Algorithms for Data Science: Lecture 3

Barna Saha

1 Concentration Inequalities

Lemma 1 (Markov's inequality)**.** *Let X be a non-negative random variable. For all $\lambda > 0$,*

$$\Pr[X > \lambda] \leq \frac{\boldsymbol{E}[X]}{\lambda}$$

Lemma 2 (Chebyshev Inequality)**.** *For all $\lambda > 0$,*

$$\Pr[|X - \boldsymbol{E}[X]| > \lambda] \leq \frac{var[X]}{\lambda^2}$$

Lemma 3 (The Chernoff Bound: Upper bound)**.** *Let $X_1, X_2, ..., X_n$ be independent random variables taking values in $\{0, 1\}$ with $\mathbb{E}[X_i] = p_i$. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$. Then the following holds*

1. *For any $\delta > 0$,*
$$\Pr[X \geq (1+\delta)\mu] < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$
2. *For $0 < \delta \leq 1$,*
$$\Pr[X \geq (1+\delta)\mu] \leq e^{-\frac{\mu\delta^2}{3}}$$

Lemma 4 (The Chernoff Bound: Lower bound)**.** *Let $X_1, X_2, ..., X_n$ be independent random variables taking values in $\{0, 1\}$ with $\mathbb{E}[X_i] = p_i$. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X]$. Then the following holds*

1. *For any $\delta > 0$,*
$$\Pr[X \leq (1-\delta)\mu] < \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$$
2. *For $0 < \delta \leq 1$,*
$$\Pr[X \leq (1-\delta)\mu] \leq e^{-\frac{\mu\delta^2}{2}}$$

2 Estimating a Parameter from a Sample

Motivating Examples:

- Estimating gene mutation: We are interested in evaluating the probability that a particular gene mutation occurs in the population.
- Popular query: We are interested in estimating the number of users searching for iPhone 7 release date in Google.
- Popular item: We are interested in the number of Amazon.com shoppers buying a particular beauty product in the last month.

Challenges:

- Given a DNA sample, a lab test can determine if it carries the mutation. However, the test is expensive and we would like to obtain a relatively reliable estimate from a small number of samples.
- We can examine the query log of every user to determine the total count of searches on iPhone 7 release date. However, that will require a huge amount time for processing.
- We can examine the items purchased for every user in the last month to find the number of users buying a particular beauty product. This will again require a long processing time.

In all the above scenarios, we would like to get a good estimate of the parameter of interest using a relatively small number of samples.

2.1 How large a sample shall we take?

Let p be the unknown probability that a particular gene mutates, or a particular user searches for iPhone 7 etc. We would like to estimate p.

Suppose we have taken n samples independently and uniformly from the entire pool of data. Let $\hat{n} = n\hat{p}$ be the number of samples having the desired property (mutation, or query with iPhone 7).

Given a sufficiently large number of samples, we expect that $\hat{p}$ to be close to the actual probability p. In particular, given a δ and aγ, we would like to obtain the minimum value of n such that

$$\Pr[\hat{p} \in [p - \delta, p + \delta]] \geq (1 - \gamma)$$

that is, we have sufficient confidence at least $(1 - \gamma)$ that our estimated probability $\hat{p}$ is not too far from the actual probability p.

2.1.1 Analysis

Define an indicator random variable X_i which is 1 if the ith sample has the desired property and 0 otherwise. Then

$$X = \sum_{i=1}^{n} X_i = n\hat{p}$$

denotes the number of sampled instances that have the desired property.

X is a sum of n independent random variables taking values in $0, 1$. We have $\mathbf{E}[X] = np$. Note that, we cannot calculate $\mathbf{E}[X]$ as we do not know p. Then by the Chernoff bound,

$$\begin{aligned}\Pr(\hat{p} < p - \delta) &= \Pr(n\hat{p} < np - n\delta) = \Pr(X < np(1 - \frac{\delta}{p})) \\ &= \Pr(X < E[X](1 - \frac{\delta}{p})) \leq e^{-E[X]\frac{\delta^2}{2p^2}} = e^{-n\frac{\delta^2}{2p}}\end{aligned}$$

Similarly,

$$\begin{aligned}\Pr(\hat{p} > p + \delta) &= \Pr(n\hat{p} > np + n\delta) = \Pr(X > np(1 + \frac{\delta}{p})) \\ &= \Pr(X > E[X](1 + \frac{\delta}{p})) \leq e^{-E[X]\frac{\delta^2}{3p^2}} = e^{-n\frac{\delta^2}{3p}}\end{aligned}$$

Therefore, by union bound,

$$\Pr[\hat{p} \notin [p-\delta, p+\delta]] \leq e^{-n\frac{\delta^2}{2p}} + e^{-n\frac{\delta^2}{3p}}$$

Since, we do not know p, let us put the trivial upper bound of $p \leq 1$ in the above equation. Then, we get

$$\Pr[\hat{p} \notin [p-\delta, p+\delta]] \leq e^{-n\frac{\delta^2}{2}} + e^{-n\frac{\delta^2}{3}} \leq 2e^{-n\frac{\delta^2}{3}}$$

Setting $\gamma = 2e^{-n\frac{\delta^2}{3}}$, we obtain a trade-off between δ, n and the confidence parameter γ.

