NLP - 문서 분류

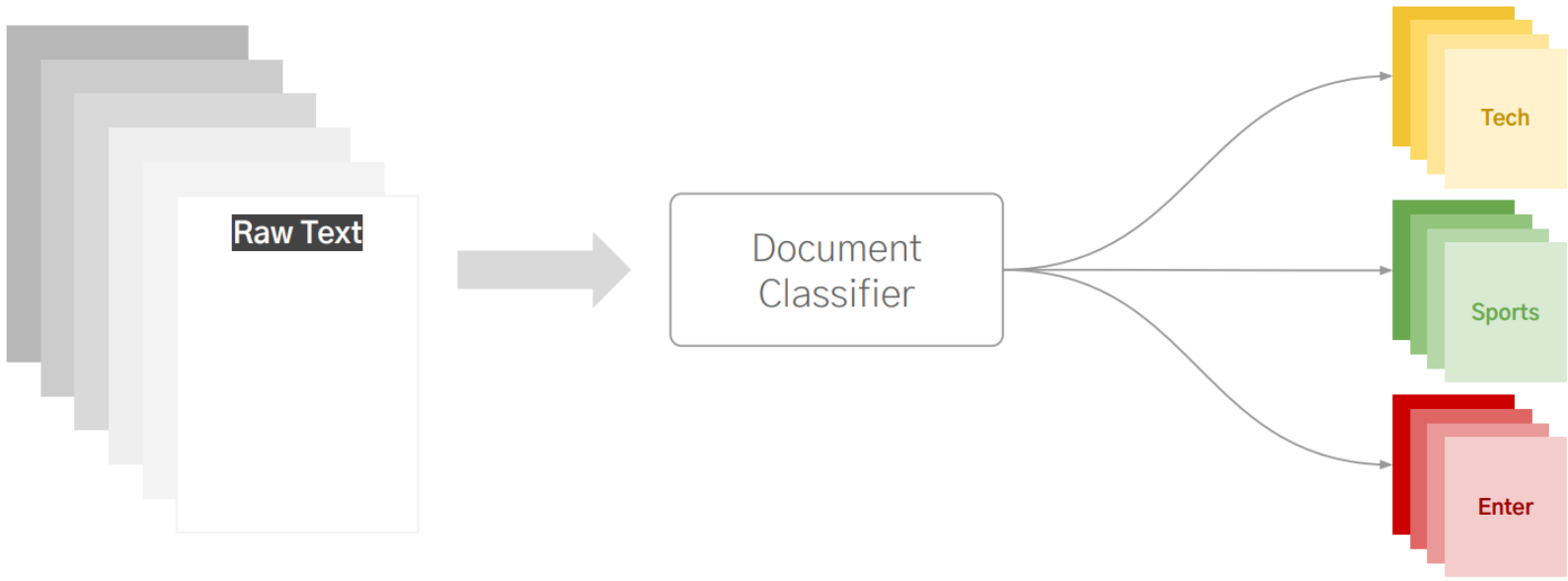
Index

Document Classification

문서 분류

문서 분류 (Document Classification)

- 문서를 사전에 구성된 그룹으로 분류하는 모델
  - 카테고리 분류, 감정 분석, 언어 탐지 등
- 텍스트 분류는 텍스트를 빠르고 비용 효율적으로 적용이 가능

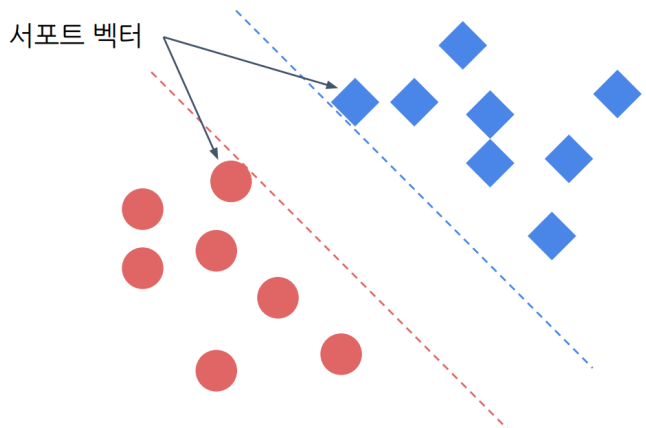


분류 모델 (1) – 나이브 베이즈 분류

- 베이즈 정리를 사용하는 분류 모델
- 학습 데이터에서 추출한 이미 알고 있는 사전 확률을 바탕으로 사후 확률을 계산하여 분류
- 성능 개선을 위한 방법
  - 불용어 처리: 분류를 판단하는데 불필요한 단어 제거
  - 원형복원: 같은 의미의 다른 표현을 원형 복원하여 표준화
  - N-gram: n개의 단어 묶음. 문맥을 포함할 수 있음
  - TF-IDF: 단순 빈도 기반이 아니라 문서 빈도-역문서 빈도에 기반하여 각 단어의 점수를 결정

