2021학년도 2학기 언어와 컴퓨터

제21강 로지스틱 회귀분석 (2)

박수지

서울대학교 인문대학 언어학과

2021년 11월 24일 수요일

오늘의 목표

1. 확률적 경사 하강법을 사용하여 로지스틱 회귀분석에서 교차엔트로피 함수를 최소로 하는 매개변수 값들을 찾을 수 있다.

준비

x_j, y, a_j, b의 의미

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

$$\begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(n)} \end{bmatrix} = \sigma\left(\begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_d^{(m)} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} + \begin{bmatrix} b \\ b \\ \vdots \\ b \end{bmatrix} \right)$$

$x_j^{(i)}$ i 번째 데이터(문서)의 j 번째 특성값

a_j j 번째 특성값에 대한 가중치

개괄

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

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

$\hat{y}$ 데이터의 설명변수(특성값) x가 주어졌을 때 반응변수(범주) y가 1일 확률 $(0 < \hat{y} < 1)$

y x에 해당하는 데이터가 실제로 속하는 정답 $(y \in \{0, 1\})$

목표

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

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

$\Rightarrow$ 방정식 $\frac{\partial}{\partial a_1} L_{CE} = 0, \ldots, \frac{\partial}{\partial a_d} L_{CE} = 0, \frac{\partial}{\partial b} L_{CE} = 0$을 만족하는 a, b의 값을 구하기

문제

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

해결

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

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

편도함수

정의

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

예시

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

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

경사 하강법

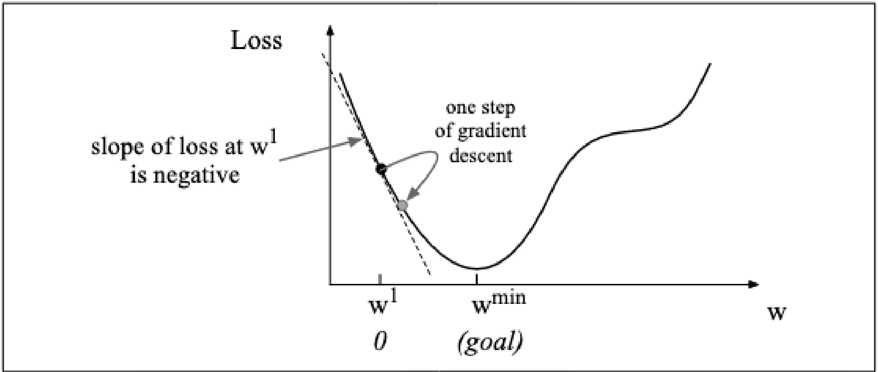

Figure 5.3 The first step in iteratively finding the minimum of this loss function, by moving w in the reverse direction from the slope of the function. Since the slope is negative, we need to move w in a positive direction, to the right. Here superscripts are used for learning steps, so w^1 means the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

경사 하강법

접선의 기울기(편도함수의 값)가 감소하는 방향으로 η만큼 움직인다.

$$\mathrm{a}_j^{(t+1)} \leftarrow \mathrm{a}_j^{(t)} - \eta \frac{\partial}{\partial \mathrm{a}_j} L_{CE}$$

$\mathrm{a}_j^{(t)}$ 현재 단계의 a_j 값

$\mathrm{a}_j^{(t+1)}$ 다음 단계의 a_j 값

η 움직이는 정도. 학습률(learning rate).

L_{CE}의 편도함수

- $\frac{\partial}{\partial a_j} L_{CE}(\hat{y}, y) = (\hat{y} - y)x_j$
- $\frac{\partial}{\partial b} L_{CE}(\hat{y}, y) = (\hat{y} - y)$

증명: SLP3 Ch.5 부록

결론

로지스틱 회귀분석 모형의 매개변수를 $\theta = (a_1, a_2, \cdots, a_d, b)$ 라고 할 때

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \nabla_\theta L_{CE}$$

$$\begin{bmatrix} a_1^{(t+1)} \\ a_2^{(t+1)} \\ \vdots \\ a_d^{(t+1)} \\ b^{(t+1)} \end{bmatrix} \leftarrow \begin{bmatrix} a_1^{(t)} \\ a_2^{(t)} \\ \vdots \\ a_d^{(t)} \\ b^{(t)} \end{bmatrix} - \eta \begin{bmatrix} (\hat{y} - y)\, x_1 \\ (\hat{y} - y)\, x_2 \\ \vdots \\ (\hat{y} - y)\, x_d \\ (\hat{y} - y) \end{bmatrix}$$

단, $\hat{y} = \sigma\left(a^{(t)} \cdot x + b^{(t)}\right) = \dfrac{1}{1 - \exp\left(-\left(a^{(t)} \cdot x + b^{(t)}\right)\right)} = \dfrac{1}{1 - \exp\left(-\left(a_1^{(t)} x_1 + a_2^{(t)} x_2 + \cdots + a_d^{(t)} x_d + b^{(t)}\right)\right)}$

확률적 경사 하강법 (SGD)

실습 코드: `https://colab.research.google.com/drive/19PcTY3eEm-Yy0NPiCIir-0M-6Uvu4K4l?usp=sharing`

function STOCHASTIC GRADIENT DESCENT($L()$, $f()$, x, y) **returns** θ
    # where: L is the loss function
    #   f is a function parameterized by θ
    #   x is the set of training inputs $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$
    #   y is the set of training outputs (labels) $y^{(1)}, y^{(2)}, \ldots, y^{(m)}$

$\theta \leftarrow 0$
repeat til done   # see caption
  For each training tuple $(x^{(i)}, y^{(i)})$ (in random order)
    1. Optional (for reporting):   # How are we doing on this tuple?
      Compute $\hat{y}^{(i)} = f(x^{(i)};\theta)$ # What is our estimated output $\hat{y}$?
      Compute the loss $L(\hat{y}^{(i)}, y^{(i)})$ # How far off is $\hat{y}^{(i)}$ from the true output $y^{(i)}$?
    2. $g \leftarrow \nabla_\theta L(f(x^{(i)};\theta), y^{(i)})$ # How should we move θ to maximize loss?
    3. $\theta \leftarrow \theta - \eta\, g$ # Go the other way instead
return θ

Figure 5.5 The stochastic gradient descent algorithm. Step 1 (computing the loss) is used to report how well we are doing on the current tuple. The algorithm can terminate when it converges (or when the gradient norm $< \epsilon$), or when progress halts (for example when the loss starts going up on a held-out set).