풀이

STEP III 1998, 문제 13

확률과 통계 • 최대가능도 추정

STEP III 1998, 13번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/98-s3-q13

핵심 아이디어

여기서 미지수는 n이고, n은 정수이므로 미적분(calculus)은 쓸모가 없다. n에 대해 미분할 수 없기 때문이다. 수열(sequence)을 최대로 만드는 표준적인 도구는 이웃한 두 항의 비(*ratio*)를 살펴보는 것이다. 비가 1보다 큰 동안 수열은 증가하고, 비가 1보다 작아지면 감소한다. 그리고 동률(두 개의 최대가능도 추정량(maximum likelihood estimate, MLE))은 어떤 지점에서 비가 정확히 1과 같을 때에만 생긴다.

풀이

공정한 동전을 n번 던질 때 나오는 앞면의 개수는 이항분포(binomial distribution)를 따르므로

$$\mathrm{P}(k\text{회 앞면}, n\text{회 던짐}) = \binom{n}{k}\frac{1}{2^n}, \qquad n \geq k,$$

이고, $n < k$이면 확률은 0이다. 앞으로 계속 $k \geq 1$이라 하자(친구는 앞면을 적어도 한 번 보았다). 따라서 후보는 $n \geq k$뿐이다.

공정한 동전에 대한 일반적인 계산

$n \geq k$에 대하여 $L(n) = \binom{n}{k}2^{-n}$이라 쓰고, 이를 k를 고정한 n의 함수로 생각한다. $n \geq k$일 때

$$\frac{L(n+1)}{L(n)} = \frac{\binom{n+1}{k}}{\binom{n}{k}} \cdot \frac{1}{2} = \frac{n+1}{n+1-k} \cdot \frac{1}{2},$$

이다. 여기서 $\binom{n+1}{k} = \frac{n+1}{n+1-k}\binom{n}{k}$를 이용하였다. $n+1-k > 0$이므로 부등호 방향을 바꾸지 않고 분모를 없앨 수 있다.

$$\frac{L(n+1)}{L(n)} \geq 1 \iff n+1 \geq 2(n+1-k) \iff n \leq 2k-1,$$

그리고 등호는 정확히 $n = 2k-1$일 때 성립한다. 따라서

$$L(k) < L(k+1) < \cdots < L(2k-1) = L(2k) > L(2k+1) > L(2k+2) > \cdots .$$

생각해 보기 비를 보면 최댓값이 어디에 있는지 어떻게 알 수 있는가?

$n < 2k-1$이면 비가 1보다 크므로 각 항이 바로 앞의 항보다 크다. 즉 수열은 $n = 2k-1$까지 순증가한다. $n = 2k-1$에서 비는 정확히 1이므로 $L(2k) = L(2k-1)$이다. $n > 2k-1$이면 비가 1보다 작으므로 그 이후로 수열은 순감소한다. 그러므로 최댓값은 정확히 두 개의 n 값에서 얻어진다.

앞면이 한 번

$k = 1$이면 두 MLE는 $n = 2k-1 = 1$과 $n = 2k = 2$이다. 직접 계산하면

$$L(1) = \tfrac{1}{2}, \quad L(2) = 2 \cdot \tfrac{1}{4} = \tfrac{1}{2}, \quad L(3) = 3 \cdot \tfrac{1}{8} = \tfrac{3}{8},$$

이고, 위의 논증은 $n = 2$ 이후로 $L(n)$이 계속 감소함을 보여 준다. 따라서 요구한 대로 MLE는 1과 2, 두 개이다.

앞면이 k번

일반적인 계산에 의해, 던진 횟수의 두 MLE는

$$n = 2k-1 \quad 과 \quad n = 2k.$$

이다.