분류 모델 (2) – 서포트 벡터 머신

- SVM(Support Vector Machines)은 제한된 양의 데이터를 처리할 때 좋은 성능을 보이는 분류 알고리즘
- 주어진 그룹에 속하는 벡터와 그룹에 속하지 않는 벡터 간 분류를 결정
- SVM는 많은 학습 데이터가 필요하지 않지만, 나이브 베이즈 분류보다 좋은 성능을 내기 위해서는 더 많은 계산 리소스가 필요



Bayes Classifier

Bayes Classifier

Bayes Classifier

- 데이터의 조건부 확률에 기반한 분류 => 데이터 중심
- 범주형 자료에만 적용 가능 : 수치형 자료(예.키, 몸무게, 주가 등)는 범주형으로 변환 필요
- 좋은 성능을 위해서는 대량 데이터가 필요

- 종류
  - Exact Bayes Classifier
    - 조건부 확률과 베이즈 확률에 기반
    - 조건이 많으면 계산이 어려움
  - Naive Bayes Classifier
    - 독립변수가 많을 때 간단히 계산

확률

- 어떤 사건이 발생할 가능성(사건 결과의 비율)
- 확률 = 가능성 = %
- 어떤 사건이 발생할 가능성을 0 ~ 1 값으로 표현한 것

확률 계산

- P(A) = A의 개수 / S의 개수
- S는 표본 공간 (Sample Space)
- A는 사건 (event)

$$P(A) = \frac{\text{관심사건}(A)}{\text{표본공간}(S)} = \frac{A\text{의개수}}{S\text{의개수}} = \frac{n_A}{n_S}$$

조건부 확률

- P(A), P(B) 두 개의 사건이 발생함
- P(B|A) = A 조건이 주어진 상태에서 B가 발생 할 확률

사례

- (확률) 동전 2번 던져서 모두 앞면이 나오는 경우
  - S = [HH, HT, TH, TT], A = [HH]
  - P(A) = 1/4 = 0.25

- (조건부확률) 동전 하나는 이미 앞면이라고 알고 있는 경우
  - S = [HH, HT, TH], A = [HH]
  - P(A) = 1/3 = 0.333

- 따라서, 조건부확률에서는 표본 공간 S가 바뀐다.

$$P(B \mid A) = \frac{n(A \cap B)}{n(A)} = \frac{P(A \cap B)}{P(A)}$$

예제: 사기 재무보고 예측

- 이전 법적문제를 가지고 있는 기업의 경우 사기일 확률은?
- 사건 A = [사기(fraud), 정직(honest)]
- 사건 B = [이전 법적문제 유(yes), 이전 법적문제 무(no)]

$$P(fraud \mid yes) = \frac{50}{230} = 0.22 = 22\,\%$$
$$P(fraud) = \frac{100}{1000} = 0.1$$

	이전 법적문제 유	이전 법적문제 무	계
사기	50	50	100
정직	180	720	900
계	230	770	1000

- 조건이 여러 개일 경우
  - P(사기 | 이전법적문제=yes,회사규모=small) = 1/2 = 50%
  - P(사기 | 이전법적문제=no,회사규모=small) = 0/3 = 0%

회사	이전 법적문제 유	회사규모	상태
1	yes	small	정직
2	no	small	정직
3	no	large	정직
4	no	large	정직
5	no	small	정직
6	no	small	정직
7	yes	small	사기
8	yes	large	사기
9	no	large	사기
10	yes	large	사기

Naïve Bayes Classifier

나이브 베이즈 분류

나이브 베이즈 분류 (Naïve Bayes Classifier)

