Chapter 3.4 Median(중간값) 구하기

Median 을 구하는 Randomized 알고리즘을 알아보기 전에,

Chapter 3.1 의 Markov's inequality 와 Chapter 3.3 의 Chebyshev's inequality 을 간략하게 알아본다. 두 inequality 는 Randomized 알고리즘의 정확도의 한계점 또는 최저점(bound)를 구하기 위해 사용한다.

Theorem 3.1 Markov's inequality
X ≥ 0 인, 랜덤변수 X 에 대해서, 다음의 식을 만족한다.

모든 a > 0 에 대해,

$$\Pr(\mathrm{X} \geq \mathrm{a}) \leq \frac{E[X]}{a}$$

Proof)
모든 a > 0 에 대해, 변수 I 를 다음과 같이 정의한다.

$$I = \begin{cases} 1, & if\ X \geq a \\ 0, & otherwise \end{cases}$$

그러면, X ≥ 0 에 대해서, 다음 식을 만족한다.

$$I \leq \frac{X}{a}$$

$$E[I] = \Pr(X \geq a) \leq E\left[\frac{X}{a}\right] = \frac{E[X]}{a}$$

Theorem 3.6 Chebyshev's inequality
모든 a > 0 에 대해서

$$\Pr(|E[X] - X| \geq\ a) \leq \frac{Var(X)}{a^2}$$

을 만족한다.

Proof)

$$\Pr(|E[X]-X| \geq a) = \Pr\left(|E[X]-X|^2 \geq a^2\right)$$

Markov's inequality 정리에 따라서,

$$\Pr(|E[X]-X|^2 \geq a^2) \leq \frac{(E[X]-X)^2}{a^2} = \frac{X^2-(E[X])^2}{a^2} = \frac{Var(X)}{a^2}$$

일반적으로, Median 을 구하는 Deterministic 알고리즘은 정렬이다. (더 빠른 알고리즘이 있지만, 복잡하므로 설명하지 않는다.) 정렬 알고리즘은 $O(nlogn)$ 시간 복잡도를 갖는다.
그러나 Randomized 알고리즘은 O(n)으로 Median 을 꽤 높은 정확도로 구할 수 있다. (오답률 최대 $n^{-\frac{1}{4}}$)

그러면, Median 를 구하는 Randomized 방법을 설명한다.

1. 입력으로 주어진 N 개의 수열을 S 라고 한다.(계산의 편의상 N 은 홀수라고 한다.)
2. 수열 S 에서 $n^{\frac{3}{4}}$ 개 만큼 샘플을 추출한다. 추출한 수열을 R 이라고 한다.
3. R 을 정렬한다. ($n^{\frac{3}{4}}logn^{\frac{3}{4}} \in O(n)$)
4. 정렬된 수열 R 에서 $(\frac{1}{2}n^{\frac{3}{4}} - \sqrt{n})$ 번째 숫자를 선택한다. 이 수를 l(L)라고 정한다.
5. 다시 수열 R 에서$(\frac{1}{2}n^{\frac{3}{4}} + \sqrt{n})$ 번째 숫자를 선택한다. 이 수를 u 라고 정한다.
6. 두 수 l(L)과 u 의 rank 를 S 에서 구한다. (각각 O(n))
7. 위의 두 수 사이에 포함되는 수를 S 에서 구한다. 뽑아낸 수 집합을 C 라고 하자.
8. $|C| \leq 4n^{\frac{3}{4}}$ 이면, C 를 정렬하여 median 을 구할 수 있다.

(8 번의 이유는 아직 정확하게 알지 못 한다.)

다음으로 위 알고리즘의 오답률이 최대 $n^{-\frac{1}{4}}$ 을 증명한다.

이 알고리즘은 다음 3 가지 조건을 만족하는 경우, 실패한다.

1) $Y_1 = |\{r \in R \mid r \geq m\}| < \frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}$
2) $Y_2 = |\{r \in R \mid r \leq m\}| < \frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}$
3) $|C| \geq 4n^{\frac{3}{4}}$

(1 번 조건이 만족하는 경우, 중간값은 u 보다 큰 값을 갖게 된다. 2 번 조건이 만족하는 경우는 중간값이 l(L)보다 작은 값을 갖게 된다.)

1 번 조건이 발생할 확률, 즉 $\Pr(Y_1) \leq \frac{1}{4}n^{-\frac{1}{4}}$ 임을 다음과 같이 증명할 수 있다.

Proof)

먼저 변수 Xi 를 다음과 같이 정의한다.

$$X_i = \begin{cases} 1 & if\ i\ th\ sample\ is\ less\ than\ or\ equal\ to\ median \\ 0 & otherwise \end{cases}$$

각각의 X_i는 서로 독립적이다. 그리고 중간값(median)보다 작은 수는 $\frac{(n-1)}{2}+1$ 개가 있으므로,

$$\Pr(X_i = 1) = \frac{\left(\frac{n-1}{2}+1\right)}{n} = \frac{1}{2}+\frac{1}{2n}$$

이다.

