2018학년도 2학기 언어와 컴퓨터

제16강 단순 베이즈 분류기와 감정분석 (1)

박수지

서울대학교 인문대학 언어학과

2018년 11월 19일 월요일

이번 주의 주제

1. 텍스트 분류
2. 단어 주머니 가정
3. 베이즈 정리
4. 단순 베이즈 분류기

복습

- 11강: 1종 오류와 2종 오류
- 14강: 조건부확률, 언어 모형
- 15강: 코퍼스 분할, 문장 생성, 평탄화

분류 대상

데이터: “문서”

- 문장, 문단, 게시물, 댓글, 트윗, …

분류 결과

고정된 범주 — 분류 중에 범주를 추가할 수 없다.

분류 예시

- 감정 분석: 긍정 & 부정
- 스팸메일 필터: 스팸 & 스팸 아님
- 저자 식별: 후보1 & 후보2 & …& 후보N
- 주제 분류: 정치 & 경제 & 사회 & 스포츠 & 연예
- …

분류 방법

규칙 기반 분류기 사람이 작성한 규칙을 사용하여 범주를 결정한다.

통계적 분류기 각 범주의 확률을 계산하여 확률이 가장 높은 것을 선택한다.

기계 학습

지도 학습 훈련 집합의 데이터와 정답을 함께 입력하여 규칙을 학습시킨 뒤 실험 집합의 데이터의 범주를 예측함

비지도 학습 훈련 집합의 데이터만 입력하여 규칙을 학습시킨 뒤 [후략]

준비

텍스트 데이터를 수치로 표현하기

- 특성(features) 값들의 벡터

특성 예시

값이 수치로 표현될 수 있어야 한다.

- 문서의 길이 (단어 수)
- “good” 이 출현한 횟수
- 해시태그 개수
- URL 포함 여부
- …

텍스트 분류에서 가장 전형적인 특성: 단어 출현 횟수

단어 주머니(Bag-of-words) 가정

문서를 단어의 주머니처럼 표현한다.
$\Rightarrow$ 단어의 위치를 고려하지 않고 빈도만 고려한다.

예시

단어 주머니 가정에 따르면 아래의 두 한줄평은 같은 문서이다.

1. 정말 좋은 영화였다. 나쁜 점이 없었다.
2. 정말 나쁜 영화였다. 좋은 점이 없었다.

……N-그램을 단어로 간주하는 방식으로 보완할 수 있다.

단어 주머니 가정의 귀결

$P(w_1 w_2 \cdots w_n)$ 대신 $P(w_1, w_2, \cdots, w_n)$을 사용하겠다.

$$
\begin{aligned}
&P(\text{“two point five”}|\text{condition}) \\
&= P(\text{“five point two”}|\text{condition}) \\
&= P(\text{“two”}, \text{“point”}, \text{“five”}|\text{condition})
\end{aligned}
$$

…하지만 $P(\text{“two”}, \text{“point”}, \text{“five”}|\text{condition})$도 구하기 어렵다.

현실

단어 하나의 조건부확률만 구할 수 있다.

- $P(\text{“two”}|\text{condition})$
- $P(\text{“point”}|\text{condition})$
- $P(\text{“five”}|\text{condition})$

단순 베이즈 가정 (Naive Bayes assumption)

각 단어의 출현 확률은 조건부독립이다.

$$P(\text{“two”}, \text{“point”}, \text{“five”}|\text{c}) = P(\text{“two”}|\text{c}) \times P(\text{“point”}|\text{c}) \times P(\text{“five”}|\text{c})$$

연습 문제

$P(\text{“Hong Kong”}|\text{c}) = P(\text{“Hong”}|\text{c}) \times P(\text{“Kong”}|\text{c})$ 으로 계산하는 것이 타당한지 생각해 보자.

분류 문제에서의 조건

범주 $c \in C = \{c_1, c_2, \ldots, c_k\}$

감정 분석에서 단순 베이즈 분류기의 재료

{긍정적인, 부정적인} 문서에서 단어가 출현할 확률

- $P(\text{“좋은”}|+)$ vs. $P(\text{“좋은”}|-)$
- $P(\text{“나쁜”}|+)$ vs. $P(\text{“나쁜”}|-)$
- $\cdots$

확률 추정 방법

단어 출현 횟수를 센다.

참조 N-그램 언어 모형

의문

- 분류 문제에서 필요한 것은 $P(\text{class}|\text{data})$ 이다.
- 그런데 지금 가진 것은 $P(\text{data}|\text{class})$ 뿐이다.

두 가지 조건부확률을 혼용해도 되는가? 안 된다.

해결책

베이즈 정리

베이즈 정리

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

증명

$$P(d|c) = \frac{P(c,d)}{P(c)} \quad \text{(1) 조건부확률의 정의}$$

$$P(c,d) = P(d|c)P(c) \quad \text{(2) (1)의 양변에 } P(c) \text{ 곱하기}$$

$$P(c|d) = \frac{P(c,d)}{P(d)} \quad \text{(3) 조건부확률의 정의}$$

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad \text{(4) (3)의 분자에 (2) 대입}$$

베이즈 정리

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

의미

c로 분류된다는 가설에 대한 믿음을 d의 관측을 통해 업데이트한다.