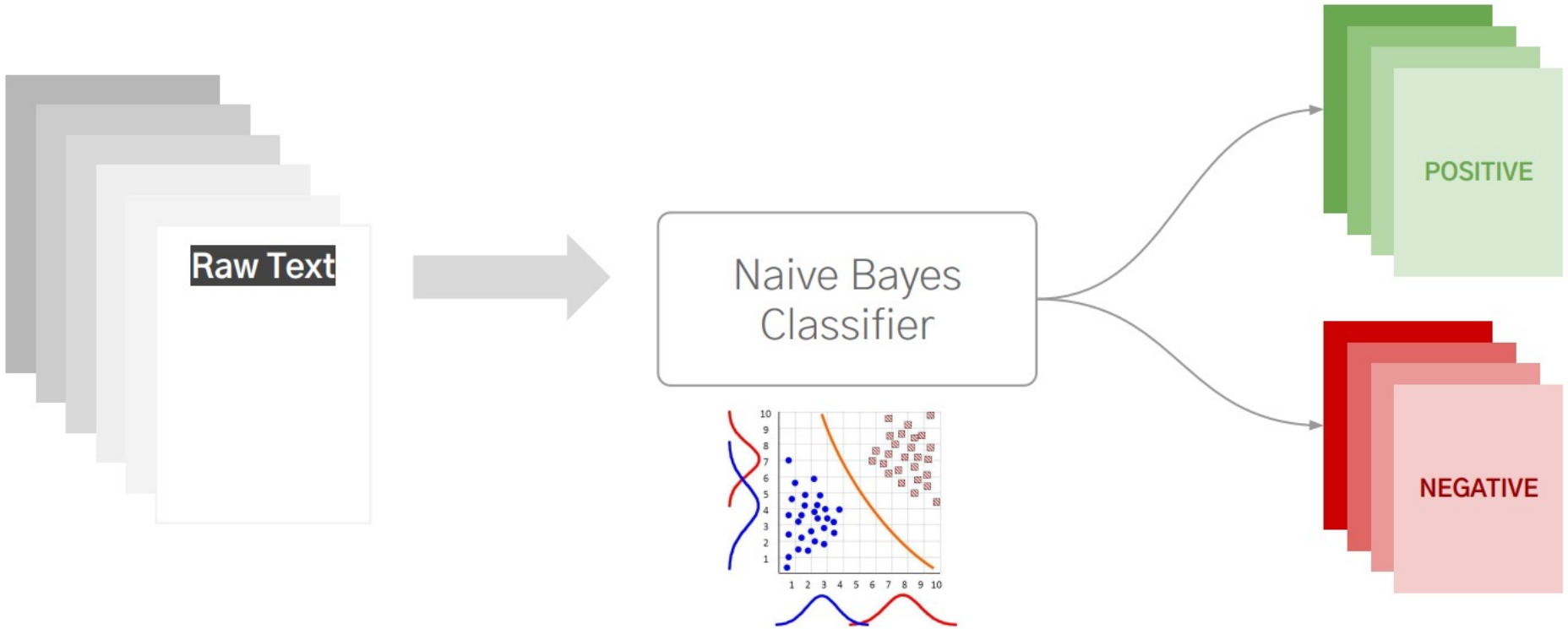
– 기계 학습분야에서, '나이브 베이즈 분류(Naïve Bayes Classification)는 특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기의 일종으로 1950 년대 이후 광범위하게 연구되고 있다.

– 통계 및 컴퓨터 과학 문헌에서 , 나이브 베이즈는 단순 베이즈, 독립 베이즈를 포함한 다양한 이름으로 알려져 있으며, 1960년대 초에 텍스트 검색 커뮤니티에 다른 이름으로 소개되기도 하였다.

– 나이브 베이즈 분류는 텍스트 분류에 사용됨으로써 문서를 여러 범주 (예: 스팸, 스포츠, 정치)중 하나로 판단하는 문제에 대한 대중적인 방법으로 남아있다. 또한, 자동 의료 진단 분야에서의 응용사례를 보면, 적절한 전처리를 하면 더 진보 된 방법들 (예: 서포트 벡터 머신 (Support Vector Machine))과도 충분한 경쟁력을 보임을 알 수 있다.

나이브베이즈 분류기 활용 감정분석

- 감정분석도 분류 문제의 하나로 볼 수 있음
- 따라서 분류모델을 활용하여 감정분석이 가능함. 대신 감정라벨이 부착된 학습용 데이터가 필요



베이즈 정리

- 사전 확률과 사후확률 사이의 관계를 조건부 확률을 이용해 계산하는 확률 이론

- 사전 확률(prior probability): 이미 알고 있는 사건이 발생할 확률

- 우도(likelihood probability): 이미 알고 있는 사건이 발생한다는 조건하에 다른 사건이 발생할 확률

- 사후 확률(posterior probability): 사전확률과 우도를 통해서 알게되는 조건부 확률

- 확률론과 통계학에서, 베이즈 정리(영어: Bayes' theorem)는 **두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 정리**다. 베이즈 확률론 해석에 따르면 베이즈 정리는 사전확률로부터 사후확률을 구할 수 있다.