따라서 Y_1은 다음과 같이 표현할 수 있다.

$$Y_1 = \sum_{i=1}^{n^{\frac{3}{4}}} X_i$$

Y_1은 Bernoulli 랜덤변수의 합이므로, Binomial 랜덤 변수이다. (n=$n^{\frac{3}{4}}$, p=$\frac{1}{2}+\frac{1}{2n}$)

$$\begin{aligned} \text{Var}[Y_1] = \text{npq} &= n^{\frac{3}{4}}\left(\frac{1}{2}+\frac{1}{2n}\right)\left(\frac{1}{2}-\frac{1}{2n}\right) \\ &= \frac{1}{4}n^{\frac{3}{4}} - \frac{1}{5n^{\frac{5}{4}}} \\ &< \frac{1}{4}n^{\frac{3}{4}} \end{aligned}$$

따라서, 위의 1)번 조건이 만족하지 않을 확률은 다음과 같이 표현할 수 있다.

$$\begin{aligned} \Pr(E_1) &= \Pr\left(Y_1 < \frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}\right) \\ &\leq \Pr\left(|Y_1 - E[Y_1]| > \sqrt{n}\right) \end{aligned}$$

$$\leq \frac{\mathrm{Var}[\mathrm{Y}_1]}{\mathrm{n}} \ (\ by\ chebyshev's\ law\ )$$
$$\leq \frac{\frac{1}{4}n^{\frac{3}{4}}}{\mathrm{n}} = \frac{1}{4}n^{-\frac{1}{4}}$$

조건 2)는 1)과 같으므로 동일하다.

조건 3)이 발생하는 경우는 두 조건을 동시에 만족하는 경우이다.
3.1) C 에 중간값보다 큰 원소가 2 $n^{\frac{3}{4}}$보다 많다.
3.2) C 에 중간값보다 작은 원소가2 $n^{\frac{3}{4}}$보다 많다.

3.1 번의 조건을 만족하기 위해서는 R 에서 u 를 선택할 때, u_{rank}가 적어도 $\frac{1}{2}n + 2n^{\frac{3}{4}}$ 보다 커야한다. 따라서, $\frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}$ 안에, $\frac{1}{2}n - 2n^{\frac{3}{4}}$개의 수가 포함되어야한다.

1 번 조건과 비슷하게 진행하면,
$X_i = \begin{cases} 1 & if\ i\ th\ sample\ is\ among\ the \frac{1}{2}n - 2n^{\frac{3}{4}}\ largest\ elements\ in\ S \\ 0 & otherwise \end{cases}$
로 정의할 수 있다.

$E[X] = np = n^{\frac{3}{4}} * \frac{\left(\frac{1}{2}n - 2n^{\frac{3}{4}}\right)}{n} = n^{\frac{3}{4}}\left(\frac{1}{2} - 2n^{\left(-\frac{1}{4}\right)}\right) = \frac{1}{2}n^{\frac{3}{4}} - 2n^{\frac{1}{2}}$

$Var[X] = npq = n^{\frac{3}{4}} * \left(\frac{1}{2} - 2n^{-\frac{1}{4}}\right)\left(\frac{1}{2} + 2n^{\frac{1}{4}}\right) = \frac{1}{4}n^{\frac{3}{4}} - 4n^{\frac{1}{4}} < \frac{1}{4}n^{\frac{3}{4}}$

따라서, 3.1 조건이 일어날 확률은 chebyshev 의 룰로 구하면,

$$\Pr(Y_{3.1}) = \Pr\left(X \geq \frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}\right)$$
$$= \Pr\left(|X - E[X]| \geq \sqrt{n}\right)$$
$$\leq \frac{Var[X]}{n} = \frac{\frac{1}{4}n^{\frac{3}{4}}}{n} = \frac{1}{4}n^{-\frac{1}{4}}$$

이와 마찬가지로, 3.2 도 같은 조건으로 발생하므로, 증명 과정을 생략한다.

이 알고리즘이 실패할 확률은 위의 조건들이 다 만족할 확률과 같다.

$$\Pr(Y_1) + \Pr(Y_2) + \Pr(Y_{3.1}) + \Pr(Y_{3.2}) \leq \frac{1}{4}n^{-\frac{1}{4}} + \frac{1}{4}n^{-\frac{1}{4}} + \frac{1}{4}n^{-\frac{1}{4}} + \frac{1}{4}n^{-\frac{1}{4}} = n^{-\frac{1}{4}}$$

이다.