

Naïve Bayes
Classification

ToBig's 9기 김수지

Naïve Bayes Classification

text data classification 중심으로

Contents

Unit 01 | 개요

Unit 02 | Statistical Concept

Unit 03 | Naïve Bayes Classification

Unit 04 | Example

Unit 05 | Laplace Smoothing

Unit 06 | Logarithm

Unit 07 | 실습

Unit 01 | 개요

나이브 베이즈 분류는 Supervised Learning 알고리즘 중 하나!

Cf) Supervised Learning 이란?

Unit 01 | 개요

나이브 베이즈 분류는 Supervised Learning 알고리즘 중 하나!

Cf) Supervised Learning 이란?

Labeled data를 이용한 학습
예를들어 타겟 변수(label)가 '커피 종류' 일 때

(샷, 물) -> (아메리카노)
(샷, 우유) -> (라떼)
(샷, 우유, 바닐라시럽) -> (바닐라라떼)

위 정보들을 train data로 주고

(샷, 우유) 가 주어졌을 때
아메리카노/라떼/바닐라라떼 중 하나로 분류를 하는 것!!

Unit 01 | 개요

Q: 로지스틱 회귀, KNN, Trees, SVM 등 분류 알고리즘이 이미 많이 있습니다.
Naïve Bayes 는 어떤 경우에 사용하면 좋나요?

- 분류에 필요한 파라미터를 추정하기 위한 training data가 적을 때 (300개 이하)
- Text data 분류에서 강세를 보임
- 실무에서 인기가 많다
(비교적 간단한 편이어서 저장 공간과 계산시간 측면에서 효율적이기 때문!)

ex) 스팸 메일 분류

Unit 01 | 개요

Q: 로지스틱 회귀, KNN, Trees, SVM 등 분류 알고리즘이 이미 많이 있습니다.
Naïve Bayes 는 어떤 경우에 사용하면 좋나요?

- 분류에 필요한 파라미터를 추정하기 위한 training data가 적을 때 (300개 이하)
- Text data 분류에서 강세를 보임
- 실무에서 인기가 많다
(비교적 간단한 편이어서 저장 공간과 계산시간 측면에서 효율적이기 때문!)

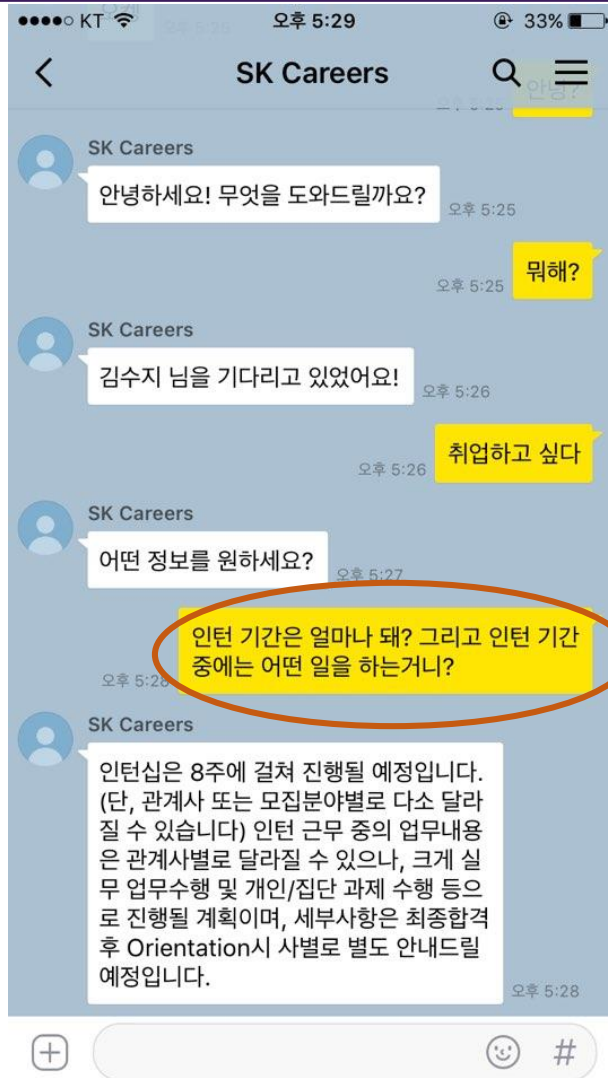
ex) 스팸 메일 분류

→ 예시가 식상해요 ㅋㅋ 실무적인 예를 들어주세요

ex) 챗봇

→ '채용' 관련한 챗봇을 만든다고 상상해봅시다!