- 확률 종류
  - 사전(prior) 확률 : $P(A)$
  - 우도(likelihood) 확률 : $P(B|A) = P(A \cap B) / P(A)$
  - 사후(posterior) 확률 :
    - $P(A|B) = P(A \cap B) / P(B)$
    - $P(A \cap B) = P(B|A) \times P(A)$
    - $P(B) = P(A)P(B|A) + P(A')P(B|A')$
    - $P(A|B) = P(A \cap B) / P(B) = [P(B|A) \times P(A)] / [P(A)P(B|A) + P(A')P(B|A')]$

Bayes' theorem

- 조건부 확률에 대한 수학적 정리

- 두 확률변수의 사전확률과 사후확률 사이의 관계를 나타내는 정리

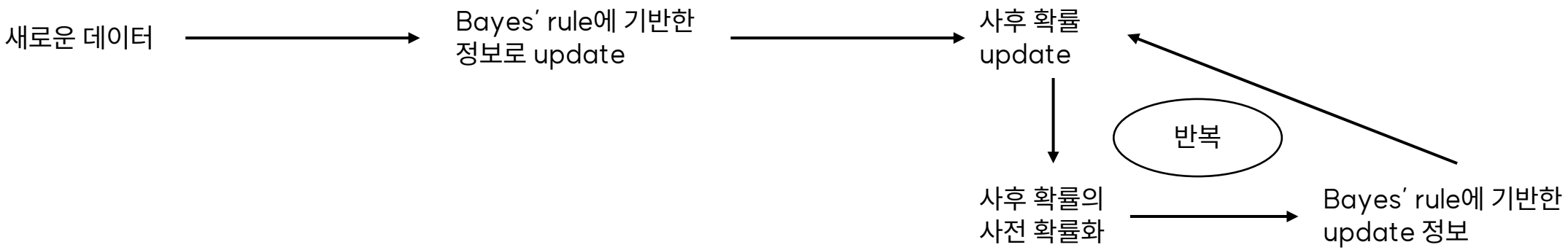
사전확률
(prior)

$$\underbrace{P(X|E)}_{\text{사후확률 (posterior)}} = \frac{P(X,E)}{P(E)} = \frac{P(E|X)P(X)}{P(E)}$$

$$= \frac{P(E|X)P(X)}{P(X|E)P(E)+P(X|\sim E)P(\sim E)}$$

Bayes' theorem

– 어떤 의사결정이나 확률을 구할 때, 계속 발전된 방향으로 업데이트 시켜나가기 때문에 머신러닝 분야나 인공지능 분야에서 베이즈 정리를 많이 사용



예제 : 유방암

- 김여사 유방조영술을 통해 유방암 검사를 받았는데, 검사결과가 양성(Positive)라고 검진되었음
- 유방암에 걸렸을 때, 유방조영술을 통해 양성(Positive)으로 나올 확률은 90%
- 유방암이 아니더라도 유방조영술이 양성일 확률은 7%
- 40-50대 여성이 유방암에 걸릴 확률은 0.8%
- 김여사가 유방암에 걸렸을 확률은?

- 경우
  - Y = [암(cancer), 정상(normal)] # 유방암
  - X = [양성(positive), 음성(negative)] # 유방조영술

