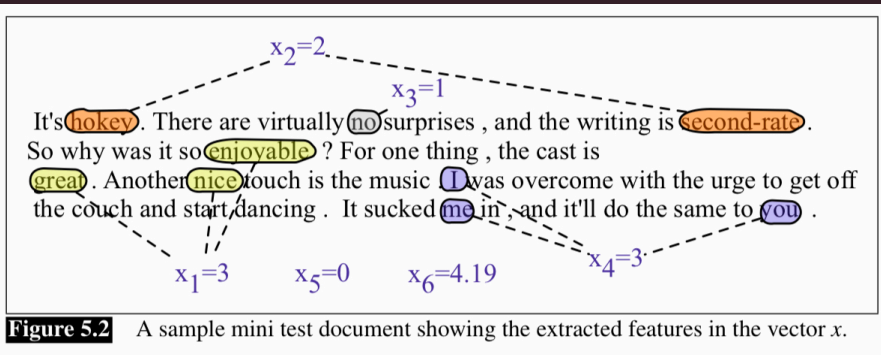
## 특성 설정

Var	Definition	Value in Fig. 5.2
$x_1$	count(positive lexicon words $\in$ doc)	3
$x_2$	count(negative lexicon words $\in$ doc)	2
$x_3$	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	1
$x_4$	count(1st and 2nd pronouns $\in$ doc)	3
$x_5$	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	0
$x_6$	log(word count of doc)	$\ln(66) = 4.19$

# 로지스틱 회귀분석 분류 예시: 영화평 감정 분류

## 특성값 계산

$$[x_1, x_2, x_3, x_4, x_5, x_6] = [3, 2, 1, 3, 0, 4.19]$$



**Figure 5.2** A sample mini test document showing the extracted features in the vector  $x$ .



# 로지스틱 회귀분석 분류 예시: 영화평 감정 분류

## 확률 계산

설정:  $[w_1, w_2, w_3, w_4, w_5, w_6] = [2.5, -5.0, -1.2, 0.5, 2.0, 0.7]$ ,  $b = 0.1$

$$\begin{aligned} p(+|x) = P(y = 1|x) &= \sigma(\mathbf{w} \cdot \mathbf{x} + b) \\ &= \sigma([2.5, -5.0, -1.2, 0.5, 2.0, 0.7] \cdot [3, 2, 1, 3, 0, 4.19] + 0.1) \\ &= \sigma(.833) \\ &= 0.70 \end{aligned} \tag{5.7}$$

$$\begin{aligned} p(-|x) = P(y = 0|x) &= 1 - \sigma(\mathbf{w} \cdot \mathbf{x} + b) \\ &= 0.30 \end{aligned}$$

# 다량의 데이터 처리하기

실제의 시험 집합은 여러 개의 데이터로 이루어져 있다.

$$\hat{y}^{(i)} = \sigma \left( \vec{w} \cdot \vec{x}^{(i)} + b \right) = \sigma \left( \left[ x_1^{(i)}, x_2^{(i)}, \dots, x_f^{(i)} \right] \cdot [w_1, w_2, \dots, w_f] + b \right)$$

시험 집합에 관측이  $m$ 개, 특성값이  $f$ 가지 있을 때: 입력값을  $(m \times f)$  행렬로 표현할 수 있다.

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix} = \sigma \left( \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_f^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_f^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \dots & x_f^{(m)} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_f \end{bmatrix} + \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix} \right)$$

로지스틱 회귀분석 모델을 학습시킨다는 것

모델 매개변수(가중치 벡터  $\vec{w}$ 와 편향  $b$ )를 학습시키는 것

모델 매개변수를 어떻게 학습하는가?

- 분류기가 예측한  $\hat{y}$ 와 정답  $y$  사이의 거리를 표현하는 손실(비용) 함수  $L$ 을 정의한다.
  - 평균제곱오차(Mean Squared Error) — 선형 회귀분석
  - 교차엔트로피오차(Cross Entropy Error) — 로지스틱 회귀분석
- 손실 함수의 값을 최소화하는 알고리즘을 실행한다.
  - (확률적) 경사하강법((Stochastic) Gradient Descent)

## 손실 함수

$\hat{y}$  예측 결과

$y$  실제 정답

$\hat{y}$ 와  $y$ 가 얼마나  
떨어져 있는가?

## 예시: 선형 회귀분석

■ 예측값 추정  $\hat{y} = \vec{w} \cdot \vec{x} + b$

■ 손실 함수  $L_{\text{MSE}}(\hat{y}, y) = \frac{1}{2} (\hat{y} - y)^2$

$$\begin{aligned}\text{Cost}(\vec{w}, b) &= \frac{1}{m} \sum_{i=1}^m L_{\text{MSE}}(\hat{y}^{(i)}, y^{(i)}) \\ &= \frac{1}{m} \sum_{i=1}^m \frac{1}{2} (\vec{w} \cdot \vec{x}^{(i)} + b - y^{(i)})^2\end{aligned}$$

$\vec{w}$ ,  $b$ 에 대해 미분하여 최솟값을 구할 수 있다.

## 로지스틱 회귀분석의 문제

평균제곱오차 손실 함수를 사용하면 최적화하기 어렵다.