Unit 01 | 개요



질문을 받아서 이 질문이
'지원서 접수'
'전형절차'
'인턴십 관련'
'일반사항'
4개의 카테고리 중
어디에 속하는지
1차 분류를 할 때
쓰이는 것이 Naïve Bayes!!

궁금한 사항에 대해서 FAQ를
먼저 확인하시면
빠른 답변을 얻을 수 있습니다

FAQ		Q & A	
지원서 접수	전형절차	인턴십 관련	일반사항

Q 기졸업자 또는 대학교 3학년 이하에 재학중인 사람도 인턴십 지원이 가능한가요?

Q 인턴사원 채용 전형에서 불합격한 경우, 다음 신입사원 채용 전형에 다시 지원할 수 있나요?

Q 인턴 기간은 얼마나 되나요? 인턴십 기간 중에는 어떤 일을 하게 되나요?

A 인턴십은 8주에 걸쳐 진행될 예정입니다. (단, 관계사 또는 모집분야별로 다소 달라질 수 있습니다) 인턴 근무 중의 업무내용은 관계사별로 달라질 수 있으나, 크게 실무 업무수행 및 개인/집단 과제 수행 등으로 진행될 계획이며, 세부사항은 최종합격 후 Orientation시 사별로 별도 안내드릴 예정입니다.

Q 근무지는 어디인가요? 지방 근무가 가능한가요?

Q 인턴 기간중의 급여는 어떻게 되나요? 복리후생으로는 어떤 것들이 있나요?

Q 인턴십 기간 중의 평가는 어떻게 이루어지나요? 신입사원 입사 여부는 언제쯤 확실하게 알 수 있나요?

Q 인턴 근무 후, 신입사원으로 입사하는 비율은 어느 정도 인가요?

Q 신입사원으로 입사하게 될 경우, 인턴 기간에 근무한 부서로 입사하게 되나요?

Q 신입사원 입사시 인턴으로 근무한 관계사 외에 다른 관계사로 입사할 수 있나요?

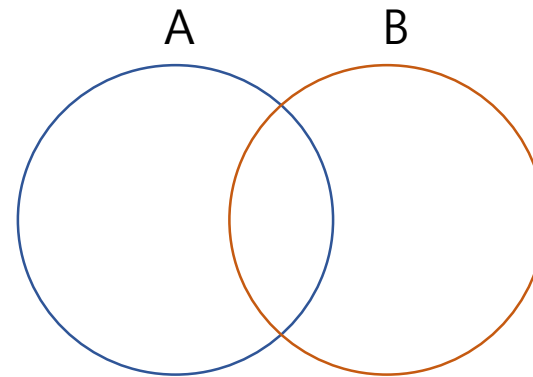
Unit 02 | Statistical Concept: Conditional Probability

Conditional Probability(조건부 확률)

Def) 사건 A, B에 대해 $P(B) \neq 0$ 일 때,
사건 B가 일어났을 때 사건 A가 일어날 확률은 다음과 같이 정의한다.

$$\cdot P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\cdot P(A \cap B) = P(A|B)P(B)$$



Unit 02 | Statistical Concept: Bayes' Theorem

Bayes' Theorem(베이즈 정리)

두 확률 변수의 사전확률과 사후확률 사이의 관계를 나타내는 정리
베이즈 정리를 이용해서 사전확률로부터 사후확률을 구할 수 있다.

Def) 사건 A, B에 대해 $P(B) \neq 0$ 일 때,

$$\cdot P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

· $P(A)$ 를 사전확률, $P(A|B)$ 를 사건 A의 사후확률이라고 한다.