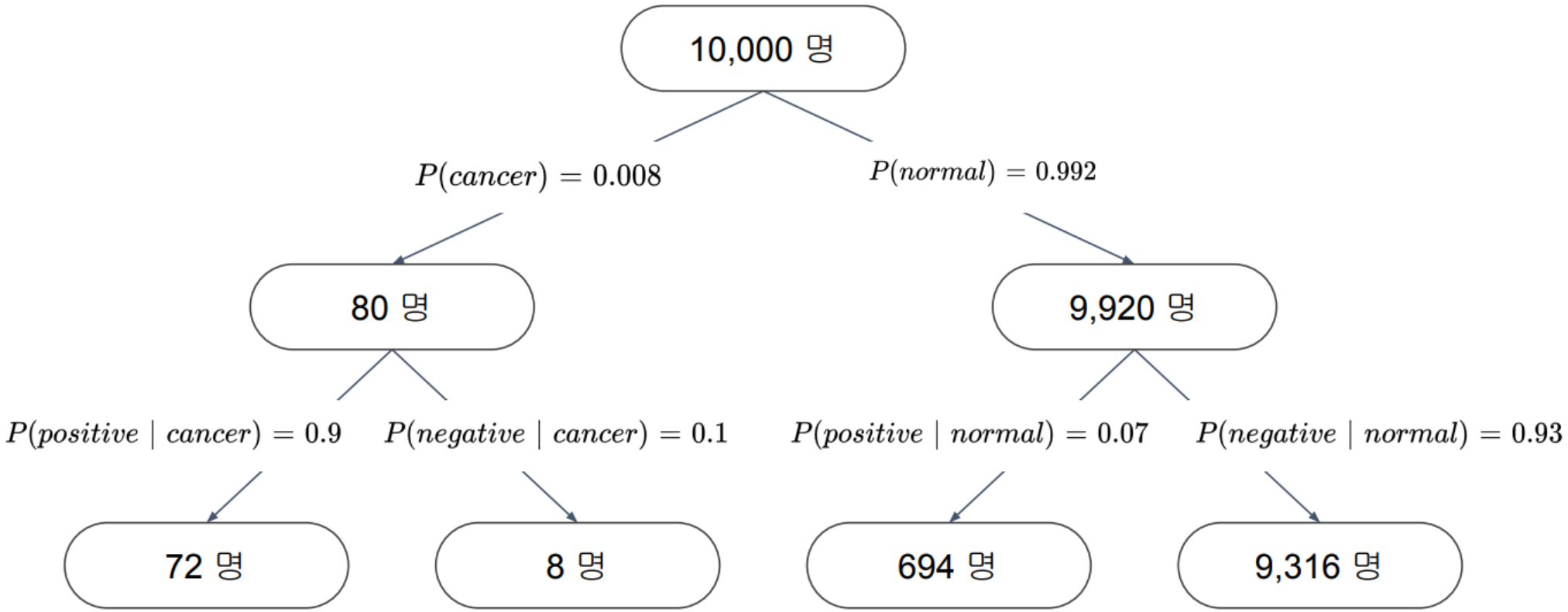
- 유방암에 걸릴 확률 (사전 확률)
  - P(cancer) = 0.008
  - P(normal) = 1 - 0.008 = 0.992

- 우도

$$P(positive \mid cancer) = \frac{P(cancer \cap positive)}{P(cancer)} = 0.9$$

$$P(positive \mid normal) = \frac{P(normal \cap positive)}{P(normal)} = 0.07$$

Exact Bayes Classifier



Exact Bayes Classifier

$$P(cancer \mid positive)$$

$$= \frac{P(cancer)P(positive|cancer)}{P(cancer)P(positive|cancer)+P(normal)P(positive|normal)}$$

$$= \frac{0.008 \cdot 0.9}{(0.008 \cdot 0.9)+(0.992 \cdot 0.07)}$$

$$= \frac{0.0072}{0.0072+0.0694}$$

$$= \frac{0.0072}{0.0766} = 0.0939$$

Naïve Bayes 정리

- 정확히 일치하는 데이터가 없어도 전체 데이터를 이용해 계산

When

$$X = \langle X_1, X_2, \cdots, X_n \rangle, \quad X_i : \text{discrete or continuous}, \quad Y : \text{discrete}$$

Naive Bayes classifier

$$P(Y = y_k | X_1 \ldots X_n) \overset{\text{Bayes}}{=} \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)}$$

Conditional Independence

$$\overset{\text{Naive Bayes}}{=} \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

예제 : 사기 재무보고 예측

- 예제(사후확률) : P_{nb}(사기|이전법전문제=yes, 회사규모=small)
- 사기 인 경우
  - 사전 확률 : P(사기) = 4/10 = 0.4
  - 우도(조건부 확률)
    - P(이전법적문제=yes | 사기) = ¾ = 0.75
    - P(회사규모=small | 사기) = ¼ = 0.25

- 정직 인 경우
  - 사전 확률 : P(정직) = 1- 0.4 = 0.6
  - 우도(조건부 확률)
    - P(이전법적문제=yes | 정직) = 1/6 = 0.17
    - P(회사규모=small | 정직) = 4/6 = 0.67

회사	이전 법적문제	회사규모	상태
1	yes	small	정직
2	no	small	정직
3	no	large	정직
4	no	large	정직
5	no	small	정직
6	no	small	정직
7	yes	small	사기
8	yes	large	사기
9	no	large	사기
10	yes	large	사기

