R 교육 세미나 2

ToBig's 7기 최희정

Naive Bayes Classification

Navie Bayes 분류기

Contents

Naive Bayes Classifier

특성들 사이의 독립을 가정하는 베이즈 정리를 적용한 확률 분류기로
주로 스팸 필터나 키워드 검색을 활용한 문서 분류에 사용되는 지도 학습 분류기

Unit 01 | Naive Bayes Intro – 조건부확률, MLE&MAP

투빅스에서 맥주를 마시는 술자리가 있었다. 그 술자리에 누가 있었을까?

(단, 희정&재석은 서로를 굉장히 싫어해서 같이 술자리를 가지지 않음)

0 재석

1 희정

어떤 기준으로 분류할까?

1. 재석일 때 맥주를 마실 확률 vs 희정일 때 맥주를 마실 확률
-> P(맥주 | 재석) vs P(맥주 | 희정)

2. 맥주를 마실 때 재석일 확률 vs 맥주를 마실 때 희정일 확률
-> P(재석 | 맥주) vs P(희정 | 맥주)

Unit 01 | Naive Bayes Intro – 조건부확률, MLE&MAP

1. 조건부확률(Conditional Probability)

: 어떤 사건이 일어난 **조건하에서** 다른 사건이 일어날 확률

사건 A가 발생했을 때, 사건 B가 발생할 확률

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

사건 A와 사건 B가 동시에 발생할 확률

$$\boxed{P(A \cap B) = P(A|B) \times P(B)}$$

Unit 01 | Naive Bayes Intro – 조건부확률, MLE&MAP

2. 최대우도추정법(Maximum Likelihood Estimation)

: 데이터(x)가 **관찰될 확률을 최대화**시키는 모수 θ값(class)을 찾는 방법

-> $$\hat{\theta}_{ML} := \arg\max_{\theta} f(x|\theta)$$

ex) 데이터 = 맥주

-> P(맥주 | 재석) vs P(맥주 | 희정)

Unit 01 | Naive Bayes Intro – 조건부확률, MLE&MAP

3. 최대사후확률(Maximum A Posteriori)

: 데이터(x)의 **사후확률을 최대화**시키는 모수 θ값(class)을 찾는 방법

–> $$Posterior = P(\theta|x) = \frac{p(\theta \cap x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{Likelihood \times Prior}{Evidence}$$

–> $$\hat{\theta}_{\mathrm{MAP}} := \underset{\theta}{\arg\max}\, f(x|\theta)p(\theta)$$ ($\because$ 모든 데이터에 대해 $p(x)$ 동일)

ex) 데이터 = 맥주

–> P(재석 | 맥주) vs P(희정 | 맥주)

–> P(맥주 | 재석) x P(재석) vs P(맥주 | 희정) x P(희정)

Unit 01 | Naive Bayes Intro - 조건부확률, MLE&MAP

0 재석

0.2

0.3

0.5

1 희정

0.6

0.2

0.2

Unit 01 | Naive Bayes Intro - 조건부확률, MLE&MAP

0 재석

0.2	0.3	0.5

1 희정

0.6	0.2	0.2

Unit 01 | Naive Bayes Intro – 조건부확률, MLE&MAP

0 재석

1 희정

0.6 | 0.2 | 0.2

MLE : 맥주가 관찰될 확률 최대화

-> P(맥주 | 재석) < P(맥주 | 희정)
-> 맥주가 있는 술자리에는 희정이가 있다!

Unit 01 | Naive Bayes Intro – 조건부확률, MLE&MAP

이 때, 투빅스 술자리의 80%에는 재석이가 참여하고
20%에는 희정이가 참여한다는 사전정보를 획득했다.

0 재석 (0.8)

1 희정 (0.2)

