2 주 차 교 육

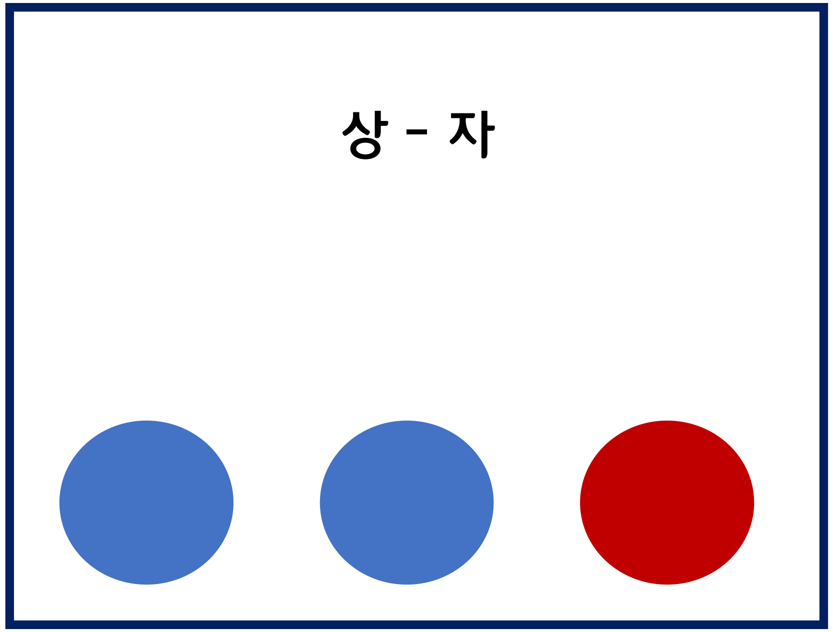
ToBig's 10기 황이은

Naive Bayes

*한줄 요약: 나이브하게(순진하게) 조건부 독립을 가정하여, 베이즈 정리와 비슷한 예측계산을 적용해 분류를 하겠다.

Contents

1.1 확률



파란 공을 잡을 확률: 2/3

빨간 공을 잡을 확률: 1/3

2/3 + 1/3 =1

1.2 조건부 확률

- $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A|B)P(B)$

- 곱셈 공식

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

A와 B가 독립시행일 경우 $P(A \cap B) = P(A)*P(B)$

어떤 사건이 일어난 조건하에서
다른 사건이 일어날 확률

1.3 베이즈 정리

*확률을 '지식 또는 믿음의 정도를 나타내는 양'으로 해석하는 확률론

*이전의 경험과 현재의 증거를 토대로 어떤 사건의 확률을 추론하는 과정

*사전확률과 사후확률 사이의 관계를 조건부확률을 이용해 계산

*실제 생활에서 사후확률만 알고 있는 경우가 많음.
이때 사전확률을 유추할 수 있음.

*결과를 관측한 뒤 원인을 추론할 수 있음

베이즈

1.3 베이즈 정리

Likelihood

Prior

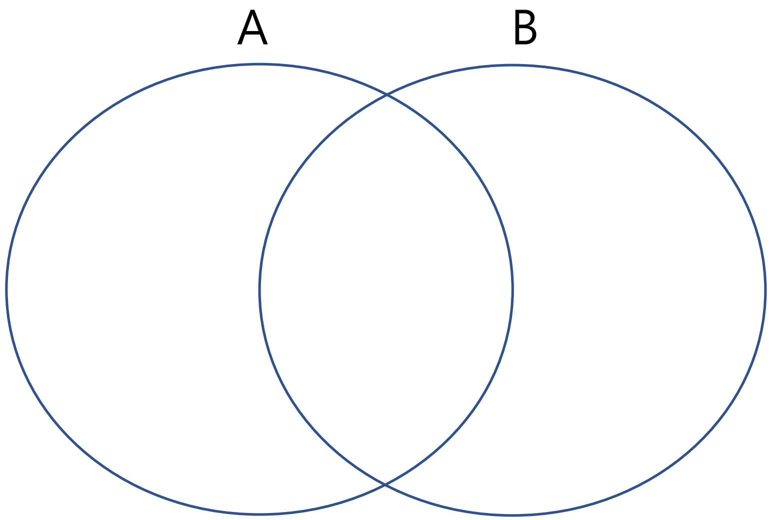
$$p(H \mid D) = \frac{p(D \mid H)\,p(H)}{p(D)}$$

Posterior

Evidence

베이즈 정리는 사전확률(prior)로부터 사후확률(posterior)을 구할 수 있다.
어떻게? 조건부 확률로!

