Priority Sampling

Lecture 20

April 4, 2019

Sampling for data reduction

- X set of n points in the plane $a_1, a_2, \ldots, a_n$.
- Want to answer queries of the form: given some shape C (say circles), how many point inside C?
- standard data structures or brute force linear search say

Sampling for data reduction

- X set of n points in the plane $a_1, a_2, \ldots, a_n$.
- Want to answer queries of the form: given some shape C (say circles), how many point inside C?
- standard data structures or brute force linear search say

Question: Suppose n is too large and we can only store k points for some $k < n$.

Sampling approach:

- S sample of size k (with replacement). Store only S
- Given query C, compute $|C \cap S|$. What should we report as an estimate for $|C \cap X|$?

Sampling for data reduction

- X set of n points in the plane $a_1, a_2, \ldots, a_n$.
- Want to answer queries of the form: given some shape C (say circles), how many point inside C?
- standard data structures or brute force linear search say

Question: Suppose n is too large and we can only store k points for some $k < n$.

Sampling approach:

- S sample of size k (with replacement). Store only S
- Given query C, compute $|C \cap S|$. What should we report as an estimate for $|C \cap X|$? $\frac{n}{k}|S \cap X|$ which is an unbiased estimator

Weighted case

- X set of n points in the plane $a_1, a_2, \ldots, a_n$. Each point a_i has a non-negative weight w_i
- Want to answer queries of the form: given some shape C (say circles), what is weight of point inside C?

Question: Suppose n is too large and we can only store k points for some $k < n$.

Sampling approach?

Weighted case

- X set of n points in the plane $a_1, a_2, \ldots, a_n$. Each point a_i has a non-negative weight w_i
- Want to answer queries of the form: given some shape C (say circles), what is weight of point inside C?

Question: Suppose n is too large and we can only store k points for some $k < n$.

Sampling approach?

- Easy to see that uniform sampling is not ideal
- Sample in proportion to weight? Say a_i sampled with $p_i = w_i/W$ where $W = \sum_i w_i$.
- What do we set the weight of the sampled points to? Can we control sample size? What is the variance?

Importance Sampling

- Decide sampling probabilities $p_1, p_2, \ldots, p_n$
- Choose a_i independently with probability p_i and if i is chosen set $\hat{w}_i = w_i/p_i$. If i is not chosen we implicitly set $\hat{w}_i = 0$.

Importance Sampling

- Decide sampling probabilities $p_1, p_2, \ldots, p_n$
- Choose a_i independently with probability p_i and if i is chosen set $\hat{w}_i = w_i / p_i$. If i is not chosen we implicitly set $\hat{w}_i = 0$.
- For any i, $\mathbf{E}[\hat{w}_i] = w_i$.

Importance Sampling

- Decide sampling probabilities $p_1, p_2, \ldots, p_n$
- Choose a_i independently with probability p_i and if i is chosen set $\hat{w}_i = w_i / p_i$. If i is not chosen we implicitly set $\hat{w}_i = 0$.
- For any i, $\mathbf{E}[\hat{w}_i] = w_i$. Hence for any C, $\mathbf{E}[\hat{w}(C \cap S)] = \mathbf{E}[w(C \cap S)]$.

Importance Sampling

- Decide sampling probabilities $p_1, p_2, \ldots, p_n$
- Choose a_i independently with probability p_i and if i is chosen set $\hat{w}_i = w_i/p_i$. If i is not chosen we implicitly set $\hat{w}_i = 0$.
- For any i, $\mathbf{E}[\hat{w}_i] = w_i$. Hence for any C, $\mathbf{E}[\hat{w}(C \cap S)] = \mathbf{E}[w(C \cap S)]$.

Question: How should we choose p_i's?

Importance Sampling

- Decide sampling probabilities $p_1, p_2, \ldots, p_n$
- Choose a_i independently with probability p_i and if i is chosen set $\hat{w}_i = w_i / p_i$. If i is not chosen we implicitly set $\hat{w}_i = 0$.
- For any i, $\mathbf{E}[\hat{w}_i] = w_i$. Hence for any C, $\mathbf{E}[\hat{w}(C \cap S)] = \mathbf{E}[w(C \cap S)]$.

Question: How should we choose p_i's?

- Choose to reduce variance for queries of interest (depends on queries)
- Expected number of chosen points is $\sum_i p_i$ and hence choose p_i's to roughly meet the memory bound.

