

2021학년도 2학기 언어와 컴퓨터

제19강 단순 베이지 분류기와 감정분석 (2)

박수지

서울대학교 인문대학 언어학과

2021년 11월 17일 수요일

오늘의 목표

- 1 분류기를 평가하는 세 가지 척도를 각각 계산할 수 있다.
- 2 단순 베이즈 분류기를 파이썬에서 구현할 수 있다.
 - a 데이터 전처리
 - b 훈련: 로그사전확률 및 로그가능도 계산
 - c 실험: 데이터에 대한 확률이 가장 높은 범주 선택

분류기를 평가하는 척도

정확도 Accuracy 전체적으로 얼마나 맞추었는가?

정밀도 Precision 특정 범주라고 예측된 것이 실제로 얼마나 맞았는가?

재현도 Recall 실제로 그 범주인 것이 얼마나 제대로 예측되었는가?

F1 정밀도와 재현도의 조화평균

주의

기준이 되는 범주에 따라 정밀도, 재현도, F1의 값이 달라진다.

Confusion matrix

	예측:양성	예측:음성
실제:양성	진양성(True Positive)	위음성(False Negative)
실제:음성	위양성(False Positive)	진음성(True Negative)

$$\text{정확도(Accuracy)} = \frac{\text{진양성(TP)} + \text{진음성(TN)}}{\text{진양성(TP)} + \text{위양성(FP)} + \text{위음성(FN)} + \text{진음성(TN)}}$$

$$\text{정밀도(Precision)} = \frac{\text{진양성(TP)}}{\text{진양성(TP)} + \text{위양성(FP)}}$$

$$\text{재현도(Recall)} = \frac{\text{진양성(TP)}}{\text{진양성(TP)} + \text{위음성(FN)}}$$

예시: 스팸 메일 분류

	예측:정상	예측:스팸
실제:정상	90	0
실제:스팸	9	1

$$\text{정확도(Accuracy)} = \frac{90 + 1}{90 + 10 + 0 + 1} = \frac{91}{100} = 91\%$$

$$\text{정상 범주에 대한 정밀도(Precision)} = \frac{90}{90 + 9} = \frac{90}{99} = 90.909090\ldots\%$$

$$\text{정상 범주에 대한 재현도(Recall)} = \frac{90}{90 + 0} = \frac{90}{90} = 100\%$$

예시: 스팸 메일 분류

	예측:스팸	예측:정상
실제:스팸	1	9
실제:정상	0	90

$$\text{정확도 (Accuracy)} = \frac{1 + 90}{1 + 0 + 9 + 90} = \frac{91}{100} = 91\%$$

$$\text{스팸 범주에 대한 정밀도 (Precision)} = \frac{1}{1 + 0} = \frac{1}{1} = 100\%$$

$$\text{스팸 범주에 대한 재현도 (Recall)} = \frac{1}{1 + 9} = \frac{1}{10} = 10\%$$

function TRAIN NAIVE BAYES(D, C) **returns** $\log P(c)$ and $\log P(w|c)$

for each class $c \in C$ # Calculate $P(c)$ terms

N_{doc} = number of documents in D

N_c = number of documents from D in class c

$logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$

$V \leftarrow$ vocabulary of D

$bigdoc[c] \leftarrow$ **append**(d) **for** d \in D **with** class c

for each word w in V # Calculate $P(w|c)$ terms

$count(w, c) \leftarrow$ # of occurrences of w in $bigdoc[c]$

$loglikelihood[w, c] \leftarrow \log \frac{count(w, c) + 1}{\sum_{w' \text{ in } V} (count(w', c) + 1)}$

return $logprior, loglikelihood, V$

function TEST NAIVE BAYES($testdoc, logprior, loglikelihood, C, V$) **returns** best c

for each class $c \in C$

$sum[c] \leftarrow logprior[c]$

for each position i in $testdoc$

$word \leftarrow testdoc[i]$

if $word \in V$

$sum[c] \leftarrow sum[c] + loglikelihood[word, c]$

return $\operatorname{argmax}_c sum[c]$

Figure 4.2 The naive Bayes algorithm, using add-1 smoothing. To use add- α smoothing instead, change the +1 to + α for loglikelihood counts in training.

