베이즈통계학 입문

wonseok Jung

1. 베이즈 추정에서는 정보를 순차적으로 사용할 수 있다.
2. 베이즈 추정은 정보를 얻을수록 더 정확해진다.

1. 베이즈 추정에서는 정보를 순차적으로 사용할 수 있다.

베이즈 추정에서는 이전의 정보를 잊어도 앞뒤가 들어 맞는다.

- 어떠한 정보 두개를 얻었을때의 베이즈에서 확률 변화

$$Prior \rightarrow_{info1} \rightarrow Posterior \rightarrow_{info2} \rightarrow AnotherPosterior$$

- Information 1 에 의해 확률이 바뀌었다.
- 새로운 Information2를 얻어 다시 확률이 바뀔때, Information1을 얻기 전의 확률 $prior$ 은 잊어도 된다.
- 질문 : 그렇다면 information1 이전의 확률 $prior$은 전혀 필요없는 확률일까?

Example - Email filter

$$Prior \rightarrow_{info1} \rightarrow Posterior \rightarrow_{info2} \rightarrow AnotherPosterior$$

- Prior and conditional probability about Link:
  - $P(spam) = 0.5$
  - $P(link \mid spam) = 0.6$,
  - $P(Notlink \mid spam) = 0.4$
  - $P(Notspam) = 0.5$
  - $P(link \mid Notspam) = 0.2$
  - $P(Notlink \mid Notspam) = 0.8$

Information1 - 링크 검출

- $P(spam) = 0.5$
- $P(link \mid spam) = 0.6$
- $P(Notspam) = 0.5$
- $P(link \mid Notspam) = 0.2$
- 스팸일 확률 : 아닐확률
  $0.5 \times 0.6 : 0.5 \times 0.2$
  $= 0.3 : 0.1$
  $= \frac{3}{4} : \frac{1}{4}$
  = 75% : 25%

Posterior을 Prior로 사용

- 위에서 구한 Posterior을 다시 Prior을 재설정한다.
- 이미 사전확률이라는 것은 근거없이 설정되어 있는 값
- 질문 : 너무 불확실한것 아닐까? 이렇게 계속 정보를 넣어가며 확률이 바뀌면 언젠가는 수렴하는가?

New Prior and condition probability about word "meet"

- $P(spam) = 0.75$
- $P(meet \mid spam) = 0.4$
- $P(Nomeet \mid spam) = 0.6$
- $P(Notspam) = 0.25$
- $P(Meet \mid Notspam) = 0.05$
- $P(Nomeet \mid Notspam) = 0.95$

Information2 - "meet" 단어검출

- $P(spam) = 0.75$
- $P(meet \mid spam) = 0.4$
- $P(Notspam) = 0.25$
- $P(Nomeet \mid Notspam) = 0.95$
- Spam일 확률 : Spam이 아닐 확률

$$0.75 \times 0.4 : 0.25 \times 0.95$$

$$= 0.3 : 0.01$$
$$= \frac{0.3}{0.31} : \frac{0.01}{0.31}$$
$$= 0.97 : 0.03$$

- Information2에 의해 새로운 Prior

두가지 방법

- Info1으로부터 수정한 Posterion을 Prior로 설정하고 또다른 Info2로 인해 수정된 Posterior

$$Prior \rightarrow_{info1} \rightarrow Posterior \rightarrow_{info2} \rightarrow AnotherPosterior$$

- Info1, Info2을 한번에 사용해서 구한 Posterior

$$Prior \rightarrow_{info1,info2} \rightarrow AnotherPosterior$$

- 두 방법을 통하여 구한 Posterior는 일치한다.

축차합리성 (영어로 무엇일까?)

- 두가지 정보를 한꺼번에 사용하여 추정한 결과와
- 첫번째 정보를 사용하여 추정 -> 그 추정 결과를 사전확률로 변경 -> 두 번째 정보를 사용하여 추정한 결과가 완전히 일치
- 이러한 성질을 축차합리성 이라고 한다.

축차합리성 -2

- 이전에 사용한 정보는 잊어도 관계없다.
  - 그것은 그로부터 얻은 사후확률에 반영

베이즈 추정은 정보를 얻을수록 더 정확해진다.

적당적당한 추측에서 더 정확한 추정으로 만들려면

-베이즈추정 정보가 더 많아질수록 더 정확한 추정을 한다.

Example - Bowl Problem

문제 :

- 눈앞에 Bowl가 하나 있다.
- Bowl A 혹은 Bowl B라는 사실은 알고 있지만 겉으로 봐서는 어느 쪽인지 알 수가 없다.
- 두개의 공을 뽑는데 처음 꺼낸 공을 다시 단지에 넣고 새로 공을 한깨 뽑는 경우 (independent)
- Bowl A
    - White ball 9, Black ball 1
- Bowl B
    - White ball 2, Black ball 8

Prior 설정 - $P(BowlA)$

- $P(BowlA) = 0.5$
- $P(black \mid BowlA) = 0.1$
- $P(white \mid BowlA) = 0.9$

Prior 설정 - $P(BowlB)$

- $P(BowlB) = 0.5$
- $P(black \mid BowlB) = 0.8$
- $P(white \mid BowlB) = 0.2$

Information - 첫번째와 두번째 공 모두 Black 관측

- BowlA : BowlB
- $0.5 \times 0.1 \times 0.1 : 0.5 \times 0.8 \times 0.8$
- $0.005 : 0.32$
- $P(BowlA) = 0.005/0.325 = 0.015384615384615384$
- $P(BowlB) = 0.32/0.325 = 0.9846153846153846$

$0.015 : 0.984$

Information - 첫번째는 Blck, 두번째는 White 공이 관측

- BowlA : BowlB
- $0.5 \times 0.1 \times 0.9 : 0.5 \times 0.8 \times 0.2$
- $= 0.045 : 0.08$
- $P(BowlA) = 0.045/0.125 = 0.36$
- $P(BowlB) = 0.08/0.125 = 0.64$

- 첫번째와 두번째 공이 모두 Black 공 관측
    - P(Bowl|B)= 0.32 / 0.325 =0.98$
- 첫번째는 Blck, 두번째는 White 공 관측
    - $P(BowlB) = 0.08/0.125 = 0.64$
- $0.98 \rightarrow 0.64$ 앞에있는 Bowl이 B일 것이라는 확률이 작아짐

최신 관측결과에 따라 결론이 달라진다.

- (n+1)번째 Posterior을 계산하기 위해 n번째까지의 공의 색을 열거한 필요는 없다.
- n번째의 Posterior에 그 모든 확률이 전부 반영되어있다.

Example n+1번째에 black 공을 뽑았을때 각 Bowl의 Posterior

(n+1)번째의 $P(BowlA)$: (n+1)번째의 $P(BowlB)$

$= a' : b'$

$= a \times 0.1 : b \times 0.8$

$= a : 8b$

- Bowl A :Bowl B 8배의 비례관계

여러 번 관측할수록 추측은 진실에 가까워진다.

- 베이즈 추정은 관측을 많이 하면 할수록 정확한 결론을 내릴 수 있다.
- (정보가 많이 있으면) 더욱 올바른 결론을 내릴 수 있다.