Unit 02 | Statistical Concept: Independence

Independence(독립)

Def) 사건 A, B는 아래의 조건 중 하나일 때 독립이다.

- $P(A \cap B) = P(A)P(B)$
- A, B 둘 중 하나의 확률값이 0 또는 1
- $P(A|B) = P(A)$

(사건(event)이 두 개 이상일 때) E_1, \dots, E_n 이 독립이면 다음이 성립한다.

- $P(E_i \cap E_j) = P(E_i)P(E_j)$
- $P(E_i \cap E_j \cap E_k) = P(E_i)P(E_j)P(E_k)$
-
- $P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_1)P(E_2) \dots P(E_n)$

Unit 02 | Statistical Concept: Conditional Independence

Conditional Independence(조건부 독립)

사건 F, G가 사건 H가 주어진 상황에서 조건부 독립이라면 다음이 성립한다.

- $P(F \cap G | H) = P(F | H) P(G | H)$

모수 θ 를 가진 분포에서 확률변수 Y_1, \dots, Y_n 들이 뽑혔다고 하자.

모두 조건부 독립이라면 위의 정의에 따라 두 가지를 이끌어낼 수 있다.

- $P(y_1, \dots, y_n | \theta) = P_{Y_1}(y_1 | \theta) \times \dots \times P_{Y_n}(y_n | \theta) = \prod_{i=1}^n P_{Y_i}(y_i | \theta)$

- $P(y_i | \theta, y_j) = P(y_i | \theta)$

- 1) Joint density를 marginal density들의 곱으로 표현할 수 있다.
- 2) θ 를 알고 있다면, Y_j 가 Y_i 에 대한 추가적인 정보를 주지 못 한다.

Unit 03 | Naïve Bayes Classification

Naïve Bayes Classification (나이브 베이즈 분류)

- 베이즈 정리를 이용하여 분류하고자 하는 대상의 각 분류별 확률을 측정하여, 그 확률이 큰 쪽으로 분류하는 방법
- 분류할 카테고리가 C_1, C_2, C_3 일 때
새로운 데이터 (x_1, \dots, x_p) 에 대해서 $P(C_k|x_1, \dots, x_p)$ 가 가장 큰 k로 분류하는 방법
- 계산의 편의를 위해서 C_k 가 주어졌을 때 x_1, \dots, x_p 사이의 독립을 가정한다.

Unit 03 | Naïve Bayes Classification

Naïve Bayes Classification (나이브 베이즈 분류)

- 베이즈 정리를 이용하여 분류하고자 하는 대상의 각 분류별 확률을 측정하여, 그 확률이 큰 쪽으로 분류하는 방법
- 분류할 카테고리가 C_1, C_2, C_3 일 때
새로운 데이터 (x_1, \dots, x_p) 에 대해서 $P(C_k|x_1, \dots, x_p)$ 가 가장 큰 k로 분류하는 방법
- 계산의 편의를 위해서 C_k 가 주어졌을 때 x_1, \dots, x_p 사이의 독립을 가정한다.

ex) 이메일에 대해서 스팸인지 아닌지 분류를 해야할 때
이메일에 들어가 있는 단어(x)들 ('여름세일', '전품목할인', '절호의 기회')에 대해서
 $P(\text{스팸} | \text{'여름세일', '전품목할인', '절호의 기회'}) > P(\text{스팸}x | \text{'여름세일', '전품목할인', '절호의 기회'})$
→ 이 메일은 스팸이군!!

Unit 03 | Naïve Bayes Classification

$P(C_k|x_1, \dots, x_p)$ 를 계산해보자

$$P(C_k|x_1, \dots, x_p) = \frac{P(x_1, \dots, x_p|C_k)P(C_k)}{P(x)} \quad \text{by Bayes' Theorem}$$

$$\propto P(x_1, \dots, x_p|C_k)P(C_k) \quad \text{by Conditional Independence}$$

$$= P(x_1|C_k) \dots P(x_p|C_k)P(C_k)$$

즉, $P(C_k|x_1, \dots, x_p)$ 에 대해서 $P(x_1|C_k), \dots, P(x_p|C_k), P(C_k)$ 를 각각 계산할 수 있으면 된다!