1.3 베이즈 정리 – 증명



$P(A|B) = P(A \cap B) / P(B)$
$P(A \cap B) = P(A|B) * P(B)$
$P(B \cap A) = P(B|A) * P(A)$
$P(B \cap A) = P(A \cap B)$
$P(A \cap B) = P(B|A) * P(A) = P(A|B) * P(B)$
$P(B|A) * P(A) = P(A|B) * P(B)$

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

1.3 베이즈 정리 – 정리

조건부 확률

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

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

베이즈 정리

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

P(B)를 모를 때

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')}$$

Unit 02 | 예시

2. 예시

Naive + **Bayes**

형용사
순진한/순수한
천진난만한

베이즈 정리

>> 나이브하게 베이즈정리와 비슷한 계산을 하겠다!

=순진하게, 억지스럽게, 난 아무것도 모른다.

각 피쳐는 원래는 서로 독립이 아니야……. 계산 엄청 복잡해지네…….몰라몰라~ 조건부 독립을 가정해서 풀래~
(실제론 베이즈 정리와 상관없음! 베이즈정리와 비슷한 예측계산이 들어가다보니 이름을 그렇게 지었을 뿐.
중요한 건 조건부 독립)

Unit 02 | 예시

2. 예시 – 분류문제

d=관측치 개수
K=class 개수

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$$f^*(x) = argmax_{Y=y} \mathrm{P}(\mathrm{X} = \mathrm{x}|\mathrm{Y} = \mathrm{y})\mathrm{P}(\mathrm{Y} = \mathrm{y})$$

$\mathrm{P}(\mathrm{X} = \mathrm{x}|\mathrm{Y} = \mathrm{y})$
$=P(x1 = sunny, x2 = warm, x3 = normal,\ x4 = strong,\ x5 = warm, x6 = same|y = Yes)$
P(Y=y) = (y=Yes)

$\mathrm{P}(\mathrm{X} = \mathrm{x}|\mathrm{Y} = \mathrm{y})$= for all x,y -> $(2^d - 1)k$
P(Y=y) for all y -> k-1

Unit 02 | 예시

2. 예시 – 분류문제

문제점: 계산량이 많아짐

$P(X = x|Y = y) =$ for all x,y -> $(2^d - 1)k$

변수가 늘어날 수록 기하급수적으로 연산량이 증가함

어떻게 얻은 변수들인데……. d의 개수를 줄이는 건 피하고 싶어

->>해결책: 조건부 독립을 가정!!!!

3. 조건부 독립(Conditional independence)