편향된 동전

이제 $0 < p < 1, q = 1-p$라 하면

$$L(n) = \binom{n}{k} p^k q^{n-k}, \qquad n \geq k.$$

이다. $n \geq k$일 때 인수 p^k는 약분되고 인수 q가 하나 더 생긴다.

$$\frac{L(n+1)}{L(n)} = \frac{(n+1)q}{n+1-k}.$$

이번에도 $n+1-k>0$이므로

$$\frac{L(n+1)}{L(n)} \gtreqless 1 \iff (n+1)(1-p) \gtreqless n+1-k \iff k \gtreqless (n+1)p \iff n+1 \lesseqgtr \frac{k}{p}.$$

이다. 따라서 L은 $n+1<k/p$인 동안 증가하고, $n+1=k/p$일 때 같은 값을 유지하며, $n+1>k/p$가 되면 감소한다. $p<1$이므로 $k/p>k$이고, 따라서 전환점은 범위 $n \geq k$ 안에 있다는 점에 주목하자.

- k/p가 정수가 **아니면** $m=\lfloor k/p \rfloor$라 하자. 그러면 $n+1<k/p$는 정확히 $n+1 \leq m$일 때 성립하므로 $L(k)<\cdots<L(m)$이고, $L(m+1)<L(m)$이며 그 뒤로도 계속 감소한다. MLE는 하나뿐이며
$$n=\left\lfloor \frac{k}{p} \right\rfloor.$$
이다.
- $k/p=N$이 정수이면 L은 $L(N-1)$까지 증가하고, $L(N)=L(N-1)$이며, 그 뒤로 감소한다. MLE는 두 개이며
$$n=\frac{k}{p}-1 \quad \text{과} \quad n=\frac{k}{p}.$$
이다. 둘 다 k 이상이다. N은 $N>k$인 정수이므로 $N-1 \geq k$이기 때문이다.

따라서 MLE가 유일할 **필요충분조건은 k/p가 정수가 아닌 것이다**. 공정한 동전에서는 $k/p=2k$가 항상 정수이므로, 앞의 소문항들에서 언제나 두 MLE $2k-1$과 $2k$가 나온 것이다.

검산

- 답이 타당하다. 앞면이 k번 나왔고 앞면이 나올 확률이 p라면 대략 k/p번 던졌으리라 예상할 수 있는데, MLE는 k/p를 내림한 값이다(정수일 때는 양옆의 동률).
- $p=\frac{1}{3}, k=1$이라 하면 $k/p=3$이다. 이때 $L(2)=2\cdot\frac{1}{3}\cdot\frac{2}{3}=\frac{4}{9}$이고 $L(3)=3\cdot\frac{1}{3}\cdot\frac{4}{9}=\frac{4}{9}$로 동률이며, $L(4)=4\cdot\frac{1}{3}\cdot\frac{8}{27}=\frac{32}{81}<\frac{4}{9}$이다.
- 컴퓨터로 분수를 이용해 $L(n)$을 정확히 계산하여, 공정한 동전에 대해 $k=1,\ldots,39$인 경우(항상 $\{2k-1,2k\}$)와, p가 분모 12 이하인 분수이고 $k \leq 14$인 924개의 쌍 (p,k)에 대해 최대가 되는 값을 구하였다. 모든 경우가 위의 규칙과 일치하였다.

흔한 실수

- n에 대해 미분하는 것. 여기서 n은 정수이고, 추가적인 도구 없이는 $\binom{n}{k}$가 다른 n 값에 대해 정의되지 않는다. 비를 살펴보는 방법이 올바른 도구이다.

- 부등식의 방향을 거꾸로 잡거나, 비가 1과 같아지는 경우가 두 번째 MLE를 만든다는 사실을 잊는 것.
- 모든 경우에 $\lfloor k/p \rfloor$라고 답하여, k/p가 정수일 때의 동률을 놓치는 것.

STEP 너머

최대가능도(maximum likelihood)는 통계학에서 추정(estimation)의 핵심 방법이다. 관측된 자료가 나올 확률을 가장 크게 만드는 모수(parameter) 값을 고르는 것이다. 보통 모수는 연속적이며, 로그가능도(log-likelihood)를 미적분으로 최대화한다. n과 같은 정수를 추정하는 문제("시행 횟수" 문제)는 이 방법이 통하지 않는 고전적인 예이며, 이 문제는 MLE가 유일하지조차 않을 수 있음을 보여 준다. Blitzstein과 Hwang의 *Introduction to Probability*(확률론 입문)는 이항분포를 잘 다루고 있으며, Grimmett과 Welsh의 *Probability: An Introduction*(확률: 입문)은 배경이 되는 확률론을 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.