Unit 03 | Naïve Bayes Classification

$P(C_k|x_1, \dots, x_p)$ 를 계산해보자

$$\begin{aligned} P(C_k|x_1, \dots, x_p) &= \frac{P(x_1, \dots, x_p|C_k)P(C_k)}{P(x)} \quad \text{by Bayes' Theorem} \\ &\propto P(x_1, \dots, x_p|C_k)P(C_k) \quad \text{by Conditional Independence} \\ &= P(x_1|C_k) \dots P(x_p|C_k)P(C_k) \end{aligned}$$

즉, $P(C_k|x_1, \dots, x_p)$ 에 대해서 $P(x_1|C_k), \dots, P(x_p|C_k), P(C_k)$ 를 각각 계산할 수 있으면 된다!

ex) X_1 ~연속형 변수(정규분포), X_2 ~범주형 변수(2가지 범주), C_1, C_2 두 가지로 분류한다고 가정.

$$C_1 \implies \begin{cases} X_1 \rightarrow \mu_{11}, \sigma_{11} \\ X_2 \rightarrow p_{11}, p_{12} \\ P(C_1) \end{cases} \qquad C_2 \implies \begin{cases} X_1 \rightarrow \mu_{21}, \sigma_{21} \\ X_2 \rightarrow p_{21}, p_{22} \\ P(C_2) \end{cases}$$

Unit 03 | Naïve Bayes Classification

데이터가 주어졌을 때 모수에 대한 추정 및 모델이 만들어지는 과정:

- 1) 데이터를 C_1, C_2 에 따라 나눈다.
- 2) C_k 에 따라 모수추정 (C_1 에 속한 데이터 수를 m , 전체를 n 이라고 하자)

$$\mu_{11} = \frac{1}{m} \sum_{i=1}^m x_{1i}, \quad \sigma_{11} = \frac{1}{m} \sum_{i=1}^m (x_{1i} - \mu_{11})^2 \quad (\text{평균과 표준편차})$$

$$p_{11} = \frac{\#(C_1, <\text{범주}_1>)}{\#(C_1)}, \quad p_{12} = \frac{\#(C_1, <\text{범주}_2>)}{\#(C_1)} \quad (\text{숫자 세기}) \quad p(C_1) = \frac{m}{n}$$

- 3) 이것으로 $P(x_1|C_k), \dots, P(x_p|C_k), P(C_k)$ 를 구할 수 있고, $P(C_k|x_1, \dots, x_p)$ 도 계산할 수 있다.
- 4) 새로운 데이터를 예측할 수 있다.

Unit 04 | Example01: numeric data

- 간단한 나이브 베이즈 분류기를 만들어서 하나의 데이터를 예측해보자!

합불(C_k)	서류 점수(X1)	석사 여부(X2)
합	10	X
합	6	0
합	7	X
불	6	X
불	4	0

합불(C_k)	서류 점수(X1)	
	평균	분산
합	7.7	2.9
불	5	1

Q: 서류에서 9점, 석사 학위가 없는 사람은 합격 했을까?
→ $P(\text{합}|(9,x))$ vs $P(\text{불}|(9,x))$ 둘 중 큰 값을 찾자!

Unit 04 | Example01: numeric data

- 간단한 나이브 베이즈 분류기를 만들어서 하나의 데이터를 예측해보자!

합불(C_k)	서류 점수(X1)	석사 여부(X2)
합	10	X
합	6	O
합	7	X
불	6	X
불	4	O

합불(C_k)	서류 점수(X1)	
	평균	분산
합	7.7	2.9
불	5	1

Q: 서류에서 9점, 석사 학위가 없는 사람은 합격 했을까?

→ $P(\text{합}|(9,x))$ vs $P(\text{불}|(9,x))$ 둘 중 큰 값을 찾자!