## 교차엔트로피 손실 함수의 목표

조건부최대가능도 추정법(Conditional maximum likelihood estimation)  
훈련 집합의 관측  $\vec{x}$ 에 대한 정답  $y$ 의 확률을 최대로 만드는 매개변수  $\vec{w}$ 와  $b$ 를 선택한다.

$$\begin{aligned} p(y|\vec{x}) \text{의 값을 최대로 만든다} &= \log p(y|\vec{x}) \text{의 값을 최대로 만든다} \\ &= -\log p(y|\vec{x}) \text{의 값을 최소로 만든다} \end{aligned}$$

똑같이  $y = 1$ 을 맞히더라도  $\hat{y} = 0.9$ 의 확률로 맞히는 것이  $\hat{y} = 0.6$ 보다 좋다!

## 로지스틱 회귀분석 분류기의 확률 표현 및 로그 값

$$\begin{aligned} p(y|x) &= \begin{cases} \hat{y}, & y = 1 \\ 1 - \hat{y}, & y = 0 \end{cases} \\ &= \hat{y}^y (1 - \hat{y})^{1-y} \quad (\hat{y} \text{는 } y \text{가 1일 확률}) \end{aligned}$$

$$\begin{aligned} \log p(y|x) &= \log [\hat{y}^y (1 - \hat{y})^{1-y}] \\ &= y \log \hat{y} + (1 - y) \log (1 - \hat{y}) \end{aligned}$$

## 교차엔트로피 손실 함수

$$\begin{aligned}L_{CE}(\hat{y}, y) &= -\log p(y|x) \\&= -[y \log \hat{y} + (1 - y) \log (1 - \hat{y})] \\&= -[y \log \sigma(\vec{w} \cdot \vec{x} + b) + (1 - y) \log (1 - \sigma(\vec{w} \cdot \vec{x} + b))]\end{aligned}$$

## 모형의 “비용”

$$\begin{aligned}\text{Cost}(\vec{w}, b) &= \frac{1}{m} \sum_{i=1}^m L_{CE}(\hat{y}^{(i)}, y^{(i)}) \\&= -\frac{1}{m} \sum_{i=1}^m \left[ y^{(i)} \log \sigma(\vec{w} \cdot \vec{x}^{(i)} + b) + (1 - y^{(i)}) \log (1 - \sigma(\vec{w} \cdot \vec{x}^{(i)} + b)) \right]\end{aligned}$$

# 개괄

통계적 분류기 일반  $P(\text{class}|\text{observation})$

로지스틱 회귀분석  $P(Y = 1|X = \vec{x}) = \sigma(\vec{w} \cdot \vec{x} + b) = \frac{1}{1 + \exp(-(\vec{w} \cdot \vec{x} + b))}$

$\hat{y}$  관측의 특성값  $x$ 가 주어졌을 때 분류기가  $y$ 를 1로 예측할 확률 ( $0 < \hat{y} < 1$ )

$y$   $x$ 에 해당하는 관측이 실제로 속하는 부류(정답) ( $y \in \{0, 1\}$ )

## 목표

$y = 1$ 일 때  $\hat{y}$ 를 1에 가깝게,  $y = 0$ 일 때  $\hat{y}$ 를 0에 가깝게 만들기

⇒ 교차엔트로피 함수  $L_{CE} = -[y \log \hat{y} + (1 - y) \log (1 - \hat{y})]$ 의 값을 최소로 만들기

⇒ 방정식  $\frac{\partial}{\partial w_1} L_{CE} = 0, \dots, \frac{\partial}{\partial w_f} L_{CE} = 0, \frac{\partial}{\partial b} L_{CE} = 0$ 을 만족하는  $\vec{w}, b$ 의 값을 구하기



## 문제

$\frac{\partial}{\partial w_j} L_{CE} = 0$ 을 만족하는  $w_j$ 의 값을 한번에 계산해 낼 수 없다.

## 해결

확률적 경사 하강법(Stochastic Gradient Descent) 알고리즘으로 해를 찾는다.

그런데  $\frac{\partial}{\partial w_j} L_{CE}$ 라는 기호가 무슨 뜻인가?

# 편도함수 벵락치기

## 정의

다변수함수를 하나의 변수에 대하여 (나머지 변수를 상수로 놓고) 미분하여 얻은 도함수

## 예시

$f(x_1, x_2, x_3) = x_1x_2 + x_3^2$ 일 때  $x_j$ 에 대한 편도함수는 아래와 같다.

- $\frac{\partial}{\partial x_1} f(x) = x_2$
- $\frac{\partial}{\partial x_2} f(x) = x_1$
- $\frac{\partial}{\partial x_3} f(x) = 2x_3$

## 통계적 분류기로서 로지스틱 회귀분석의 작동 과정

- 1 훈련 집합의 각 문서를 특성값들의 벡터  $\vec{x}$ 로 나타낸다.
- 2 분류기가  $\vec{x}$ 를 1로 분류할 확률  $\hat{y} = P(y = 1|\vec{x})$ 를  $\sigma(\vec{w} \cdot \vec{x} + b)$ 로 나타낸다.
- 3 확률 추정값  $\hat{y}$ 과 실제 정답  $y$  사이의 “거리”를 교차엔트로피 손실 함수로 정의한다.
- 4 교차엔트로피 손실 함수의 값을 최소로 만들기 위해 편도함수의 값이 0이 될 때의 모형 매개변수  $\vec{w}$ ,  $b$ 의 값을 계산한다.

## 남은 문제

- 1 텍스트를 어떻게 수치화된 벡터  $\vec{x}$ 로 나타내는가? — Feature Engineering
- 2 부류가 0, 1 이외에 세 개 이상 존재하는 경우 확률을 어떻게 추정하는가? — Multinomial logistic regression
- 3 교차엔트로피 손실 함수의 편도함수가 0이 되는 지점을 어떻게 찾는가? — Gradient Descent