정렬된 1-차원 데이터의 k-구간 최소 WCSS 분할

문제 정의

정렬된 스칼라 데이터

$$x_1 \le x_2 \le \cdots \le x_n \quad (x_i \in \mathbb{R})$$

를 k개의 **연속 구간** $C_1, \ldots, C_k$ 로 나누어
군집 내 제곱편차 합 (Within-Cluster Sum of Squares, WCSS)

$$\Phi = \sum_{m=1}^{k} \sum_{x \in C_m} (x - \mu_m)^2, \qquad \mu_m = \frac{1}{|C_m|} \sum_{x \in C_m} x$$

을 최소화한다.
(크기 1인 군집의 분산은 0으로 정의한다.)

최적 분할은 구간 분할임

정리 (연속성):
위 문제의 전역 최적 해는 항상 각 군집이 지표 구간
$C-m = x-a-m, x-a-m+1, \ldots, x-b-m$ 형태이다.

증명 개요:
군집 평균을 $\mu-1 < \mu-2 < \cdots < \mu-k$ 로 두고,
교차 군집이 존재한다고 하자.
즉, $x-i, x-j \in C-p$, $x-\ell \in C-q$ $(p < q)$ 이고

$$x_i < x_\ell < x_j$$

인 경우가 존재한다고 하자. 그러면

$$|x_\ell - \mu_p| < |x_\ell - \mu_q|$$

이 되어 **최근접 평균 할당 조건**
$|x-\mu-C(x)| \le |x-\mu-C'|$ 을 위배하게 된다.
즉, 최적해에는 교차 군집이 존재할 수 없으며,
따라서 군집은 반드시 연속 구간이다. ■

동적 계획법 점화식

상태 정의

$$V(i,k) = \min\left(\mathrm{WCSS} | [x_1, \ldots, x_i] \text{를 } k\text{개 구간으로 분할}\right)$$

점화식

마지막 군집이 $x-t+1 \sim x-i$ 라 하면,
앞쪽 $x-1 \sim x-t$ 는 $(k-1)$ 개 구간으로 최적으로 분할되어야 한다.
(→ 최적 부분 구조)

$$V(i,k) = \min_{t<i} \left(V(t,k-1) + \mathrm{WCSS}(x_{t+1:i}) \right)$$

구현 메모

- 누적합 $S-1(j) = \sum -p \le jx-p$,
  $S-2(j) = \sum -p \le jx-p^2$ 를 미리 구해

  $$\mathrm{WCSS}(l,r) = S_2(r) - S_2(l-1) - \frac{[S_1(r) - S_1(l-1)]^2}{r-l+1}$$

  으로 계산.
- 기본값:

  $$V(0,0)=0, \quad V(i,1)=\mathrm{WCSS}(1,i), \quad V(\cdot,\cdot)=\infty \text{ (불가능 상태)}$$

- 복잡도:
  - 단순 구현: $O(n^2k)$
  - Divide-and-Conquer DP 사용시 $O(nk)$ 가능

요약

1. 연속성 정리를 통해 탐색 공간을 **구간 분할**로 축소
2. 마지막 경계 t만 정하면 나머지는 부분 최적해 → **최적 부분 구조**
3. 위 점화식과 누적합 계산으로 효율적인 DP 구현 가능