- $P(\text{서류}=9|\text{합})P(\text{석사}=x|\text{합})P(\text{합}) = 0.1239 * \frac{2}{3} * \frac{3}{5} = 0.05$
- $P(\text{서류}=9|\text{불})P(\text{석사}=x|\text{불})P(\text{불}) = 0.0002 * \frac{1}{2} * \frac{2}{5} = 0.00004$

Unit 04 | Example02: text data

다음과 같이 5개의 학습 문서가 주어지고, 카테고리는 '지원서 접수', '인턴십' 두 개가 존재한다고 하자!

	input으로 들어온 문장	문장의 키워드	카테고리
1	지원서 제출이 안되는데 왜 그럴까요?	지원서, 제출	지원서 접수
2	지원서 제출이 잘 되었는지 확인을 하고 싶은데요	지원서, 제출, 확인	지원서 접수
3	지원서 작성시 임시저장과 제출은 어떤 차이가 있나요?	지원서, 제출, 작성, 임시저장, 차이	지원서 접수
4	인턴 기간은 얼마나 되나요? 인턴 기간 중에는 어떤 일을 하게 되나요?	인턴, 기간, 일	인턴십
5	인턴 기간 중의 급여는 어떻게 되나요? 복리후생으로는 어떤 것들이 있나요?	인턴, 기간, 급여, 복리후생	인턴십

Q: 지원서는 어떻게 제출하는 건가요?

→ (지원서, 제출) 의 분류별 확률을 구해보자!

- $\text{count}(\text{지원서 접수}) = 10$
- $\text{count}(\text{인턴십}) = 7$
- $\text{count}(\text{지원서}, \text{지원서 접수}) = 3$
- $\text{count}(\text{제출}, \text{지원서 접수}) = 3$
- $\text{count}(\text{지원서}, \text{인턴십}) = 0$
- $\text{count}(\text{제출}, \text{인턴십}) = 0$

Unit 04 | Example02: text data

- 주어진 질문이 '지원서 접수'일 확률은 $P(\text{지원서 접수} \mid \text{words}) = \frac{P(\text{words} \mid \text{지원서 접수}) * P(\text{지원서 접수})}{P(\text{words})}$
- 주어진 질문이 '인턴십'일 확률은 $P(\text{인턴십} \mid \text{words}) = \frac{P(\text{words} \mid \text{인턴십}) * P(\text{인턴십})}{P(\text{words})}$
- 위 두 확률은 대소만 비교하기 때문에 아래 두 값만 구하면 된다!
 $P(\text{words} \mid \text{지원서 접수}) * P(\text{지원서 접수}) \quad \dots (1)$
 $P(\text{words} \mid \text{인턴십}) * P(\text{인턴십}) \quad \dots (2)$

Unit 04 | Example02: text data

주어진 질문이 '지원서 접수'일 확률은 $P(\text{지원서 접수} \mid \text{words}) = \frac{P(\text{words} \mid \text{지원서 접수}) * P(\text{지원서 접수})}{P(\text{words})}$

주어진 질문이 '인턴십'일 확률은 $P(\text{인턴십} \mid \text{words}) = \frac{P(\text{words} \mid \text{인턴십}) * P(\text{인턴십})}{P(\text{words})}$

위 두 확률은 대소만 비교하기 때문에 아래 두 값만 구하면 된다!

$$P(\text{words} \mid \text{지원서 접수}) * P(\text{지원서 접수}) \quad \dots (1)$$

$$P(\text{words} \mid \text{인턴십}) * P(\text{인턴십}) \quad \dots (2)$$

(1) $P(\text{지원서, 제출} \mid \text{지원서 접수}) * P(\text{지원서 접수})$

$$= \frac{P(\text{지원서} \mid \text{지원서 접수}) * P(\text{제출} \mid \text{지원서 접수}) * P(\text{지원서 접수})}{P(\text{지원서 접수})} = \frac{3}{10} * \frac{3}{10} * \frac{3}{5} = 0.054$$



$$\frac{P(\text{지원서} \mid \text{지원서 접수})}{P(\text{지원서 접수})} = \frac{\text{count}(\text{지원서}, \text{지원서 접수})}{\text{count}(\text{지원서 접수})} = \frac{3}{10}$$

