풀이

STEP II 1998, 문제 12

확률과 통계 • 베이즈 정리

> **STEP II 1998, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/98-s2-q12

이 문제는 *Beyond STEP*의 12.2 절 **기저율 오류**에서 단계별 안내와 함께 자세히 다룬다:
https://beyondstep.keithmcnulty.org/ko/ch12/base-rate.html

핵심 아이디어

이 문제는 베이즈 정리(Bayes' theorem)를 두 번 쓰는 문제이다. 첫 번째 부분은 고전적인 "희귀병" 역설이다. 병이 드물면 꽤 정확한 검사라 해도 양성 판정의 대부분이 위양성(false positive)일 수 있다. 두 번째 부분의 요령은 STEP을 치르는 사람들이 전체 인구에서 무작위로 뽑은 표본이 아니라는 점을 알아차리는 것이다. 이들은 AL 양성 집단이므로, 첫 번째 부분의 답이 두 번째 검사의 **사전 확률**(prior)이 된다.

풀이

어떤 사람이 수학염(mathematitis)에 걸려 있는 사건을 D, 검사 AL이 양성인 사건을 A, 검사 STEP이 양성인 사건을 B라 하자. 주어진 조건은 다음과 같다.

$$\mathrm{P}(D) = \tfrac{1}{100}, \qquad \mathrm{P}(A \mid D) = \tfrac{9}{10}, \qquad \mathrm{P}(A \mid D') = \tfrac{1}{11}.$$

프랭키의 확률

전확률 공식(law of total probability)에 의해

$$\begin{aligned}\mathrm{P}(A) &= \mathrm{P}(A \mid D)\,\mathrm{P}(D) + \mathrm{P}(A \mid D')\,\mathrm{P}(D') \\ &= \frac{9}{10}\cdot\frac{1}{100} + \frac{1}{11}\cdot\frac{99}{100} = \frac{9}{1000} + \frac{9}{100} = \frac{99}{1000}.\end{aligned}$$

이다. 그러면 베이즈 정리에 의해

$$\mathrm{P}(D \mid A) = \frac{\mathrm{P}(A \mid D)\,\mathrm{P}(D)}{\mathrm{P}(A)} = \frac{9/1000}{99/1000} = \frac{1}{11}.$$

이다. 따라서 AL 결과가 양성으로 나온 뒤에도 프랭키가 환자일 확률은 $\frac{1}{11}$에 불과하다. 1000명으로 이루어진 집단에서 환자 약 9명이 양성 판정을 받지만, 환자가 아닌 990명 중에서도 약 90명이 양성 판정을 받으므로, 위양성이 진양성을 압도한다.

p 구하기

이제 AL이 양성으로 나온 사람들로 관심을 좁히자. 이 집단 안에서는

$$\mathrm{P}(D \mid A) = \tfrac{1}{11}, \qquad \mathrm{P}(D' \mid A) = \tfrac{10}{11}.$$

이다. STEP에 관해 주어진 정보는 $\mathrm{P}(B \mid D \cap A) = 0.9$, $\mathrm{P}(B \mid D' \cap A) = p$로 해석한다. 즉 STEP의 결과는 그 사람이 환자인지 아닌지에만 의존하고, AL의 결과에는 의존하지 않는다.

> ? **생각해 보기** 왜 $\mathrm{P}(D) = \frac{1}{100}$에서 다시 시작하지 않는가?

STEP은 AL 양성인 사람들에게만 시행되고, 그 집단에서 환자의 비율은 더 이상 1%가 아니기 때문이다. 그 비율은 $\frac{1}{11}$이며, 이것이 바로 첫 번째 부분에서 구한 값이다. 첫 번째 검사의 사후 확률(posterior)이 두 번째 검사의 사전 확률이 된다.

AL 양성 집단 안에서

$$\mathrm{P}(D \cap B \mid A) = \frac{1}{11} \cdot \frac{9}{10} = \frac{9}{110},$$
$$\mathrm{P}(D' \cap B \mid A) = \frac{10}{11} \cdot p = \frac{10p}{11}.$$

이다. 두 검사에서 모두 양성인 사람들 중 절반이 환자라고 했으므로 $\mathrm{P}(D \mid A \cap B) = \frac{1}{2}$이다. 이는 위의 두 확률이 같다는 뜻이다.

$$\frac{9}{110} = \frac{10p}{11} \quad \Longrightarrow \quad p = \frac{9}{100} = 0.09.$$

검산

- 오즈(odds) 형태: 병에 걸렸을 사전 오즈는 $1 : 99$이다. AL은 이를 $\frac{0.9}{1/11} = 9.9$배 하여 $9.9 : 99 = 1 : 10$을 주며, 이는 $\frac{1}{11}$과 일치한다. 오즈가 $1 : 1$이 되려면 STEP이 10배를 곱해야 하므로 $\frac{0.9}{p} = 10$이고 $p = 0.09$이다.
- 200만 명이 두 검사를 모두 거치는 상황을 ($p = 0.09$로) 시뮬레이션한 결과, AL 양성자 중 환자의 비율은 0.0909, 두 검사 모두 양성인 사람 중 환자의 비율은 0.5003이었다.

흔한 실수

- STEP의 사전 확률로 $\mathrm{P}(D) = 0.01$을 쓰는 것. 그러면 $p = 0.0009$가 되어 백 배나 작은 값이 나온다.
- $\mathrm{P}(A \mid D) = 0.9$와 $\mathrm{P}(D \mid A)$를 혼동하는 것. 첫 번째 부분의 요점 자체가 이 둘이 매우 다르다는 것이다.

STEP 너머

베이즈 정리를 오즈 형태, 즉 사후 오즈가 사전 오즈에 가능도비(likelihood ratio)를 곱한 것과 같다는 형태로 쓰면, 순차적인 검사는 거의 기계적인 계산이 된다. 독립적인 증거 하나하나가 각자의 인수만큼 오즈를 곱해 줄 뿐이다. 이것이 베이즈 추론(Bayesian inference)의 기초이며, 실제로 의료 선별 검사 프로그램이 설계되는 방식의 바탕이기도 하다. 우리가 둔 가정, 즉 질병 여부가 주어졌을 때 두 검사가 조건부 독립(conditionally independent)이라는 가정은 "나이브 베이즈(naive Bayes)" 분류기의 바탕이 되는 가정과 같다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 조건부 확률(conditional probability)과 베이즈 규칙을 이와 같은 예제를 많이 들어 훌륭하게 다룬다.

Beyond STEP 풀이: https://beyondstep.keithmcnulty.org/ko/. 풀이 글은 CC0 1.0이다. 문제 자체의 저작권은 © Cambridge University Press & Assessment에 있으며 여기에 싣지 않는다. 영어 원문을 기계 번역하였으며 아직 검토되지 않았다.