풀이

STEP I 2004, 문제 12

확률과 통계 • 베이즈 정리

> **STEP I 2004, 12번 문제**
>
> 더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:
>
> step.maths.org/questions/04-s1-q12

핵심 아이디어

그룹 1에서 뽑은 표본은 칩이 그룹 1에 있다는 **조건 아래에서** 그 칩이 기계 A에서 나왔을 조건부 확률(conditional probability)을 근사적으로 알려 준다. 따라서 베이즈 정리(Bayes' theorem)를 이용해 이 조건부 확률을 λ, p, q로 나타내고, 이것을 관찰된 비율과 같다고 놓은 뒤 λ에 대해 푼다. 분류가 잘못되는 경우에 달라지는 것은 각 기계에서 나온 칩이 그룹 1에 들어갈 확률뿐이며, 방법은 똑같다.

풀이

공장의 생산품 중에서 칩 하나를 임의로 고른다. 그 칩이 기계 A에서 만들어졌다는 사건을 A라 하고(따라서 $\mathrm{P}(A) = \lambda$, $\mathrm{P}(A') = 1 - \lambda$이다), 칩이 완벽하다는 사건을 P, 칩이 그룹 1에 분류된다는 사건을 G라 하자. 문제에서 $\mathrm{P}(P \mid A) = p$, $\mathrm{P}(P \mid A') = q$라고 주어져 있다.

그룹 1에서 임의로 뽑은 칩은 생산품 전체에서 임의로 뽑은 칩을 G라는 조건 아래에서 본 것과 같다. 표본이 크므로 관찰된 비율 $\frac{2}{5}$는 $\mathrm{P}(A \mid G)$의 좋은 추정값이며, λ의 추정값은 $\mathrm{P}(A \mid G)$를 정확히 $\frac{2}{5}$로 만드는 값이다.

완벽한 분류

이 경우 G는 P와 같은 사건이다. 베이즈 정리에 의해

$$\mathrm{P}(A \mid P) = \frac{\mathrm{P}(P \mid A)\,\mathrm{P}(A)}{\mathrm{P}(P \mid A)\,\mathrm{P}(A) + \mathrm{P}(P \mid A')\,\mathrm{P}(A')} = \frac{\lambda p}{\lambda p + (1-\lambda)q}.$$

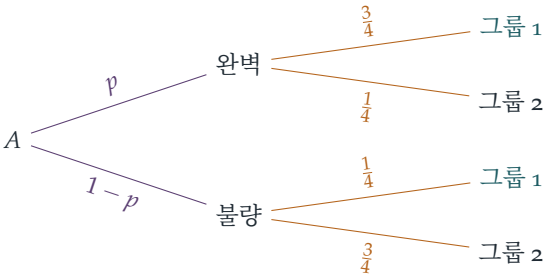
이것을 $\frac{2}{5}$와 같다고 놓고 양변에 분모를 곱하면(분모는 양수이다)

$$5\lambda p = 2\lambda p + 2(1-\lambda)q$$
$$3\lambda p + 2\lambda q = 2q$$
$$\lambda = \frac{2q}{3p+2q},$$

이고, $3p + 2q > 0$이므로 이것이 구하는 추정값이다.

잘못된 분류

이제 칩이 그룹 1에 들어가는 경우는 완벽한 칩이 올바르게 분류되는 경우와 불량 칩이 잘못 분류되는 경우 두 가지이다. 분류 오류는 칩을 만든 기계와 독립적으로 일어난다고 가정한다. 즉 칩이 완벽한지 아닌지가 주어지면, 그 칩의 분류는 어느 기계에서 나왔는지에 의존하지 않는다. 아래 수형도는 기계 A에서 나온 칩의 가지를 보여 준다. 기계 B는 p 대신 q를 넣으면 똑같다.



그룹 1로 가는 두 가지를 더하면

$$\mathrm{P}(G \mid A) = \tfrac{3}{4}p + \tfrac{1}{4}(1-p) = \frac{1+2p}{4}, \qquad \mathrm{P}(G \mid A') = \tfrac{3}{4}q + \tfrac{1}{4}(1-q) = \frac{1+2q}{4}.$$

따라서 다시 베이즈 정리에 의해(인수 $\frac{1}{4}$는 약분된다)

$$\mathrm{P}(A \mid G) = \frac{\lambda(1+2p)}{\lambda(1+2p) + (1-\lambda)(1+2q)}.$$

생각해 보기 증거가 달라졌는가? $\frac{2}{5}$를 여전히 사용해야 하는가?

그렇다. 표본은 실제로 그룹 1에서 뽑았고, 그 중 $\frac{2}{5}$는 실제로 기계 A에서 나왔다. 달라진 것은 칩이 그룹 1에 들어가는 방식에 대한 우리의 모형뿐이다. 따라서 $\mathrm{P}(A \mid G)$에 대한 새 식

을 $\frac{2}{5}$와 같다고 놓는다.

$$\begin{aligned} 5\lambda(1+2p) &= 2\lambda(1+2p) + 2(1-\lambda)(1+2q) \\ \lambda\big[3(1+2p) + 2(1+2q)\big] &= 2(1+2q) \\ \lambda &= \frac{2(1+2q)}{5+6p+4q}. \end{aligned}$$

이것이 λ의 새 추정값이다. 이 답은 첫 번째 답에서 p와 q를 그룹 1에 들어갈 새 확률인 $\frac{1+2p}{4}$와 $\frac{1+2q}{4}$로 바꾼 것임에 주목하자.

검산

- $p = q$이면 두 기계는 품질로 구별되지 않으므로 그룹 1은 생산 비율을 그대로 반영해야 하고 λ는 $\frac{2}{5}$가 되어야 한다. 두 공식 모두 이 값을 준다: $\frac{2p}{5p} = \frac{2}{5}$, $\frac{2(1+2p)}{5+10p} = \frac{2}{5}$.
- 두 추정값 모두 분자가 양수이고 분모보다 작으므로 0과 1 사이에 엄격하게 놓인다.
- $\lambda = 0.3$, $p = 0.6$, $q = 0.8$이고 분류가 잘못되는 경우에 칩 4×10^5개로 모의실험(simulation)을 하였더니, 그룹 1 중 기계 A에서 나온 비율이 0.2667로, 공식의 값 0.2661과 비교된다. 모의실험에서 얻은 비율로 공식을 거꾸로 풀면 $\lambda \approx 0.301$을 되찾는다.

흔한 실수

- $\frac{2}{5}$를 조건부 확률 $\mathrm{P}(A \mid G)$가 아니라 $\mathrm{P}(G \mid A)$나 λp와 같다고 놓는 것.
- 두 번째 부분에서 그룹 1에 잘못 들어간 불량 칩을 빠뜨리는 것, 즉 $\frac{3}{4}p$만 사용하는 것.

STEP 너머

관찰된 자료에 "정확히 들어맞는" 모수(parameter) 값을 고르는 것은 **적률법**(method of moments)이다. 여기서는 이것이 최대가능도 추정(maximum likelihood estimation)과 일치하는데, 최대가능도 추정은 대학 첫 통계학 강의에서 만나게 된다. 잘못 분류하는 기계는 측정 오차(measurement error), 즉 오분류(misclassification)의 간단한 모형이다. 이는 의학 검사에서 심각한 실제 문제로, 위양성(false positive)과 위음성(false negative)이 바로 이런 식으로 관찰된 비율을 왜곡한다. 베이즈 정리와 조건부 확률에 대해 더 알고 싶다면 Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)를 보라.

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.