Unit 04 | Example02: text data

주어진 질문이 '지원서 접수'일 확률은 $P(\text{지원서 접수} \mid \text{words}) = \frac{P(\text{words} \mid \text{지원서 접수}) * P(\text{지원서 접수})}{P(\text{words})}$

주어진 질문이 '인턴십'일 확률은 $P(\text{인턴십} \mid \text{words}) = \frac{P(\text{words} \mid \text{인턴십}) * P(\text{인턴십})}{P(\text{words})}$

위 두 확률은 대소만 비교하기 때문에 아래 두 값만 구하면 된다!

$$P(\text{words} \mid \text{지원서 접수}) * P(\text{지원서 접수}) \quad \dots (1)$$

$$P(\text{words} \mid \text{인턴십}) * P(\text{인턴십}) \quad \dots (2)$$

$$(1) P(\text{지원서, 제출} \mid \text{지원서 접수}) * P(\text{지원서 접수})$$

$$= P(\text{지원서} \mid \text{지원서 접수}) * P(\text{제출} \mid \text{지원서 접수}) * P(\text{지원서 접수}) = \frac{3}{10} * \frac{3}{10} * \frac{3}{5} = 0.054$$

$$(2) P(\text{지원서, 제출} \mid \text{인턴십}) * P(\text{인턴십})$$

$$= P(\text{지원서} \mid \text{인턴십}) * P(\text{제출} \mid \text{인턴십}) * P(\text{인턴십}) = \frac{0}{7} * \frac{0}{7} * \frac{2}{5} = 0$$

주어진 질문이 '지원서 접수'일 확률이 0.054로 더 크기 때문에 '지원서 접수'로 분류!!

Unit 05 | Laplace Smoothing

나이브베이지 알고리즘의 문제점1

학습 데이터에 없는 단어가 주어졌을 때, 등장하는 빈도가 모두 0이 되어 어느 것에도 분류할 수 없다.

Q: 지원서를 팩스로 제출해도 되나요?

→ (지원서, 팩스, 제출)의 분류별 확률을 구해보자!

$\text{count}(\text{팩스}, \text{지원서 접수})=0$

$\text{count}(\text{팩스}, \text{인턴십})=0$

Unit 05 | Laplace Smoothing

나이브베이지 알고리즘의 문제점1

학습 데이터에 없는 단어가 주어졌을 때, 등장하는 빈도가 모두 0이 되어 어느 것에도 분류할 수 없다.

Q: 지원서를 팩스로 제출해도 되나요?

→ (지원서, 팩스, 제출)의 분류별 확률을 구해보자!

count(팩스, 지원서 접수)=0

count(팩스, 인터넷)=0

(1) $P(\text{지원서, 팩스, 제출} \mid \text{지원서 접수}) * P(\text{지원서 접수})$

따옴? 0?

$$= P(\text{지원서} \mid \text{지원서 접수}) * P(\text{팩스} \mid \text{지원서 접수}) * P(\text{제출} \mid \text{지원서 접수}) * P(\text{지원서 접수}) = \frac{3}{10} * \frac{0}{10} * \frac{3}{10} * \frac{3}{5} = 0$$

Unit 05 | Laplace Smoothing

나이브베이지 알고리즘의 문제점1

학습 데이터에 없는 단어가 주어졌을 때, 등장하는 빈도가 모두 0이 되어 어느 것에도 분류할 수 없다.

Q: 지원서를 팩스로 제출해도 되나요?

→ (지원서, 팩스, 제출)의 분류별 확률을 구해보자!

count(팩스, 지원서 접수)=0

count(팩스, 인턴십)=0

(1) $P(\text{지원서, 팩스, 제출} \mid \text{지원서 접수}) * P(\text{지원서 접수})$

$$= P(\text{지원서} \mid \text{지원서 접수}) * P(\text{팩스} \mid \text{지원서 접수}) * P(\text{제출} \mid \text{지원서 접수}) * P(\text{지원서 접수}) = \frac{3}{10} * \frac{0}{10} * \frac{3}{10} * \frac{3}{5} = 0$$