Unit 01 | Naive Bayes Intro - 조건부확률, MLE&MAP

0 재석

0.2x0.8
= 0.16

0.3x0.8
= 0.24

+

0.5x0.8
= 0.4

1 희정

0.6x0.2
= 0.12

0.2x0.2
= 0.04

+

0.2x0.2
= 0.04

Unit 01 | Naive Bayes Intro – 조건부확률, MLE&MAP

0 재석

0.2x0.8
= 0.16

0.3x0.8
= 0.24

0.5x0.8
= 0.4

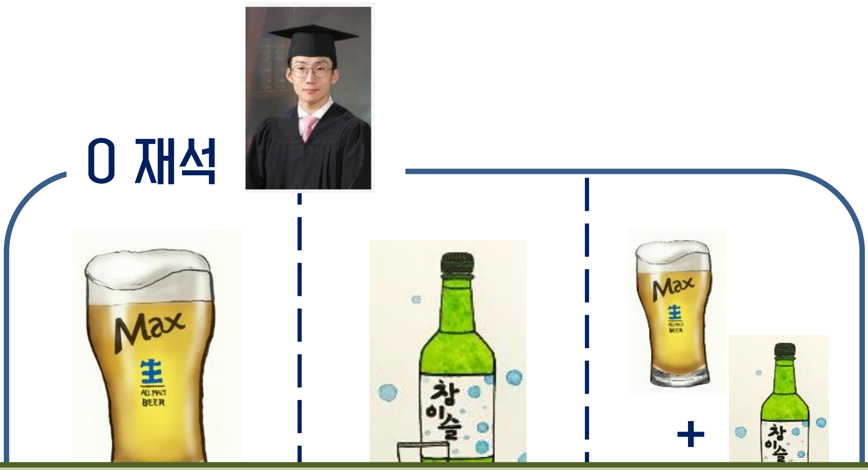
1 희정

0.6x0.2
= 0.12

0.2x0.2
= 0.04

0.2x0.2
= 0.04

Unit 01 | Naive Bayes Intro – 조건부확률, MLE&MAP



MAP : 맥주의 사후확률 최대화

–> P(맥주 | 재석) x P(재석) > P(맥주 | 희정) x P(희정)

–> 맥주가 있는 술자리에는 재석이가 있다!

Unit 02 | Naive Bayes Classifier

Naive Bayes Classifier

: 데이터(x)의 **사후확률을 최대화**하는 모수 θ값(class)로 분류하는 분류기

$$\rightarrow \quad f^*(x) = \underset{\theta}{\operatorname{argmax}}\, p(\theta|x) = \underset{\theta}{\operatorname{argmax}}\, p(x|\theta)p(\theta)$$

Unit 02 | Naive Bayes Classifier

술자리	맥주(X1)	소주(X2)	소맥(X3)	참석자(θ)
1	1	0	0	재석
2	1	1	0	희정
3	0	1	0	재석
⋮	⋮	⋮	⋮	⋮

$$\rightarrow \quad f^*(x) = \underset{\theta}{\operatorname{argmax}}\, p(x|\theta)p(\theta)$$
$$= \underset{\theta}{\operatorname{argmax}}\, \boldsymbol{p(x_1, x_2, x_3|\theta)}p(\theta)$$

Unit 02 | Naive Bayes Classifier

데이터가 n개의 feature를 가진다면,

$$f^*(x) = \underset{\theta}{\mathrm{argmax}}\, \boxed{p(x_1, \cdots, x_n | \theta)}\, p(\theta)$$

⇓

필요한 parameter의 개수 = (2^n – 1) x 2

Unit 02 | Naive Bayes Classifier

데이터가 n개의 feature를 가진다면,

$$f^*(x) = \underset{\theta}{\operatorname{argmax}} \boxed{\boldsymbol{p(x_1, \cdots, x_n | \theta)}}\, p(\theta)$$

필요한 parameter의 개수 = (2^n – 1) x 2

한 feature에서 나올 수 있는 경우의 수

Unit 02 | Naive Bayes Classifier

데이터가 n개의 feature를 가진다면,

$$f^*(x) = \underset{\theta}{\mathrm{argmax}}\, \boxed{\boldsymbol{p(x_1, \cdots, x_n|\theta)}}\, p(\theta)$$

필요한 parameter의 개수 = (2^n – 1) x 2

feature의 개수

데이터가 n개의 feature를 가진다면,

$$f^*(x) = \underset{\theta}{\mathrm{argmax}}\, \boxed{p(x_1, \cdots, x_n|\theta)}\, p(\theta)$$

$$\Downarrow$$

필요한 parameter의 개수 = (2^n − 1) x 2

$\downarrow$

나머지 parameter를 이용해
추정가능한 parameter의 수

Unit 02 | Naive Bayes Classifier

데이터가 n개의 feature를 가진다면,

$$f^*(x) = \underset{\theta}{\mathrm{argmax}}\ \boxed{\boldsymbol{p(x_1, \cdots, x_n|\theta)}}\ p(\theta)$$

필요한 parameter의 개수 = (2^n – 1) x 2

Class의 개수

Unit 02 | Naive Bayes Classifier

데이터가 n개의 feature를 가진다면,

$$f^*(x) = \underset{\theta}{\operatorname{argmax}}\ \boxed{\boldsymbol{p(x_1, \cdots, x_n|\theta)}}\ p(\theta)$$

↓

필요한 parameter의 개수 = (2^n - 1) x 2

-> feature가 늘어남에 따라 필요한 parameter의 개수 급격히 증가

-> 그렇다면 이러한 문제를 어떻게 해결할 수 있을까?