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

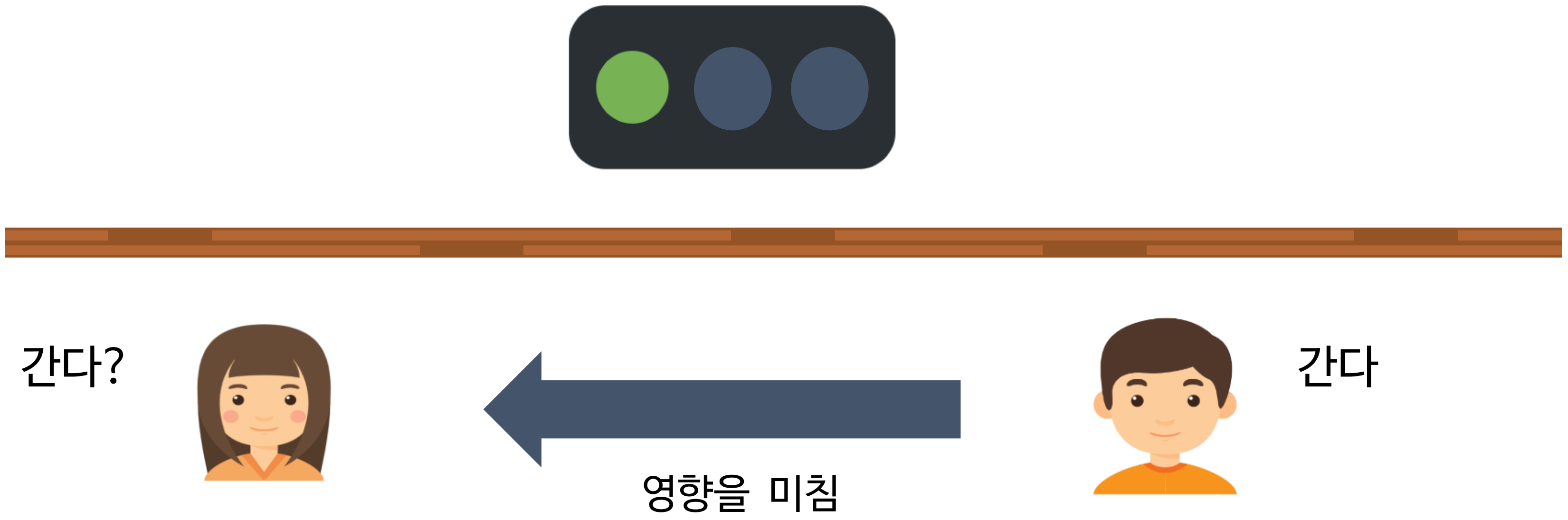
즉, b를 제외한 값이 좌변과 값이 동일하면, a와 b는 조건부 독립이다!

조건부 독립 예시:
P(천둥 |비, 번개) = P(천둥 | 번개)

비가 오는 사실이 additional한 정보를 주진 않음
따라서, 천둥 & 비는 독립이다!

3. 조건부 독립(Conditional independence)

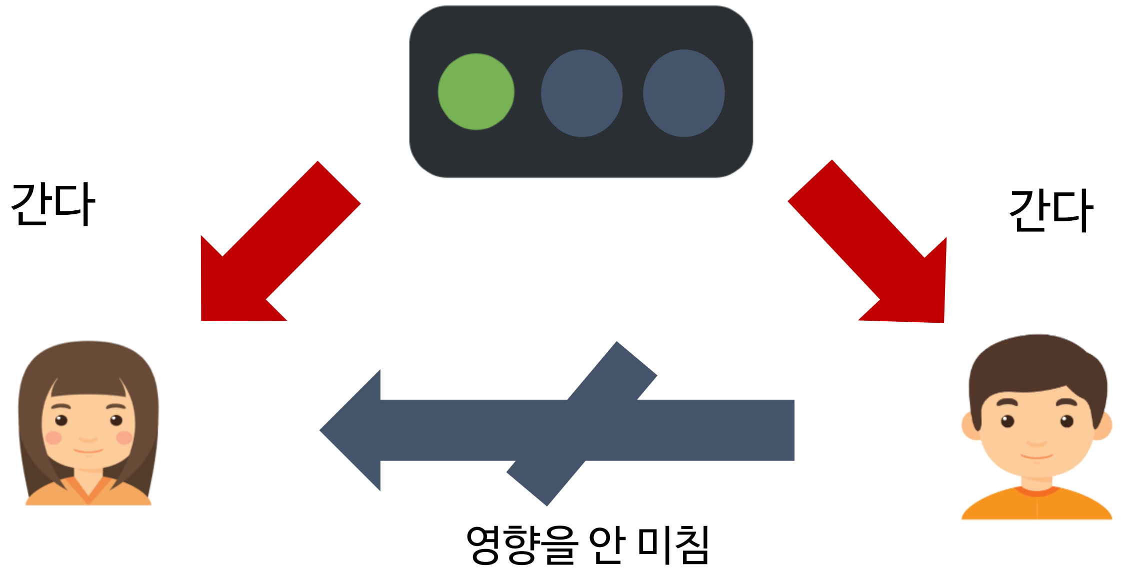
신호등 행인 예시



신호등(Y)이 없으면 나(x1)는 옆에 있는 행인(x2)에 영향을 받을 수 있음

3. 조건부 독립(Conditional independence)