$P(c|d)$ c의 사후확률(posterior probability)
- 데이터 관측 후의 확률(믿음)

$P(c)$ c의 사전확률(prior probability)
- 데이터 d 관측 전의 확률(믿음)

$P(d|c)$ c의 가능도(likelihood)
- 데이터 d에 대한 정보

통계적 분류기 일반

주어진 문서 $d = w_1 w_2 \cdots w_n$ 가 긍정적$(+)$일 확률 $P(+|d)$과 부정적$(-)$일 확률 $P(-|d)$을 비교하여 더 높은 쪽으로 분류한다.

단순 베이즈 분류기

$P(+|d)$를 아래와 같은 방법으로 계산한다.

$$
\begin{aligned}
P(+|d) &= \frac{P(d|+)P(+)}{P(d)} && \text{(1) 베이즈 정리} \\
&= \frac{P(w_1 w_2 \cdots w_n|+)P(+)}{P(d)} && \text{(2) } d\text{의 내용} \\
&= \frac{P(w_1, w_2, \cdots, w_n|+)P(+)}{P(d)} && \text{(3) 단어 주머니 가정} \\
&= \frac{P(w_1|+) \times \cdots \times P(w_n|+) \times P(+)}{P(d)} && \text{(4) 단순 베이즈 가정}
\end{aligned}
$$

무엇을 어떻게 비교할 것인가?

$$P(+|d) = \frac{P(w_1|+) \times \cdots \times P(w_n|+) \times P(+)}{P(d)}$$

$$P(-|d) = \frac{P(w_1|-) \times \cdots \times P(w_n|-) \times P(-)}{P(d)}$$

관찰

공통분모 $P(d)$는 대소 비교에 영향을 주지 않는다.
$\Rightarrow$ 분자만 비교하면 된다.

$$P(+|d) \geq P(-|d) \iff P(+)\prod_{i=1}^{n} P(w_i|+) \geq P(-)\prod_{i=1}^{n} P(w_i|-)$$

무엇을 어떻게 비교할 것인가?

문제: underflow

컴퓨터에서 확률을 많이 곱하다 보면 0이 되어버릴 수 있다.
$\Rightarrow$ 확률에 로그를 취해서 비교한다.

$$P(+|d) \geq P(-|d)$$

$$\iff \log P(+|d) \geq \log P(-|d)$$

(로그함수가 단순증가함수이므로)

$$\iff \log \left[P(+) \prod_{i=1}^{n} P(w_i|+) \right] \geq \log \left[P(-) \prod_{i=1}^{n} P(w_i|-) \right]$$

$$\iff \log P(+) + \sum_{i=1}^{n} \log P(w_i|+) \geq \log P(-) + \sum_{i=1}^{n} \log P(w_i|-)$$

(곱의 로그는 로그의 합과 같으므로)

계산해야 할 것

로그사전확률 $\log P(+)$

로그가능도 $\sum_{i=1}^{n} \log P(w_i|+)$

추정하는 방법

개수를 센다.

사전확률 $P(+) = \dfrac{\text{(훈련 집합의 긍정적 문서 개수)}}{\text{(훈련 집합 전체의 문서 개수)}}$

가능도 $P(w_i|+) = \dfrac{\text{(훈련 집합의 긍정적 문서 전체에서 } w_i \text{의 빈도)}}{\text{(훈련 집합의 긍정적 문서 전체의 단어 개수)}}$

주의

코퍼스 전체가 아닌 훈련 집합에서만 세어야 한다.

문제

긍정적 문서에만 출현하는 단어가 존재할 수 있다.
⇒ 부정적 문서의 확률이 0이 되지 않도록 평탄화가 필요하다.

가능도 추정 방법: 평탄화 추가

$$P(w_i|+) = \frac{\text{(훈련 집합의 긍정적 문서 전체에서 } w_i \text{의 빈도)} + 1}{\text{(훈련 집합의 긍정적 문서 전체의 단어(token) 개수)} + |V|}$$

$|V|$ 훈련 집합 전체의 단어(type) 가짓수

오늘의 내용

- 감정 분석 & 통계적 분류기: $P(+|d)$와 $P(-|d)$를 비교하는 문제
- 단순 베이즈 분류기: $P(+|d)$ 등을 아래와 같이 표현하기

$$P(+|d) = \frac{P(w_1|+) \times \cdots \times P(w_n|+) \times P(+)}{P(d)}$$

  - 베이즈 정리
  - 단어 주머니 가정: 단어의 위치를 무시하고 횟수만 고려한다.
  - 단순 베이즈 가정: 모든 단어의 출현을 조건부독립으로 가정한다.
- 실제적인 문제
  - 확률의 값은 훈련 집합에서의 출현 빈도로 추정한다.
  - 확률의 곱 대신 확률의 로그 값의 합을 계산한다.
  - 단어의 조건부확률을 구할 때 평탄화를 적용한다.

남은 문제

- 학습된 분류기의 성능을 측정하기

  척도 정확도(Accuracy), 정밀도(Precision), 재현율(Recall)
- 이 모든 과정을 파이썬에서 구현하기
  - 으악!
    - 겁내지 말아요.