2021학년도 2학기 언어와 컴퓨터

제22강 로지스틱 회귀분석 (3)

박수지

서울대학교 인문대학 언어학과

2021년 11월 29일 월요일

오늘의 목표

1. TF-IDF의 값을 계산할 수 있다.
2. `sklearn` 라이브러리에서 로지스틱 회귀분석 모형을 훈련시킬 수 있다.

지난 시간에 한 일

아래와 같은 형태에서 a_j, b의 값을 추정하였다.

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

실제적인 문제

위와 같은 행렬을 어떻게 얻는가?

특성값의 예시

- 긍정 어휘의 개수
- 부정 어휘의 개수
- 길이의 로그 값
- …

전통적인 방법

TF-IDF(Term Frequency - Inverse Document Frequency) 점수

개념

$\text{TF-IDF}(t, d) = \text{tf}_{t,d} \times \text{idf}_t$

$\text{tf}_{t,d}$ $\log_{10}[[\text{단어 t가 문서 d에 출현한 횟수}] + 1]$

- 문서에 자주 나타날수록 중요한 단어(주제어)다.
  - 문서에 한 번도 나오지 않은 $\text{tf}_{t,d}$의 값은 0이다.
  - tf값이 높을수록 중요하다.

idf_t $\log_{10} \frac{[\text{전체 문서의 개수}]}{[\text{단어 t를 포함하는 문서의 개수}]}$

- 이 문서 저 문서에 다 나오는 단어(예: the)은 정보량이 적다.
  - 모든 문서에 나오면 idf_t 값이 0이 된다.
  - idf값이 높을수록 중요하다.

$x_i^{(j)}$ 어휘 목록의 i번째 단어와 코퍼스의 j번째 문서의 TF-IDF 값

- 실습 코드 https://colab.research.google.com/drive/1AEUWk9MyA1PI8llZ7zVJD1kRqkyqxl_G?usp=sharing
- sklearn 라이브러리

로지스틱 회귀분석

- 로지스틱 함수
- 교차엔트로피 손실 함수
- 경사하강법
- TF-IDF 벡터화

다음 시간에 할 것

SLP3e Ch.6 벡터 의미론