신호등 행인 예시



반면, 신호등(Y)이 주어지면, 나(x1)와 옆에 있는 행인(x2)사이에 조건부 독립이 성립!!!
=옆에 사람이 가든 말든 상관 없음. 나는 간다.
여기에서 Y는 latent variable(내부 동인)이라고 함

Unit 03 | 조건부 독립

3. 조건부 독립(Conditional independence)을 적용 한 후

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

1. 알아야할 파라미터 개수의 변화: 길이2개 -> 1개로 변함

$P(X,Y) = P(X)(Y)$ $\mathrm{P}(\mathrm{X} = x_i|\mathrm{Y} = \mathrm{y})$ -> $(2 \quad - 1)dk$

2. 식의 변화
개별 피쳐 간의 곱셈으로 변함 (독립이니까)

$$f^*(x) = argmax_{Y=y}\mathrm{P}(\mathrm{X} = \mathrm{x}|\mathrm{Y} = \mathrm{y})\mathrm{P}(\mathrm{Y} = \mathrm{y})$$
$$\approx argmax_{Y=y}P(Y = y)\, \Pi_{1\le i\le d}\mathrm{P}(\mathrm{X} = x_i|\mathrm{Y} = \mathrm{y})$$

문제점 발생!
'밖에 나가서 논다' 이게 내부 동인이 되어서 -> X1, X2, X3에 영향을 미친다? ?
그게 가능하다면…….당신은 엑스멘….

4.1 나이브 베이즈

가정: 종속변수(Y)가 주어졌을 때, 입력 변수들이 모두 독립이다. (조건부 독립)

왜 나이브 베이즈라고 불르는 걸까?
결과가 주어졌을 때, 예측 변수 벡터의 정확한 조건부 확률은 각 조건부 확률의 곱으로(조건부 독립을 가정) 충분히 잘 추정 할 수 있다는 단순한 가정을 기초로 하기 때문!

수학적으로 나이브 베이즈 함수를 표현하면,

$$f^*(x) = argmax_{Y=y} \mathrm{P}(\mathrm{X} = \mathrm{x} | \mathrm{Y} = \mathrm{y}) \mathrm{P}(\mathrm{Y} = \mathrm{y})$$

$$\approx argmax_{Y=y} P(Y = y)\ \Pi_{1 \le i \le d} \mathrm{P}(\mathrm{X} = x_i | \mathrm{Y} = \mathrm{y})$$

4.2 나이브 베이즈 코드

```
data(HouseVotes84, package = "mlbench")
model <- naiveBayes(Class ~ ., data = HouseVotes84)
model
```

종속변수 데이터

```
Conditional probabilities:
            V1
Y                    n          y
  democrat   0.3953488 0.6046512
  republican 0.8121212 0.1878788

            V2
Y                    n          y
  democrat   0.4979079 0.5020921
  republican 0.4932432 0.5067568

            V3
Y                    n          y
  democrat   0.1115385 0.8884615
  republican 0.8658537 0.1341463
```

각 피쳐에 대한 조건부 확률을
계산해준다.