예제 : 사기 재무보고 예측

$$P(\text{사기} \mid \text{이전법전문제} = yes,\ \text{회사규모} = small)$$

$$= \frac{P(Fraud)\cdot P(Yes|Fraud)\cdot P(Small|Fraud)}{P(Fraud)\cdot P(Yes|Fraud)\cdot P(Small|Fraud)+P(Nrmal)\cdot P(Yes|Nrmal)\cdot P(Small|Nrmal)}$$

$$= \frac{0.4\cdot 0.75\cdot 0.25}{(0.4\cdot 0.75\cdot 0.25)+(0.6\cdot 0.17\cdot 0.67)} = 0.53$$

	정직 (0.6)	사기 (0.4)	P(x|정직)	P(x|사기)
법적 (yes)	1	3	1/6 = 0.17	3/4 = 0.75
법적 (no)	5	1	5/6 = 0.83	1/4 = 0.25
규모 (small)	4	1	4/6 = 0.67	1/4 = 0.25
규묘 (large)	2	3	2/6 = 0.33	3/4 = 0.75

- Exact bayes classifier
  - P(사기|이전법적문제=yes, 회사규모=small) = ½ = 0.5

회사	이전 법적문제	회사규모	상태
1	yes	small	정직
2	no	small	정직
3	no	large	정직
4	no	large	정직
5	no	small	정직
6	no	small	정직
7	yes	small	사기
8	yes	large	사기
9	no	large	사기
10	yes	large	사기

예제 : 사기 재무보고 예측

- 예제(사후확률) :
  - P_{nb}(사기|이전법적문제=no, 회사규모=small) = ?
  - P_{nb}(정직|이전법적문제=no, 회사규모=small) = ?

	정직 (0.6)	사기 (0.4)	P(x|정직)	P(x|사기)
법적 (yes)	1	3	1/6 = 0.17	3/4 = 0.75
법적 (no)	5	1	5/6 = 0.83	1/4 = 0.25
규모 (small)	4	1	4/6 = 0.67	1/4 = 0.25
규묘 (large)	2	3	2/6 = 0.33	3/4 = 0.75

회사	이전 법적문제	회사규모	상태
1	yes	small	정직
2	no	small	정직
3	no	large	정직
4	no	large	정직
5	no	small	정직
6	no	small	정직
7	yes	small	사기
8	yes	large	사기
9	no	large	사기
10	yes	large	사기

예제 : 사기 재무보고 예측

- P_{nb}(사기|이전법적문제=no, 회사규모=small)

$$= \frac{\frac{4}{10} \cdot \frac{1}{4} \cdot \frac{1}{4}}{\left(\frac{4}{10} \cdot \frac{1}{4} \cdot \frac{1}{4}\right) + \left(\frac{6}{10} \cdot \frac{5}{6} \cdot \frac{4}{6}\right)} = 0.07$$

- P_{nb}(정직|이전법적문제=no, 회사규모=small) = ?

$$= \frac{\frac{6}{10} \cdot \frac{5}{6} \cdot \frac{4}{6}}{\left(\frac{6}{10} \cdot \frac{5}{6} \cdot \frac{4}{6}\right) + \left(\frac{4}{10} \cdot \frac{1}{4} \cdot \frac{1}{4}\right)} = 0.93$$

회사	이전 법적문제	회사규모	상태
1	yes	small	정직
2	no	small	정직
3	no	large	정직
4	no	large	정직
5	no	small	정직
6	no	small	정직
7	yes	small	사기
8	yes	large	사기
9	no	large	사기
10	yes	large	사기

Naïve Bayes Classifier 장단점

- 장점
  - 범주형 변수 처리
  - 단순형, 계산 효율성
  - 좋은 분류성능

- 단점
  - 많은 데이터 필요
  - 값이 0일 확률 처리 : Laplace smoothing

Naïve Bayes 개선 – Laplace Smoothing

- 입력 텍스트가 기존에 계산한 확률이 존재하지 않을경우 0으로 계산될 수 있음

- 예시 : w_n이라는 신규 단어가 입력되는 경우, 각 확률은 0으로 계산됨
  - P(정상 메일 | 입력 텍스트) = P(w1 | 정상 메일) * P(w2 | 정상 메일) * P(w3 | 정상 메일) * P(w_n | 정상 메일) * P(정상 메일)
  - P(스팸 메일 | 입력 텍스트) = P(w1 | 스팸 메일) * P(w2 | 스팸 메일) * P(w3 | 스팸 메일) * P(w_n | 정상 메일) * P(스팸 메일)