(2) $P(\text{지원서, 팩스, 제출} \mid \text{인턴십}) * P(\text{인턴십})$

$$= P(\text{지원서} \mid \text{인턴십}) * P(\text{팩스} \mid \text{인턴십}) * P(\text{제출} \mid \text{인턴십}) * P(\text{인턴십}) = \frac{0}{7} * \frac{0}{7} * \frac{0}{7} * \frac{2}{5} = 0$$

Unit 05 | Laplace Smoothing

Laplace Smoothing

categorical한 데이터의 확률을 부드럽게 만들어주는 기법

사건을 예측할 때 사전에 관측되지 않은 카테고리에도 0이라는 확률을 주지 않는 역할

- 분모, 분자에 적당한 숫자를 더해 주어서 0이 되는 것을 막자!
- 보통 빈도에 1씩 더해주는데, 빈도를 더해주는 공식은 아래와 같다.

$$P(x|c) = \frac{\text{count}(x, c) + 1}{\sum_{x \in v} \text{count}(x, c) + |v|}$$

- $|v|$ 는 학습데이터에서 나오는 유일한 단어의 수가 된다.

Unit 05 | Laplace Smoothing

Laplace Smoothing

categorical한 데이터의 확률을 부드럽게 만들어주는 기법

사건을 예측할 때 사전에 관측되지 않은 카테고리에도 0이라는 확률을 주지 않는 역할

- 분모, 분자에 적당한 숫자를 더해 주어서 0이 되는 것을 막자!
- 보통 빈도에 1씩 더해주는데, 빈도를 더해주는 공식은 아래와 같다.

$$P(x|c) = \frac{\text{count}(x, c) + 1}{\sum_{x \in v} \text{count}(x, c) + |v|}$$

- $|v|$ 는 학습데이터에서 나오는 유일한 단어의 수가 된다.
- $P(\text{지원서}, \text{팩스}, \text{제출} | \text{지원서 접수}) * P(\text{지원서 접수})$
= $P(\text{지원서} | \text{지원서 접수}) * P(\text{팩스} | \text{지원서 접수}) * P(\text{제출} | \text{지원서 접수}) * P(\text{지원서 접수})$
= $\frac{3+1}{10+6} * \frac{0+1}{10+6} * \frac{3+1}{10+6} * \frac{3}{5} = 0.002$
- ‘팩스’라는 단어는 학습되지 않았지만 ‘지원서 접수’ 카테고리 분류할 수 있다.

Unit 06 | Logarithm

나이브베이즈 알고리즘의 문제점2

$P(\text{지원서 접수} \mid \text{words})$ 는 각 단어의 확률의 곱으로 계산되는데

각 확률은 1이하이기 때문에 검사해야 할 단어가 많을 경우 값이 0에 가까워져 구분이 어려워 질 수 있다.

$$P(C_k \mid x_1, \dots, x_p) = P(x_1 \mid C_k) \dots P(x_p \mid C_k) P(C_k) = 0.00000000\dots$$

Unit 06 | Logarithm

나이브베이즈 알고리즘의 문제점2

$P(\text{지원서 접수} \mid \text{words})$ 는 각 단어의 확률의 곱으로 계산되는데

각 확률은 1이하이기 때문에 검사해야 할 단어가 많을 경우 값이 0에 가까워져 구분이 어려워 질 수 있다.

$$P(C_k | x_1, \dots, x_p) = P(x_1 | C_k) \dots P(x_p | C_k) P(C_k) = 0.00000000 \dots$$

Logarithm

\log 를 이용해서 언더플로우를 방지한다

\log 함수는 단조 증가하므로 대소 관계는 변하지 않는다

$$\begin{aligned} & \log(P(x_1 | C_k) \dots P(x_p | C_k) P(C_k)) \\ &= \log(P(x_1 | C_k)) + \dots + \log(P(x_p | C_k)) + \dots + \log(P(C_k)) \end{aligned}$$

Unit 07 | 실습

실습 데이터 구성

스팸인지 아닌지 labeled 된 문자 메시지로 구성. 총 5572개의 관측치
70%는 train data로 만들어 NBC모델을 학습시키고, 나머지 30%는 test data로 모델의 성능을 테스트해보자

	type	text
1	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	ham	Ok lar... Joking wif u oni...
3	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over 18's
4	ham	U dun say so early hor... U c already then say...
5	ham	Nah I don't think he goes to usf, he lives around here though

Unit 07 | 실습

① text 전처리

Step1: Corpus(말뭉치) 만들기

Step2: text 전처리

Q: text전처리를 왜 해줘야하죠?