4.3 라플라스 스무딩

중간에 0인 값이 있을 때, 모든 확률을 0으로 만들어버리기 때문에 이를 방지하기 위한 것
카테고리형 자료일 때 쓰임(ex. 스팸 메일 분류)

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

분모, 분자에 적당한 수를 더해 0이 되는 것을 막자.
$|v|$는 학습데이터에서 나오는 유일한 단어의 수

4.3 라플라스 스무딩

중간에 0인 값이 있을 때, 모든 확률을 0으로 만들어버리기 때문에 이를 방지하기 위한 것
카테고리형 자료일 때 쓰임(ex. 스팸 메일 분류)

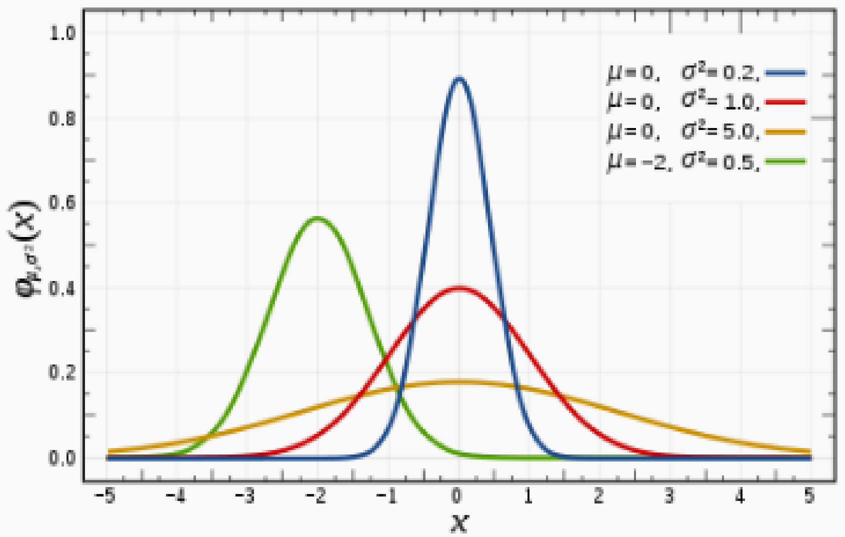
```
model <- naiveBayes(Class ~ ., data = HouseVotes84, laplace = 3)
```

5.1 부록 : 나이브 베이즈 vs 로지스틱 회귀

나이브베이즈 인풋을 연속형으로 받을 수 있다?

->가우시안 분포(정규분포)를 가정!
-> 자료로 부터 μ(평균), θ(분산)추정 가능

확률 밀도 함수



붉은 색은 표준정규분포

Unit 05 | 부록 – 나이브베이즈 vs 로지스틱 회귀

- Derivation from the naïve Bayes to the logistic regression
  - $P(Y)\prod_{1\le i\le d} P(X_i|Y) = \pi_k \prod_{1\le i\le d} \frac{1}{\sigma_k^i C} \exp(-\frac{1}{2}\left(\frac{X_i-\mu_k^i}{\sigma_k^i}\right)^2)$
- With naïve Bayes assumption
  - $P(Y=y|X) = \frac{P(X|Y=y)P(Y=y)}{P(X)} = \frac{P(X|Y=y)P(Y=y)}{P(X|Y=y)P(Y=y)+P(X|Y=n)P(Y=n)}$

$$= \frac{P(Y=y)\prod_{1\le i\le d} P(X_i|Y=y)}{P(Y=y)\prod_{1\le i\le d} P(X_i|Y=y) + P(Y=n)\prod_{1\le i\le d} P(X_i|Y=n)}$$

Unit 05 | 부록 – 나이브베이즈 vs 로지스틱 회귀

- With naïve Bayes assumption
  - $P(Y=y|X) = \frac{P(X|Y=y)P(Y=y)}{P(X)} = \frac{P(X|Y=y)P(Y=y)}{P(X|Y=y)P(Y=y)+P(X|Y=n)P(Y=n)}$

    $= \frac{P(Y=y)\prod_{1\le i\le d} P(X_i|Y=y)}{P(Y=y)\prod_{1\le i\le d} P(X_i|Y=y) + P(Y=n)\prod_{1\le i\le d} P(X_i|Y=n)}$