- 분자와 분모에 일정 상수 (k)를 더하여 신규 단어가 출현했을 때 0으로 계산되는 것을 방지

$$P(w_i \mid positive) = \frac{k + count(w_i, positive)}{2k + \sum_{w \in V}(w, positive)}$$

Naïve Bayes 개선 – Log 이용 언더플로우 방지

- 확률을 계산하고 확률간의 곱으로 연산이 이루어짐
- 1이하 값으로 이루어지는 값을 계속 해서 곱하면 소수점 이하로 계속 작아서 계산할 수 없는 범위 이하로 작아지는 것을 언더플로우(underflow)라고 함
- Log의 성질을 활용하여 곱셈을 덧셈으로 변환하여 underflow를 방지

$$\log A \cdot B = \log A + \log B$$

$$\prod_i P(word_i|Pos) = exp\left[\sum_i \{\log P(word_i|Pos)\}\right]$$

예제 : 스팸 필터링

	메일로부터 토큰화 및 정제된 단어들	분류
1	me free lottery	스팸
2	free get free you	스팸
3	you free scholarship	정상
4	free to contact me	정상
5	you won award	정상
6	you ticket lottery	스팸

- 위 표로 스팸/정상 메일을 학습 = 확률을 계산
- “free lottery”라는 토큰이 있는 메일이 스팸일 확률은?

예제 : 스팸 필터링

- 우리가 구하고자 하는 것 (목표)
  - P(Normal | Words) = 입력 텍스트가 있을 때 정상 메일일 확률
  - P(Spam | Words) = 입력 텍스트가 있을 때 스팸 메일일 확률

- 계산 방법

$$P(Normal \mid Words) = \frac{P(Words|Normal) \cdot P(Normal)}{P(Words|Normal) \cdot P(Normal) + P(Words|Spam) \cdot P(Spam)}$$

$$P(Spam \mid Words) = \frac{P(Words|Spam) \cdot P(Spam)}{P(Words|Normal) \cdot P(Normal) + P(Words|Spam) \cdot P(Spam)}$$

- 입력되는 각 단어의 조건부 확률의 곱으로 표현 가능

$$P(Words \mid Normal) \cdot P(Normal) = P(w1 \mid Normal) \cdot P(w2 \mid Normal) \cdot P(Normal)$$

$$P(Words \mid Spam) \cdot P(Spam) = P(w1 \mid Spam) \cdot P(w2 \mid Spam) \cdot P(Spam)$$

예제 : 스팸 필터링

	메일로부터 토큰화 및 정제된 단어들	분류
1	me free lottery	스팸
2	free get free you	스팸
3	you free scholarship	정상
4	free to contact me	정상
5	you won award	정상
6	you ticket lottery	스팸

$$P(Normal \mid free, lottery)$$

$$= \frac{P(free|Normal) \cdot P(lottery|Normal) \cdot P(Normal)}{P(free|Normal) \cdot P(lottery|Normal) \cdot P(Normal) + P(free|Spam) \cdot P(lottery|Spam) \cdot P(Spam)}$$

$$P(Spam \mid free, lottery)$$

$$= \frac{P(free|Spam) \cdot P(lottery|Spam) \cdot P(Spam)}{P(free|Spam) \cdot P(lottery|Spam) \cdot P(Spam) + P(free|Normal) \cdot P(lottery|Normal) \cdot P(Normal)}$$

예제 : 스팸 필터링

	메일로부터 토큰화 및 정제된 단어들	분류
1	me free lottery	스팸
2	free get free you	스팸
3	you free scholarship	정상
4	free to contact me	정상
5	you won award	정상
6	you ticket lottery	스팸

Laplace Smoothing 적용

tokens | 분류	spam	normal	P(w|spam)	P(w|normal)
award	0	1	4.55%	13.64%
contact	0	1	4.55%	13.64%
free	3	2	31.82%	22.73%
get	1	0	13.64%	4.55%
lottery	2	0	22.73%	4.55%
me	1	1	13.64%	13.64%
scholarship	0	1	4.55%	13.64%
ticket	1	0	13.64%	4.55%
to	0	1	4.55%	13.64%
won	0	1	4.55%	13.64%
you	2	2	22.73%	22.73%
합계	10	10		

$$P(free \mid spam) = \frac{k+free}{2 \cdot k+spam} = \frac{0.5+3}{2 \cdot 0.5+10} = 31.82\,\%$$

$$P(lottery \mid spam) = \frac{k+lottery}{2 \cdot k+spam} = \frac{0.5+2}{2 \cdot 0.5+10} = 22.73\,\%$$

