2021학년도 2학기 언어와 컴퓨터

제20강 로지스틱 회귀분석 (1)

박수지

서울대학교 인문대학 언어학과

2021년 11월 22일 월요일

오늘의 목표

1. 회귀분석이 무엇을 하고자 하는지를 설명할 수 있다.
2. 단순 베이즈 분류기를 사용한 분류와 로지스틱 회귀분석을 사용한 분류의 차이를 설명할 수 있다.

선형 회귀분석

$$y = a \cdot x + b$$

예시 (Pagel et al. 2007)

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

자주 쓰이는 단어일수록 형태가 보존된다.

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

x 단어 빈도의 로그 값

y 인도유럽어에서 어휘 교체가 일어난 비율의 로그 값

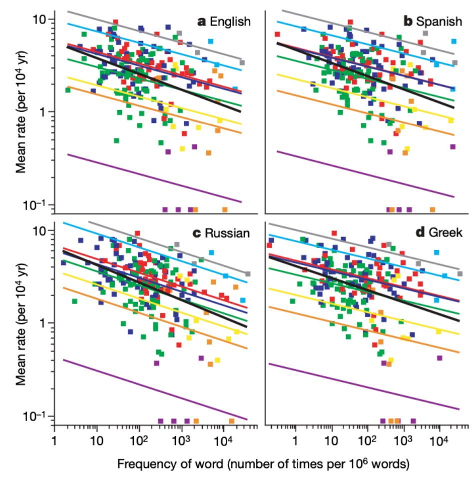


Figure 3 | Frequency of meaning-use plotted against estimated rate of lexical evolution for 200 basic meanings in four Indo-European languages.

선형 회귀분석

$$y = a \cdot x + b$$
$$= a_1 x_1 + a_2 x_2 + \cdots + a_d x_d + b$$

로지스틱 회귀분석

$$z = a \cdot x + b$$
$$P(y = 1 | x) = \sigma(z) = \sigma(a \cdot x + b)$$

반응 변수 y

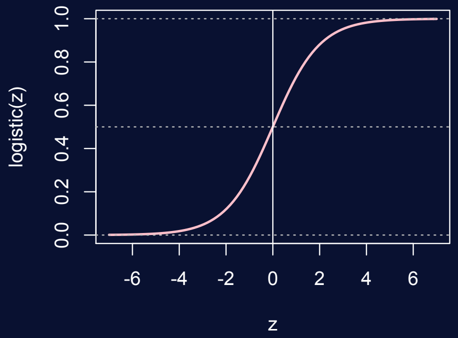
설명 변수 $x = [x_1, x_2, \cdots x_d]$

가중치 $a = [a_1, a_2, \cdots a_d]$

절편 b

로지스틱 함수

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



관측이 n 개, 설명 변수가 d 개 있을 때: 특성값을 $(n \times d)$ 행렬로 표현

$$\begin{bmatrix} z^{(1)} \\ z^{(2)} \\ \vdots \\ z^{(n)} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \cdots & x_d^{(n)} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} + \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix}$$

$x_j^{(i)}$ i 번째 데이터(문서)의 j 번째 특성값

로지스틱 회귀분석

문서 d의 표현: 특성값의 벡터 $[x_1, x_2, \cdots, x_d] \in \mathbb{R}^d$

긍정적일 확률 $P(y=1|x) = \dfrac{1}{1+\exp(-(a{\cdot}x+b))} = \dfrac{1}{1+\exp\left(-\left(\sum_{j=1}^{d} a_j x_j + b\right)\right)}$

부정적일 확률 $P(y=0|x) = 1 - \dfrac{1}{1+\exp(-(a{\cdot}x+b))}$

필요한 것

특성값 x_j의 가중치 a_j $(j = 1, 2, \cdots, d)$ 및 절편 b

단순 베이즈 분류기

문서 d의 표현: 단어의 연쇄 $w_1 w_2 \cdots w_l$

긍정적일 확률 $$P(+|d) = \frac{P(w_1|+) \times \cdots \times P(w_l|+) \times P(+)}{P(d)}$$

부정적일 확률 $$P(-|d) = \frac{P(w_1|-) \times \cdots \times P(w_l|-) \times P(-)}{P(d)}$$

필요한 것

감정 범주 c가 주어졌을 때 단어 w_v가 출현할 확률 $P(w_v|c)$ $(v = 1, 2, \cdots, |V|)$

분류기의 두 가지 유형

생성 가능도 $P(d|c)$를 먼저 계산해서 활용한다.