Example 1. *If $\delta = \frac{1}{5}$ and $\gamma \leq \frac{1}{10}$, then we must have $e^{-\frac{n}{75}} \leq \frac{1}{20}$, or $n \geq 75 \ln 20 < 224$*

3 Extensions of Reservoir Sampling

Recall the *reservoir sampling* problem that we learnt in the first class. We have to sample s elements uniformly at random from $1, 2, ..., N$ where N is unknown. The algorithm was simple.

- Maintain the first s items $a_1, a_2, .., a_s$ in the reservoir.
- For $t = s+1, ...$
  - Sample the t element with probability $\frac{s}{t}$.
    - If the t element is sampled then
      - Select a position j in $\{1, 2, .., k\}$ uniformly at random and replace the jth element in the reservoir with the newly sampled t-th element.

Exercise 1. *What is the expected number of insertions in the reservoir?*

Drawback. The above algorithm is extremely sequential. It processes one element at a time. Can we obtain a faster distributed algorithm? For example, such an algorithm in used in *Cloudera ML*, an open-source collection of data preparation and machine learning algorithms for Hadoop.

3.1 Distributed Reservoir Sampling

For every item a_i select a random number R_i chosen uniformly at random (from the uniform distribution) from $[0, 1]$, and keep the s elements that have the largest values.

Can be implemented in the MapReduce framework (to be discussed later).

As a sequential algorithm, it has higher update time. You may need to maintain a min-heap to extract and compare with the element in the reservoir with minimum value.

3.1.1 Analysis

Lemma 5. *Consider two random variables U_1 and U_2 chosen according to uniform distribution from $[0, 1]$, then $\Pr(U_1 \leq U_2) = \frac{1}{2}$.*

Proof.

$$\Pr(U_1 \leq U_2) = \int_{U_2=0}^{1} \int_{U_1=0}^{U_2} dU_1 dU_2 = \int_{U_2=0}^{1} U_2 dU_2 = \frac{1}{2}$$

□

Lemma 6. *Consider n random variables $U_1, U_2, ..., U_n$ chosen according to uniform distribution from $[0,1]$, then $\Pr(U_1 \leq U_2 \leq .... \leq U_n) = \frac{1}{n!}$.*

Proof.

$$\Pr(U_1 \leq U_2 \leq .... \leq U_n) = \int_{U_n=0}^{1} \int_{U_{n-1}=0}^{U_n} .... \int_{U_2=0}^{U_3} \int_{U_1=0}^{U_2} dU_1 dU_2 ... dU_n = \frac{1}{n!}$$

□

Therefore, if we arrange the items in non-decreasing order according to the value of the associated random variable, we get a random permutation.

Hence, the probability that an item is in the last s positions (indicating s largest values) is

$$\Pr(\text{item } a_i \text{ is in the last } s \text{ positions}) = \frac{s(n-1)!}{n!} = \frac{s}{n}.$$

3.2 Weighted Reservoir Sampling

In the weighted reservoir sample, every item in the set has an associated weight, and we want to sample such that the probability that an item is selected is proportional to its weight. Therefore, if item i has weight w_i and there are N items with N being unknown, we want that the ith item is selected with probability proportional to $\frac{w_i}{W}$ where $W = \sum_{i=1}^{N} w_i$.

The weighted reservoir sampling is based on the same idea as the distributed reservoir sampling algorithm described above. For each item i in the stream, we compute a score as follows: first, generate a random number U_i between 0 and 1 following the uniform distribution, and then take the w_ith root of U_i. Return the s items with the highest score as the sample. Items with higher weights will tend to have scores that are closer to 1, and are thus more likely to be picked than items with smaller weights.

The analysis of the algorithm is based on the following lemma.

Lemma 7. *Consider two random variables U_1 and U_2 chosen according to uniform distribution from $[0,1]$. For some $w_1, w_2 > 0$, set $X_1 = U_1^{\frac{1}{w_1}}$ and $X_2 = U_2^{\frac{1}{w_2}}$ then $\Pr(X_1 \leq X_2) = \frac{w_2}{w_1+w_2}$.*

Proof.

$$\begin{aligned}
\Pr(X_1 \leq X_2) &= \Pr(U_1^{\frac{1}{w_1}} \leq U_2^{\frac{1}{w_2}}) \\
&= \Pr(U_1 \leq U_2^{\frac{w_1}{w_2}}) \\
&= \int_{U_2=0}^{1} \int_{U_1=0}^{U_2^{\frac{w_1}{w_2}}} dU_1 dU_2 \\
&= \int_{U_2=0}^{1} U_2^{\frac{w_1}{w_2}} dU_2 \\
&= \frac{w_2}{w_1 + w_2}
\end{aligned}$$

□

Exercise 2. *Complete the proof of the weighted reservoir sampling.*