  - $P(Y=y|X) = \frac{\pi_1 \prod_{1\le i\le d}\frac{1}{\sigma_1^i C}\exp(-\frac{1}{2}\left(\frac{X_i-\mu_1^i}{\sigma_1^i}\right)^2)}{\pi_1 \prod_{1\le i\le d}\frac{1}{\sigma_1^i C}\exp(-\frac{1}{2}\left(\frac{X_i-\mu_1^i}{\sigma_1^i}\right)^2)+\pi_2 \prod_{1\le i\le d}\frac{1}{\sigma_2^i C}\exp(-\frac{1}{2}\left(\frac{X_i-\mu_2^i}{\sigma_2^i}\right)^2)}$

    $= \frac{1}{1+\frac{\pi_2 \prod_{1\le i\le d}\frac{1}{\sigma_2^i C}\exp(-\frac{1}{2}\left(\frac{X_i-\mu_2^i}{\sigma_2^i}\right)^2)}{\pi_1 \prod_{1\le i\le d}\frac{1}{\sigma_1^i C}\exp(-\frac{1}{2}\left(\frac{X_i-\mu_1^i}{\sigma_1^i}\right)^2)}}$

Unit 05 | 부록 – 나이브베이즈 vs 로지스틱 회귀

- Assuming the same variable of the two classes, $\sigma_2^i = \sigma_1^i$
  - $P(Y=y|X) = \dfrac{1}{1+\dfrac{\pi_2 \prod_{1\le i\le d} \frac{1}{\sigma_2^i C} \exp(-\frac{1}{2}\left(\frac{X_i-\mu_2^i}{\sigma_2^i}\right)^2)}{\pi_1 \prod_{1\le i\le d} \frac{1}{\sigma_1^i C} \exp(-\frac{1}{2}\left(\frac{X_i-\mu_1^i}{\sigma_1^i}\right)^2)}} = \dfrac{1}{1+\dfrac{\pi_2 \prod_{1\le i\le d} \exp(-\frac{1}{2}\left(\frac{X_i-\mu_2^i}{\sigma_2^i}\right)^2)}{\pi_1 \prod_{1\le i\le d} \exp(-\frac{1}{2}\left(\frac{X_i-\mu_1^i}{\sigma_1^i}\right)^2)}}$

$$= \frac{1}{1+\dfrac{\pi_2 \exp(-\sum_{1\le i\le d}\{\frac{1}{2}\left(\frac{X_i-\mu_2^i}{\sigma_2^i}\right)^2\})}{\pi_1 \exp(-\sum_{1\le i\le d}\{\frac{1}{2}\left(\frac{X_i-\mu_1^i}{\sigma_1^i}\right)^2\})}}$$

$$= \frac{1}{1+\dfrac{\exp(-\sum_{1\le i\le d}\left\{\frac{1}{2}\left(\frac{X_i-\mu_2^i}{\sigma_2^i}\right)^2\right\}+log\pi_2)}{\exp(-\sum_{1\le i\le d}\{\frac{1}{2}\left(\frac{X_i-\mu_1^i}{\sigma_1^i}\right)^2\}+log\pi_1)}}$$

Unit 05 | 부록 – 나이브베이즈 vs 로지스틱 회귀

- Assuming the same variable of the two classes, $\sigma_2^i = \sigma_1^i$
  - $P(Y=y|X) = \dfrac{1}{1+\exp(-\sum_{1\le i\le d}\left\{\frac{1}{2}\left(\frac{X_i-\mu_2^i}{\sigma_2^i}\right)^2\right\}+log\pi_2+\sum_{1\le i\le d}\left\{\frac{1}{2}\left(\frac{X_i-\mu_1^i}{\sigma_1^i}\right)^2\right\}-log\pi_1)}$
  - $= \dfrac{1}{1+\exp(-\frac{1}{2(\sigma_1^i)^2}\sum_{1\le i\le d}\left\{\left(X_i-\mu_1^i\right)^2-\left(X_i-\mu_2^i\right)^2\right\}+log\pi_2-log\pi_1)}$
  - $= \dfrac{1}{1+\exp(-\frac{1}{2(\sigma_1^i)^2}\sum_{1\le i\le d}\left\{2\left(\mu_2^i-\mu_1^i\right)X_i+{\mu_2^i}^2-{\mu_2^i}^2\right\}+log\pi_2-log\pi_1)}$