- $P(w_v|+)$를 통해 긍정적인 문서를 **생성**할 수 있다.

식별 확률 $P(c|d)$를 직접 계산한다.

- 긍정적인 문서가 어떻게 생겼는지 알 수 없으나 부정적인 문서와 **구별**할 수는 있다.

기계학습 분류기의 네 가지 요소

1. 특성 표현: 관측된 데이터를 벡터로 표현한다.
2. 분류 함수: 관측된 데이터가 속할 부류를 추정한다.
3. 목적 함수: 훈련 집합에서 오차를 최소화한다.
4. 최적화 알고리듬: 목적 함수를 최적화한다.

예: 단순 베이즈 분류기의 특성 표현

$x_j^{(i)}$ V의 j번째 단어가 i번째 문서에 출현한 횟수

단어 출현 횟수 이외의 정보도 특성이 될 수 있다.

로지스틱 회귀분석 분류 예시: 영화평 감정 분류

특성 설정

Var	Definition	Value in Fig. 5.2
x_1	count(positive lexicon) $\in$ doc)	3
x_2	count(negative lexicon) $\in$ 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)	$\ln(64) = 4.15$

로지스틱 회귀분석 분류 예시: 영화평 감정 분류

특성값 계산

$$[x_1, x_2, x_3, x_4, x_5, x_6] = [3, 2, 1, 3, 0, 4.15]$$

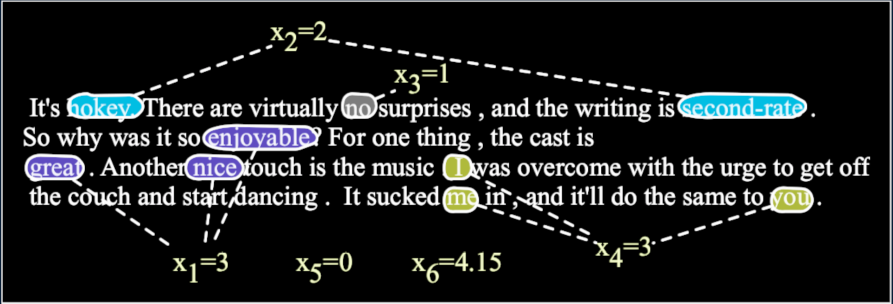


Figure 5.2 A sample mini test document showing the extracted features in the vector x.

로지스틱 회귀분석 분류 예시: 영화평 감정 분류

확률 계산

설정: $[a_1, a_2, a_3, a_4, a_5, a_6] = [2.5, -5.0, -1.2, 0.5, 2.0, 0.7]$, $b = 0.1$

$$
\begin{aligned}
p(+|x) = P(Y=1|x) &= \sigma(w \cdot x + b) \\
&= \sigma([2.5, -5.0, -1.2, 0.5, 2.0, 0.7] \cdot [3, 2, 1, 3, 0, 4.15] + 0.1) \\
&= \sigma(1.805) \\
&= 0.86 \\
p(-|x) = P(Y=0|x) &= 1 - \sigma(w \cdot x + b) \\
&= 0.14
\end{aligned}
$$

문제 1: 특성을 어떻게 설정하는가?

- 훈련 집합의 자료를 보고 언어학적 직관과 지식으로 판단한다.
- 개발 집합에서 오류를 분석하여 특성 설정이 잘 되었는지 확인한다.

좋은 특성을 찾으려면…

- 데이터를 잘 관찰하고 이해하자.
- 기존 연구를 열심히 찾자.

문제 2: 모형 매개변수(가중치와 절편)를 어떻게 학습하는가?