Unit 03 | Naive Bayes Assumption

$$\boldsymbol{p(x_1, x_2, \cdots, x_n | \theta)} = p(x_1 | \theta) p(x_2, \cdots, x_n | \theta, x_1)$$

$$= p(x_1 | \theta) p(x_2 | \theta, x_1) p(x_3, \cdots, x_n | \theta, x_1, x_2)$$

$$= \cdots$$

$$= p(x_1 | \theta) p(x_2 | \theta, x_1) \cdots p(x_n | \theta, x_1, x_2, \cdots, x_{n-1})$$

Unit 03 | Naive Bayes Assumption

만약 조건부확률을 이렇게 표현할 수 있다면?

$$\boldsymbol{p(x_1, x_2, \cdots, x_n|\theta)} = p(x_1|\theta)p(x_2, \cdots, x_n|\theta, x_1)$$

$$= p(x_1|\theta)p(x_2|\theta)p(x_3, \cdots, x_n|\theta)$$

$$= \cdots$$

$$= \boxed{\boldsymbol{p(x_1|\theta)p(x_2|\theta)\cdots p(x_n|\theta)}}$$

필요한 parameter의 개수 = (2-1) x n x 2

Unit 03 | Naive Bayes Assumption

조건부독립(Confidential Independence)

: 사건 a와 b가 조건 c에 대해 조건부독립이다.

= 사건 c가 주어졌을 때, 사건 a와 b가 서로 독립이다.

$$p(a|b,c) = p(a|c)$$

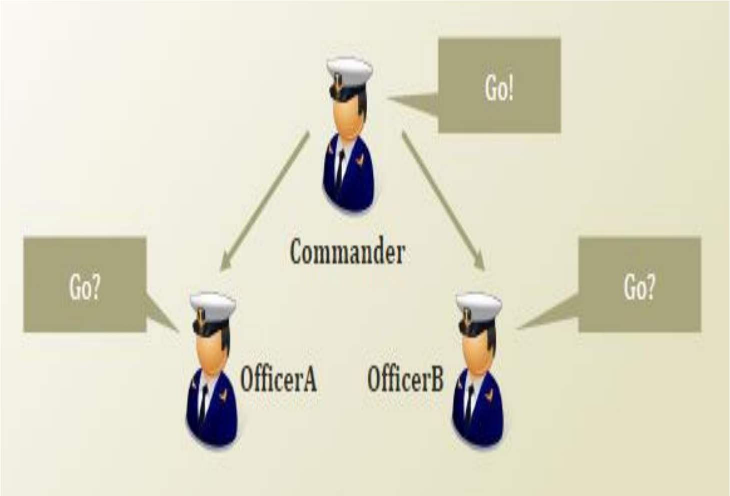
$$p(a,b|c) = p(a|b,c)p(b|c) = p(a|c)p(b|c)$$

$$\Rightarrow a \perp b \mid c$$

Unit 03 | Naive Bayes Assumption

조건부독립(Confidential Independence)

ex) Commander가 "Go!"라고 말했을 때, officer A와 officer B가 Go할 확률은 조건부 독립인가?



P(officer A=Go | officer B=Go, Commander=Go!)

= P(officer A=Go | Commander=Go!)

-> Commander가 "Go!"라고 말했을 때, officer A와 officer B가 Go할 확률은 **조건부 독립**이다.

Unit 03 | Naive Bayes Assumption

Naive Bayes Assumption

: 모든 feature $x_1, x_2, \cdots, x_n$은 class에 대해 **조건부독립**이다.