Unit 05 | 부록 – 나이브베이즈 vs 로지스틱 회귀

- Naïve Bayes classifier
  - $P(Y|X) = \frac{1}{1+\exp(-\frac{1}{2(\sigma_1^i)^2}\sum_{1\le i\le d}\left\{2(\mu_2^i-\mu_1^i)X_i+{\mu_2^i}^2-{\mu_2^i}^2\right\}+log\pi_2-log\pi_1)}$
  - Assumption to get this formula
    - Naïve Bayes assumption, Same variance assumption between classes
    - Gaussian distribution for P(X|Y)
    - Bernoulli distribution for P(Y)
  - # of parameters to estimate = 2*2*d+1=4d+1
    - With the different variances between classes
- Logistic Regression
  - $P(Y|X) = \frac{1}{1+e^{-\dot{\theta}^T x}}$
  - Assumption to get this formula
    - Fitting to the logistic function
  - # of parameters to estimate = d+1
- Who is the winner?
  - Really??? There is no winner... Why?

Together, modeling joint prob.

결론: 문제 상황을 보고 판단해서 써야한다!

5.2 부록 : 베이즈 에러

Training data를 완벽히 학습하였을 때(즉, 그 사건의 underlying density function 을 알고 있을 때),
그 Training data에 대해 가장 확률이 높은 Class Label을 선택하는 방법에서 발생하는
이론적 최소 오차

베이즈 에러는 모든 기계 학습에서 가능한 이론적 최소 오차로,
어떤 알고리즘이나 모델이 베이즈 에러에 비해 얼마나 오차가 큰지를 비교하는 경우가 많다.
이 베이즈 에러까지 오분류율을 감소시키는 함수를 optimize(최적화)함수 라고함

Unit 05 | 부록 – 나이브베이즈 vs 로지스틱 회귀

5.2 부록 : 베이즈 에러

*베이즈 에러가 0이 아닌 이상,
베이즈 에러까지의 오분류율 달성하게 해준다

*베이즈 에러는 데이터 양과 상관 없다.
(=데이터가 많아도 베이즈 에러는 줄일 수 없다.)

*실제 데이터에서는 못 구한다! ㅠㅠㅠㅠ
->시뮬레이션 환경을 통해 추정한다.

정리

베이즈 Therom: 사전확률과 사후확률 사이의 관계를 조건부확률을 이용해 계산

나이브 베이즈: 나이브하게 베이즈정리와 비슷한 예측 계산을 적용한것
=현실은 너무 복잡하니까 조건부 독립을 가정해, 조건부 확률을 풀겠다.

장점:
*입력 공간의 차원이 높을 때 유리
*텍스트에서 강점
*가우시안 나이브베이즈를 활용하면 input이 연속형일때도 사용가능!

단점:
*희귀한 확률이 나왔을 때 (라플라스 스무딩)
*조건부 독립이라는 가정 자체의 비현실적인 문제점

과제

1. 다음의 나이브베이즈 문제를 풀어보세요.

문서번호	주요단어	문서분류
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

1.1 입력문서가 {fast, furious, fun}만을 주요단어로 가질 때, 이 문서는 얼마의 확률로 어떤 문서로 분류되는가?
1.2 어떠한 문제점이 있고, 이를 해결하기 위해 어떻게 할 것인가? (방법론만 제시)
*식과 함께 답을 같이 제출해주세요~

*TIP: 나이브 베이즈 함수가 어떻게 생겼죠? ?

참고자료

⭐ 인공지능 및 기계학습 개론1
https://www.edwith.org/machinelearning1_17/joinLectures/9738

상상 머신러닝
https://www.youtube.com/watch?v=hO9SVW6nnhM

데이터 과학을 위한 통계

데이터 과학 입문

패턴 인식과 머신러닝

Q & A

들어주셔서 감사합니다.