통계적 분류기

$\arg \max_{c \in C} P(c|d)$ 구하기 (d 는 데이터, C 는 모든 범주의 집합)

$\max_{c \in C} P(c|d)$ $P(c|d)$ 의 최댓값

■ 예: $\max_{x \in \mathbb{R}} [-(x-1)^2 + 3] = 3$

$\arg \max_{c \in C} P(c|d)$ $P(c|d)$ 가 최댓값일 때 c 의 값

■ 예: $\arg \max_{x \in \mathbb{R}} [-(x-1)^2 + 3] = 1$

단순 베이즈 분류기

$$P(c|d) = \dots = \frac{P(c) \prod_{i=1}^K P(w_i|c)}{P(d)}$$

$$\arg \max_{c \in C} P(c|d) = \dots = \arg \max_{c \in C} \left[\log P(c) + \sum_{i=1}^N \log P(w_i|c) \right]$$

연습 문제

위에서 ... 에 들어갈 식을 도출해 보자.

단순 베이즈 분류기의 목표

$$\arg \max_{c \in C} \left[\log P(c) + \sum_{i=1}^N \log P(w_i | c) \right] \text{ 구하기}$$

계산할 것

$\log P(c)$ 범주 c 의 **로그사전확률**

- 훈련 집합에서 범주 c 에 속한 문서가 차지하는 비율

$\log P(w_i | c)$ 단어 w_i 에 대한 c 의 **로그가능도**

- 범주 c 에서 단어 w_i 가 차지하는 비율 (+평탄화)

분류기 훈련

로그사전확률: 훈련 집합에서 범주 c 에 속한 문서가 차지하는 비율

$$\log P(c) = \log \frac{N_c}{N_{\text{doc}}}$$

N_c 범주 c 에 속한 문서의 개수

N_{doc} 훈련 집합 전체 문서의 개수

코딩에 필요한 것

- 훈련 집합: 문서(단어의 목록)의 목록
- 훈련 집합의 문서를 범주별로 분류해서 저장하기

분류기 훈련

로그가능도: 범주 c 에서 단어 w 가 차지하는 비율 (+평탄화)

$$\log P(w|c) = \log \frac{\text{count}(w, c) + 1}{\sum_{v \in V} [\text{count}(v, c) + 1]} = \log \frac{\text{count}(w, c) + 1}{\sum_{v \in V} \text{count}(v, c) + |V|}$$

$\text{count}(w, c)$ 범주 c 에서 단어 w 가 출현한 횟수
 V 훈련 집합의 어휘 목록

코딩에 필요한 것

- 범주별로 문서를 합쳐서 단어(token)를 세기
- 훈련 집합에 출현한 어휘(type) 목록 만들기

훈련 결과 시험

훈련 결과 얻은 것

로그사전확률 $\log P(c)$, 로그가능도 $\log P(w|c)$

시험할 것

실험 집합의 문서 `testdoc`에 대한 $\arg \max_{c \in C} \log P(c|\text{testdoc})$

$$\begin{aligned} & \arg \max_{c \in C} \log P(c|\text{testdoc}) \\ &= \arg \max_{c \in C} \left[\log P(c) + \sum_{w \in \text{testdoc} \cap V} \log P(w|c) \right] \end{aligned}$$

실습 코드: <https://colab.research.google.com/drive/1C8lM22-5ILMCH0CemzXKBCWxrHyc2yIJ>

주어진 것

- 코퍼스(훈련 집합 + 실험 집합)
- 범주 목록

계산할 것

- 로그사전확률 — 범주별
- 로그가능도 — 단어별, 범주별
- 사후확률이 가장 높은 범주 — 문서별

남은 문제

- 모든 단어(혹은 형태소)를 특성으로 쓰는 것이 적절한가?
- 세 가지 이상의 범주에 대한 분류

다른 분류기

로지스틱 회귀분석 / 최대 엔트로피 분류기

- $P(d|c)$ 를 사용하지 않고 $P(c|d)$ 를 직접 추정한다.
- 단어 이외의 특성을 다양하게 사용할 수 있다.

다음 주까지 할 일

SLP3 5장 읽어 오기