-> $$p(x_1, x_2, \cdots, x_n|\theta) = p(x_1|\theta)p(x_2|\theta)\cdots p(x_n|\theta)$$

-> 필요한 parameter의 개수 = (2^n – 1) x 2

⬇

(2-1)xnx2

Unit 04 | Naive Bayes Algorithm

Naive Bayes Classifier

1. feature 간 **조건부독립** 가정

2. 주어진 feature에 대한 **사후확률** 계산

3. Naive Classifier (사후확률을 **최대화**하는 Class로 분류)

Unit 04 | Naive Bayes Algorithm

Movie	word	class
1	fun, couple, love, love	comedy
2	fast, furious, shoot	action
3	couple, fly, fast, fun, fun	comedy
4	furious, shoot, shoot, fun	action
5	fly, fast, shoot, love	action

-> 이런 데이터로 Naive Bayes Classifier를 학습시켰을 때,

"fun, furious, fast"를 포함하는 새로운 데이터는 어떤 class로 분류될까?

Unit 04 | Naive Bayes Algorithm

Movie	word	class
1	fun, couple, love, love	comedy
2	fast, furious, shoot	action
3	couple, fly, fast, fun, fun	comedy
4	furious, shoot, shoot, fun	action
5	fly, fast, shoot, love	action

〈 사후확률 비교 〉

$$p(comedy|fun, furious, fast) = p(fun, furious, fast|comedy) \times p(comedy)$$

vs $$p(action|fun, furious, fast) = p(fun, furious, fast|action) \times p(action)$$

Unit 04 | Naive Bayes Algorithm

Movie	word	class
1	fun, couple, love, love	comedy
2	fast, furious, shoot	action
3	couple, fly, fast, fun, fun	comedy
4	furious, shoot, shoot, fun	action
5	fly, fast, shoot, love	action

	fun	couple	love	fast	furious	shoot	fly	
comedy	3	2	2	1	0	0	1	9
action	1	0	1	2	2	4	1	11

Unit 04 | Naive Bayes Algorithm

	fun	couple	love	fast	furious	shoot	fly	
comedy	3	2	2	1	0	0	1	9
action	1	0	1	2	2	4	1	11

1. $p(comedy|fun, furious, fast) = p(fun, furious, fast|comedy) \times p(comedy)$

→ $p(comedy|fun, furious, fast) = p(fun|comedy) \times p(furious|comedy) \times p(fast|comedy) \times p(comedy)$

$$= \frac{3}{9} \times \frac{0}{9} \times \frac{1}{9} \times \frac{2}{5} = 0$$

Unit 04 | Naive Bayes Algorithm

	fun	couple	love	fast	furious	shoot	fly	
comedy	3	2	2	1	0	0	1	9
action	1	0	1	2	2	4	1	11

2. $\boldsymbol{p(action|fun, furious, fast) = p(fun, furious, fast|action) \times p(action)}$

→ $p(action|fun, furious, fast) = p(fun|action) \times p(furious|action) \times p(fast|action) \times p(action)$

$$= \frac{1}{11} \times \frac{2}{11} \times \frac{2}{11} \times \frac{3}{5} = 0.0018$$

Unit 04 | Naive Bayes Algorithm

	fun	couple	love	fast	furious	shoot	fly	
comedy	3	2	2	1	0	0	1	9
action	1	0	1	2	2	4	1	11

$$p(comedy|fun, furious, fast) < p(action|fun, furious, fast)$$

-> "fun, furious, fast"를 포함하는 새로운 데이터는 action으로 분류된다.

Unit 05 | Naive Bayes 문제점 및 보완

Naive Bayes 문제점

1 빈도 수를 이용해서 확률을 계산하는데 새롭게 입력된 데이터가 학습 문서에 없는 단어를 가지고 있는 경우 확률이 0이 되어 버리는 문제

2 확률은 항상 1보다 작기 때문에 입력 벡터를 구성하는 요소가 많으면, 조건부 확률의 값이 너무 작아져 값의 비교가 어려운 underflow현상이 발생하는 문제

Unit 05 | Naive Bayes 문제점 및 보완

Naive Bayes 문제점 및 보완

1 빈도 수를 이용해서 확률을 계산하는데 새롭게 입력된 데이터가 학습 문서에 없는 단어를 가지고 있는 경우 확률이 0이 되어 버리는 문제

→ Laplace Smoothing

2 확률은 항상 1보다 작기 때문에 입력 벡터를 구성하는 요소가 많으면, 조건부 확률의 값이 너무 작아져 값의 비교가 어려운 underflow현상이 발생하는 문제

→ Log 변환

Unit 05 | Naive Bayes 문제점 및 보완

1. Laplace Smoothing

: 단어 빈도에 모두 **+1**을 적용해 새로운 단어의 조건부확률이 0이 되는 것을 방지
(단, 조건부확률의 분모는 **+feature의 개수** 적용)

