Probability & Computing

Lecture 6

21/01/2020

1 Outline

- Review of Markov's Inequality
- Chebyshev's Inequality
- Randomized Median-find

2 Review of Markov's Inequality

Exercise 1 *Let $X \geq 0$, $E[X] = 10$. What can you say about $Pr[X \geq 50]$?*

Let $X \geq -10$ and $E[X] = 10$. What can you say about $Pr[X \geq 50]$?

Let $X \leq 20$ and $E[X] = 10$. What can you say about $Pr[X \leq 0]$?

Concentration around the mean using Markov:

Let $X = X_1 + \ldots + X_n$, where $Pr(X_i = 1) = p$, $Pr(X_i = 0) = 1-p$. Consider the values $p = 1/2$. What can we say about $Pr(X > 0.7n)$?

Using Markov's inequality directly, we can say that $Pr(X > 0.7n) < \dfrac{0.5n}{0.7n} = 5/7$. This is however a bad bound, and it is possible to use symmetry to improve this to $5/14$, which is still not a good bound for large values of n. We will improve this a lot in the next section by using the variance of X.

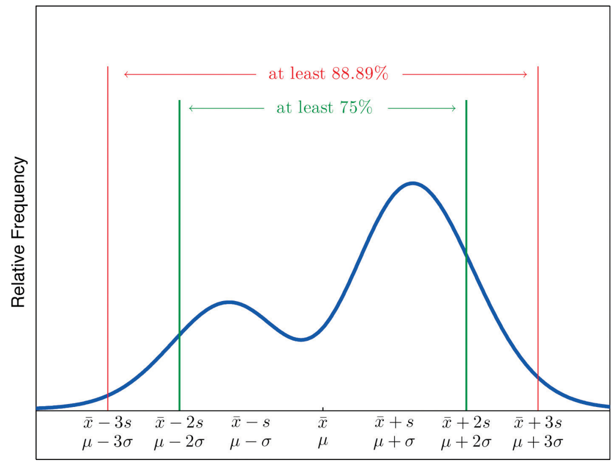
3 Chebyshev's Inequality

Theorem 1

Let X be a random variable with expectation μ and standard deviation σ. Then we have:

$$Pr[|X-\mu| \geq k\sigma \leq \frac{1}{k^2}.$$

For example, $Pr[|X-\mu| \geq 5\sigma] \leq \frac{1}{25}$, and $Pr[|X-\mu| \geq 10\sigma] \leq \frac{1}{100}$.



Exercise 2 *Suppose that the lifetime of light-bulbs from a shop is distributed randomly with a mean of 200 hours. If the variance is known to be at most 100, show that with probability at least 96%, the bulb will last for at least 150 hours.*

Variance of Binom(n,p):

Let X be the sum of n identical Bernoulli variables $X_1, \ldots, X_n$ with probability p. Then, we have $Var[X] = nVar[X_i] = n(p-p^2)$.

In particular, when $p = 1/2$, we get $Var[X] = \frac{n}{4}$ and $\sigma[X] = \frac{\sqrt{n}}{2}$. We can now apply Chebyshev's inequality to obtain:

$$Pr(X > 0.7n) \leq Pr(|X - 0.5n| > 0.2n) \leq \frac{1}{0.16n}$$

where we used $0.2n = 0.4\sqrt{n}\sigma[X]$.

4 Median-Find Problem

Input: $A = a_1, a_2, \ldots, a_n$

Output: $b_{n/2}$, where $b_1 < b_2 < \ldots < b_n$ is the sorted order of the a_is.

We now describe a randomized algorithm. We use parameters $r \sim n^{3/4}$, $k \sim n^{1/8}$, $s = \frac{k\sqrt{r}}{2} \sim \frac{n^{1/2}}{2}$, and $t = \frac{4sn}{r} \sim 2n^{3/4}$.

Algorithm:

1. *Pick $X_1, \ldots, X_r$ independently and u.a.r. from the a_is.*
2. *Sort the X_is to obtain $Y_1, \ldots, Y_r$.*
3. *Set $l = Y_{r/2-s}$ and $u = Y_{r/2+s}$.*
4. *Compare every element with l, u.*
5. *Let $L = \{a \in A : a < l\}, U = \{a \in A : a > u\}$.*

   If $|L| > n/2$ or $|U| > n/2$, ABORT.
6. *$M = \{a \in A : l \leq a \leq u\}$.*

   If $|M| > t$, ABORT.
7. *Sort M and return the element whose rank in M is $n/2 - |L|$.*

Analysis of running time:

If the algorithm does not fail, it returns the correct answer in Step 7. Step 4 takes $2n$ comparisons, and the remaining steps together take time $O(n^{3/4} \log n)$, thus the total time is $2n + o(n)$.

Analysis of failure probability:

There are two steps in which the algorithm can fail: in step 5, it may happen that the median lies outside the interval $[l, u]$, or in step 6, it may happen that $|M|$ is too large. We will now bound the probabilities of these two "bad events".

Let E_1 be the event that the l is larger than the median, that is: $Y_{r/2-s} > b_{n/2}$; and let E_2 be the event that $|M \cap \{b_1, \ldots, b_{n/2}\}| > t/2$.

Proposition 1 *For* $s = k\frac{\sqrt{r}}{2}$ *and* $t = \frac{4sn}{r}$*, we have:*

$Prob(E_1) \leq \frac{1}{k^2}$*, and* $Prob(E_2) \leq \frac{1}{k^2}$.

Proof:

The first type of bad event:

The event $E_1 : Y_{r/2-s} > b_{n/2}$ is equivalent to: $|X \cap L| \leq r/2 - s$, where $L = \{b_1, \ldots, b_{n/2}\}$.

We can write $|X \cap L|$ as a sum of indicator variables. Let $Y_i = 1$ if $X_i \in L$ and $Y_i = 0$ otherwise. Thus, $|R \cap L| = \sum_{i=1}^{r} Y_i$; we shall denote this sum by Y for convenience.

We have $E[Y] = \frac{r}{2}$ and $\sigma(Y) = \sqrt{r}/2$.

Thus, $Pr(E_1) = Pr(|X \cap L| \leq r/2 - s) = Pr(Y \leq r/2 - k\sqrt{r}/2) \leq \frac{1}{k^2}$.

The second type of bad event:

The event $|M \cap \{b_1, \ldots, b_{n/2}\}| > t/2$ is equivalent to: $|X \cap T| \geq r/2 - s$, where $T = \{b_1, b_2, \ldots, b_{n/2-t/2}\}$.

Let $Z = |X \cap T|$ so that $E[Z] = \frac{r}{n}(n/2 - t/2) = \frac{r}{2} - \frac{rt}{2n} = r/2 - 2s$ and we know that $\sigma[Z] \leq \sqrt{r}/2$.

Thus, $Pr(E_2) = Pr(Z \geq r/2 - s) \leq Pr(|Z - E[Z]| \leq s) \leq \frac{1}{k^2}$. This completes the proof of Proposition 1.

Finally, we find that the probability that the algorithm fails is at most $2Pr(E_1 \cup E_2) \leq \frac{4}{k^2} = \frac{4}{n^{1/4}}$.