예제 : 스팸 필터링

	메일로부터 토큰화 및 정제된 단어들	분류
1	me free lottery	스팸
2	free get free you	스팸
3	you free scholarship	정상
4	free to contact me	정상
5	you won award	정상
6	you ticket lottery	스팸

	Log(P(w|spam))	Log(P(w|normal))
award	-3.0910	-1.9924
contact	-3.0910	-1.9924
free	-1.1451	-1.4816
get	-1.9924	-3.0910
lottery	-1.4816	-3.0910
me	-1.9924	-1.9924
scholarship	-3.0910	-1.9924
ticket	-1.9924	-3.0910
to	-3.0910	-1.9924
won	-3.0910	-1.9924
you	-1.9924	-1.4826

예제 : 스팸 필터링

	메일로부터 토큰화 및 정제된 단어들	분류
1	me free lottery	스팸
2	free get free you	스팸
3	you free scholarship	정상
4	free to contact me	정상
5	you won award	정상
6	you ticket lottery	스팸

- P(free | Normal) * P(lottery | Normal) * P(Normal)
  = Exp(Log(P(free | Normal) * P(lottery | Normal) * P(Normal)))
  = Exp(Log(P(free | Normal)) + Log(P(lottery | Normal)) + Log(P(Normal)))
  = Exp((−1.4816) + (−3.0910) + (−0.6931)) = Exp(−5.2658)
  = 0.0052 = 0.52%

- P(free | Spam) * P(lottery | Spam) * P(Spam)
  = Exp(Log(P(free | Spam) * P(lottery | Spam) * P(Spam)))
  = Exp(Log(P(free | Spam)) + Log(P(lottery | Spam)) + Log(P(Spam)))
  = Exp((−1.1451) + (−1.4816) + (−0.6931) = Exp(−3.3199)
  = 0.0362 = 3.62%

Log 이용 언더플로우 방지

예제 : 스팸 필터링

$$P(free \mid Spam) \cdot P(lottery \mid Spam) \cdot P(Spam)$$

$$= Exp(\log(P(free \mid Spam) \cdot P(lottery \mid Spam) \cdot P(Spam)))$$

$$= Exp(\log(P(free \mid Spam)) + \log(P(lottery \mid Spam)) + \log(P(Spam)))$$

$$= Exp((-1.4816) + (-3.0910) + (-0.6931)) = Exp(-5.2658)$$

$$= 0.0052 = 0.52\,\%$$

$$P(free \mid normal) \cdot P(lottery \mid normal) \cdot P(normal)$$

$$= Exp(\log(P(free \mid normal) \cdot P(lottery \mid normal) \cdot P(normal)))$$

$$= Exp(\log(P(free \mid normal)) + \log(P(lottery \mid normal)) + \log(P(normal)))$$

$$= Exp((-1.1451) + (-1.4816) + (-0.6931) = Exp(-3.3199)$$

$$= 0.0362 = 3.62\,\%$$

예제 : 스팸 필터링

	메일로부터 토큰화 및 정제된 단어들	분류
1	me free lottery	스팸
2	free get free you	스팸
3	you free scholarship	정상
4	free to contact me	정상
5	you won award	정상
6	you ticket lottery	스팸

- free, lottery가 포함된 메일이 스팸일 확률

$$= \frac{3.62\%}{0.52\%+3.62\%} = 87.5\ \%$$

- free, lottery가 포함된 메일이 정상일 확률

$$= \frac{0.52\%}{0.52\%+3.62\%} = 12.5\ \%$$

Log 이용 언더플로우 방지

예제 : 스팸 필터링

- 토큰별 조건부 확률 계산

tokens | 분류	spam	normal	P(w|spam)	P(w|normal)	Log(P(w|spam))	Log(P(w|normal))
award	0	1	4.55%	13.64%	-3.0910	-1.9924
contact	0	1	4.55%	13.64%	-3.0910	-1.9924
free	3	2	31.82%	22.73%	-1.1451	-1.4816
get	1	0	13.64%	4.55%	-1.9924	-3.0910
lottery	2	0	22.73%	4.55%	-1.4816	-3.0910
me	1	1	13.64%	13.64%	-1.9924	-1.9924
scholarship	0	1	4.55%	13.64%	-3.0910	-1.9924
ticket	1	0	13.64%	4.55%	-1.9924	-3.0910
to	0	1	4.55%	13.64%	-3.0910	-1.9924
won	0	1	4.55%	13.64%	-3.0910	-1.9924
you	2	2	22.73%	22.73%	-1.9924	-1.4826
합계	10	10				