〈 기존 조건부확률의 문제점 〉

ex) 영화 class분류

-> $p(comedy|fun, furious, fast) = p(fun|comedy) \times p(furious|comedy) \times p(fast|comedy) \times p(comedy)$

$$= \frac{3}{9} \times \frac{0}{9} \times \frac{1}{9} \times \frac{2}{5} = 0$$

-> $p(action|fun, furious, fast) = p(fun|action) \times p(furious|action) \times p(fast|action) \times p(action)$

$$= \frac{1}{11} \times \frac{2}{11} \times \frac{2}{11} \times \frac{3}{5} = 0.0018$$

Unit 05 | Naive Bayes 문제점 및 보완

1. Laplace Smoothing

〈 기존 조건부확률의 문제점 보완 〉

–〉 $p(comedy|fun, furious, fast) = p(fun|comedy) \times p(furious|comedy) \times p(fast|comedy) \times p(comedy)$

$$= \frac{3+1}{9+7} \times \frac{0+1}{9+7} \times \frac{1+1}{9+7} \times \frac{2}{5} = 0.004688$$

–〉 $p(action|fun, furious, fast) = p(fun|action) \times p(furious|action) \times p(fast|action) \times p(action)$

$$= \frac{1+1}{11+7} \times \frac{2+1}{11+7} \times \frac{2+1}{11+7} \times \frac{3}{5} = 0.0005487$$

$\therefore p(comedy|fun, furious, fast) > p(action|fun, furious, fast)$

–〉 "fun, furious, fast"를 포함하는 새로운 데이터는 comedy로 분류된다.

Unit 05 | Naive Bayes 문제점 및 보완

2. Log 변환

: 사후확률에 **로그**를 취해 feature 개수가 많을 때 사후확률 값이 0이 되는 것을 방지

〈 기존 사후확률의 문제점 〉

$$p(\theta|x_1, x_2, \cdots, x_n) = p(\theta)\, p(x_1, x_2, \cdots, x_n|\theta)$$

$$= p(\theta)\, p(x_1|\theta) p(x_2|\theta) \cdots p(x_n|\theta)$$

$$= p(\theta) \prod_{i=1}^{n} p(x_i|\theta)$$

-> $\lim_{n\to\infty} p(\theta) \prod_{i=1}^{n} p(x_i|\theta) = 0$ $\quad (\because p(x_i|\theta) < 1)$

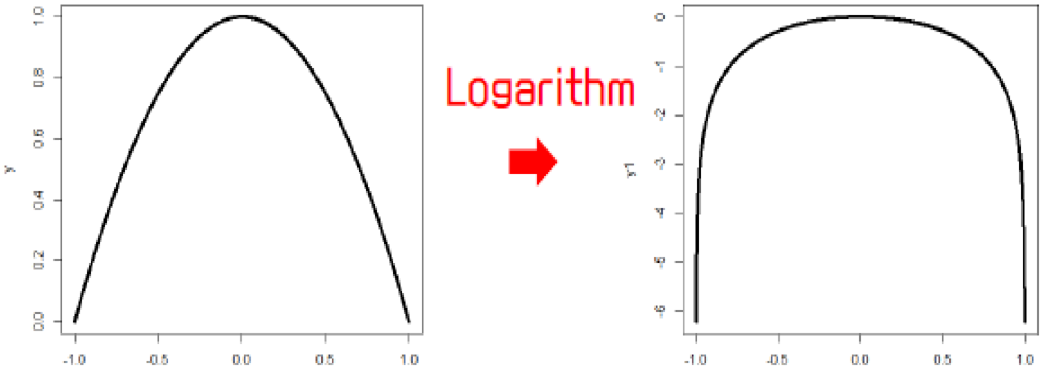
Unit 05 | Naive Bayes 문제점 및 보완

2. Log 변환

〈 기존 사후확률의 문제점 보완 〉

사후확률 = $p(\theta)\prod_{i=1}^{n} p(x_i\,|\theta)$

−〉 **로그사후확률** = $\ln \boldsymbol{p}(\boldsymbol{\theta})\prod_{i=1}^{n} \boldsymbol{p}(\boldsymbol{x_i}\,|\boldsymbol{\theta}) = \boldsymbol{p}(\boldsymbol{\theta}) + \sum_{i=1}^{n} \ln \boldsymbol{p}(\boldsymbol{x_i}\,|\boldsymbol{\theta})$



Q &A

들어주셔서 감사합니다.