- 정답과 예측 사이의 거리를 표현하는 손실 함수를 정의한다.
  - 평균제곱오차(Mean Squared Error) — 선형 회귀분석
  - 교차엔트로피오차(Cross Entropy Error) — 로지스틱 회귀분석
- 손실 함수의 값을 최소화하는 알고리듬을 실행한다.
  - (확률적) 경사하강법((Stochastic) Gradient Descent)

교차엔트로피 손실 함수

손실 함수

$\hat{y}$ 예측 결과
y 실제 정답

$\hat{y}$와 y가 얼마나 떨어져 있는가?

예시: 선형 회귀분석

- 반응 변수 추정 $\hat{y} = a \cdot x + b$
- 손실 함수 $L_{MSE}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$

$$\begin{aligned} Cost(a, b) &= \frac{1}{n}\sum_{i=1}^{n} L_{MSE}\left(\hat{y}^{(i)}, y^{(i)}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(a \cdot x^{(i)} + b - y^{(i)}\right)^2 \end{aligned}$$

a, b에 대해 미분하여 최솟값을 구할 수 있다.

교차엔트로피 손실 함수

로지스틱 회귀분석의 문제

평균제곱오차 손실 함수를 사용하면 최적화하기 어렵다.

교차엔트로피 손실 함수의 목표

$\log p(y|x)$ 의 값을 최대로 한다 = $-\log p(y|x)$ 의 값을 최소로 한다.

교차엔트로피 손실 함수

$$
\begin{aligned}
p(y|x) &= \begin{cases} \hat{y}, & y = 1 \\ 1 - \hat{y}, & y = 0 \end{cases} \\
&= \hat{y}^{y} (1 - \hat{y})^{1-y} \quad (\hat{y}\text{는 } y\text{가 } 1\text{일 확률}) \\
\log p(y|x) &= \log \left[\hat{y}^{y} (1 - \hat{y})^{1-y} \right] \\
&= y \log \hat{y} + (1 - y) \log (1 - \hat{y}) \\
L_{CE}(\hat{y}, y) &= -\log p(y|x) = -\left[y \log \hat{y} + (1 - y) \log (1 - \hat{y}) \right]
\end{aligned}
$$

교차엔트로피 손실 함수

$$
\begin{aligned}
L_{CE}(\hat{y}, y) &= -\left[y \log \hat{y} + (1-y)\log(1-\hat{y})\right] \\
&= -\left[y \log \sigma(a \cdot x + b) + (1-y)\log(1-\sigma(a \cdot x + b))\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cost}(a, b) &= \frac{1}{n}\sum_{i=1}^{n} L_{CE}\left(\hat{y}^{(i)}, y^{(i)}\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\Big[y^{(i)} \log \sigma(a \cdot x^{(i)} + b) \\
&\qquad + (1-y^{(i)})\log\left(1-\sigma(a \cdot x^{(i)} + b)\right)\Big]
\end{aligned}
$$

경사 하강법

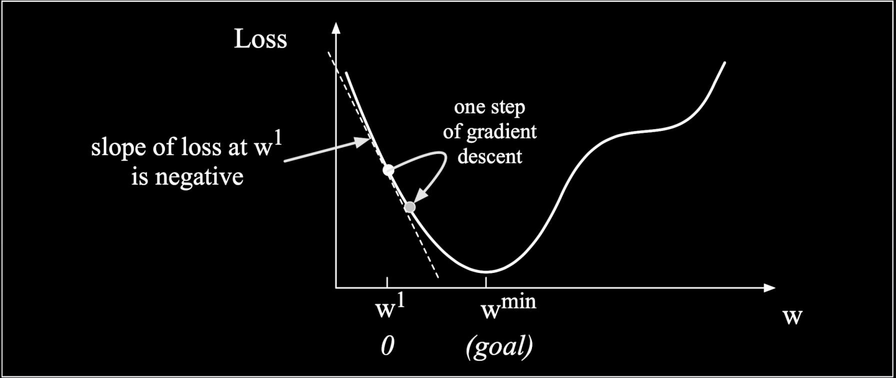


Figure 5.3 The first step in iteratively finding the minimum of this loss function, by moving w in the reverse direction from the slope of the function. Since the slope is negative, we need to move w in a positive direction, to the right. Here superscripts are used for learning steps, so w^1 means the initial value of w (which is 0), w^2 at the second step, and so on.