A: NBC모델은 단어 단위로 학습합니다. 그러니 일괄적으로 소문자로 바꿔주고, 어근추출도 하고 구두점도 없애서 컴퓨터가 알아보기 쉽게 바꿔줘야겠죠?!

- ① tm_map()을 이용해 말뭉치를 mapping한다.
- ② tolower()을 이용해 주어진 문장들을 모두 소문자로 바꿔준다.
- ③ removeNumbers()를 이용해 문장 안 숫자들을 제거한다.
- ④ stopwords()를 이용해 to, and, but, or과 같이 의미 없는 단어들을 제거한다.
- ⑤ removePunctuation을 이용해 구두점 제거한다.

Step3: text 어근추출(stemming)

Step4: text 토큰화(tokenization)


Unit 07 | 실습

전처리 후 얻은 sms_dtm

행은 메시지, 열은 전체 메시지에서 한번이라도 등장한 단어예요.
행렬 속 숫자가 의미하는 것은 '메시지에 해당 단어가 등장한 횟수'입니다.
저희 데이터는 5574행, 6597열로 구성된 sparse matrix입니다.

	text	apple		red	zoo
1	I love apple. Because apple is red!	2	..	1	0
2	I love tobigs!!	0	..	0	0
3	I want to go zoo.	0	..	0	1
..					
5574					

행렬 중 이렇게
대부분의 원소들이
0으로 채워진 행렬을
sparse matrix라고 해요!!



Unit 07 | 실습

② train data/test data 나누기

① train:test=7:3으로 나눌건데

위에서 얻은 sms_dtm을 단순히 70%되는 지점으로 딱 잘랐어요.

② sms_dtm의 열이 너무 많아서 학습 속도는 느려지고 컴퓨터 메모리만 차지하는 것 같아요.

그래서 등장 빈도가 5 미만인 단어들은 삭제합니다.

그랬더니 열이 6579->1158로 대폭 줄었어요!

③ NBC모델은 단어의 빈도는 관심이 없고, 등장여부만 중요해요.

그래서 위에서 얻은 sms_dtm을

convert_counts라는 간단한 사용자 함수를 만들어서 변형시킬 겁니다! 아래 처럼요!

	text	apple		red	zoo
1	I love apple. Because apple is red!	1	..	1	0
2	I love tobigs!!	0	..	0	0
3	I want to go zoo.	0	..	0	1
..					
5574					

apple이 2번 등장했음에도 불구하고 1로 바뀌었어요. 등장했으면 1, 아니면 0인 matrix예요

Unit 07 | 실습

③ NB모델 train data로 train시키기

④ NB모델 test data로 예측하고 성능평가하기

text라서 전처리하고 train/test 나누는데 오래 걸린거예요.
Naïve bayes는 함수 한 줄로 쓱쓱

<필요한 패키지>

```
install.packages("tm")
```

```
install.packages("SnowballC")
```

```
install.packages("wordcloud")
```

```
install.packages("naivebayes")
```

```
install.packages("caret")
```

Unit 08 | 과제

파이썬을 이용하여 Naive Bayes Classification 해보기

train.csv를 불러와 NB모델을 훈련시키고 예측한 뒤, 정확도까지 구합니다.

train.csv의 target변수는 Survived(0 or 1)입니다.

train/test set으로 나눌 때는 앞서 배운 k-fold cross validation을 써서 정확도를 구해주시기 바랍니다.

Hint

```
from sklearn.naïve_bayes import GussianNB  
from sklearn.model_selection import Kfold  
from sklearn.model_selection import cross_val_score
```

Q & A

들어주셔서 감사합니다.