Importance Sampling in Streaming Setting

Setting:

- points $a_1, \ldots, a_n$ with weights arriving in stream
- have a memory size of k
- want to maintain a k-sample (to utilize memory as well as possible) such that we can estimate $w(C \cap X)$ accurately

Priority Sampling

[Duffield,Lund,Thorup]

- Queries are arbitrary subset sums so no structure there to exploit
- Focus on streaming aspect and using memory

Priority Sampling

[Duffield,Lund,Thorup]

- Queries are arbitrary subset sums so no structure there to exploit
- Focus on streaming aspect and using memory

Scheme:

1. For each $i \in [n]$ set priority $q_i = w_i/u_i$ where u_i is chosen uniformly (and independently from other items) at random from $[0, 1]$.
2. S is the set of items with the k highest priorities.
3. τ is the $(k+1)$'st highest priority. If $k \geq n$ we set $\tau = 0$.
4. If $i \in S$, set $\hat{w}_i = \max\{w_i, \tau\}$, else set $\hat{w}_i = 0$.

Priority Sampling

[Duffield,Lund,Thorup]

- Queries are arbitrary subset sums so no structure there to exploit
- Focus on streaming aspect and using memory

Scheme:

1. For each $i \in [n]$ set priority $q_i = w_i/u_i$ where u_i is chosen uniformly (and independently from other items) at random from $[0, 1]$.
2. S is the set of items with the k highest priorities.
3. τ is the $(k+1)$'st highest priority. If $k \geq n$ we set $\tau = 0$.
4. If $i \in S$, set $\hat{w}_i = \max\{w_i, \tau\}$, else set $\hat{w}_i = 0$.

Claim: Can maintain S, τ in streaming setting

Priority Sampling

Intuition: from uniform weight case

- Suppose $w_i = 1$ for all i. Then sampling k without repetition can be done via adaptation of reservoir sampling.
- A different approach: pick a uniformly random $r_i \in [0, 1]$ for each i. And pick top k in terms of r_i values (simulates random permutation) but can be done in streaming fashion. Many other distributions would work too and picking top k according to $1/r_i$ works too.
- Why $1/r_i$? What is the expected value of τ?

Priority Sampling: Properties

Lemma

$\mathbf{E}[\hat{w}_i] = w_i$.

Priority Sampling: Properties

Lemma

$\mathbf{E}[\hat{w}_i] = w_i$.

Fix i. Let $A(\tau')$ be the event that the k'th highest priority among items $j \neq i$ is τ'. Note that $i \in S$ if $q_i = w_i/u_i \geq \tau'$ and if $i \in S$ then $\hat{w}_i = \max\{w_i, \tau'\}$, otherwise $\hat{w}_i = 0$. To evaluate $\Pr[i \in S \mid A(\tau')]$ we consider two cases.

Case 1: $w_i \geq \tau'$. Here we have $\Pr[i \in S \mid A(\tau')] = 1$ and $\hat{w}_i = w_i$.

Case 2: $w_i < \tau'$. Then $\Pr[i \in S \mid A(\tau')] = \frac{w_i}{\tau'}$ and $\hat{w}_i = \tau'$.

In both cases we see that $E[\hat{w}_i] = w_i$.

Variance

Lemma

$$Var[\hat{w}_i] = \mathbf{E}[\hat{v}_i] \text{ where } \hat{v}_i = \begin{cases} \tau \max\{0, \tau - w_i\} & \text{if } i \in S \\ 0 & \text{if } i \notin S \end{cases}$$

Variance

Lemma

$$Var[\hat{w}_i] = \mathbf{E}[\hat{v}_i] \text{ where } \hat{v}_i = \begin{cases} \tau \max\{0, \tau - w_i\} & \text{if } i \in S \\ 0 & \text{if } i \notin S \end{cases}$$

Fix i. We define $A(\tau')$ to be the event that τ' is the k'th highest priority among elements $j \neq i$.

Show that

$$E[\hat{v}_i \mid A(\tau')] = E[\hat{w}_i^2 \mid A(\tau')] - w_i^2.$$

Since u_i is independent of τ' we can remove conditioning

Variance

$$E[\hat{v}_i \mid A(\tau')] = E[\hat{w}_i^2 \mid A(\tau')] - w_i^2.$$

$$\begin{aligned}
\mathbf{E}[\hat{v}_i \mid A(\tau')] &= \Pr[i \in S \mid A(\tau')] \times \mathbf{E}[\hat{v}_i \mid i \in S \wedge A(\tau')] \\
&= \min\{1, w_i/\tau'\} \times \tau' \max\{0, \tau' - w_i\} \\
&= \max\{0, w_i\tau' - w_i^2\}.
\end{aligned}$$

Variance

$$E[\hat{v}_i \mid A(\tau')] = E[\hat{w}_i^2 \mid A(\tau')] - w_i^2.$$

$$\begin{aligned}
\mathbf{E}[\hat{v}_i \mid A(\tau')] &= \Pr[i \in S \mid A(\tau')] \times \mathbf{E}[\hat{v}_i \mid i \in S \wedge A(\tau')] \\
&= \min\{1, w_i/\tau'\} \times \tau' \max\{0, \tau' - w_i\} \\
&= \max\{0, w_i\tau' - w_i^2\}.
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}\left[\hat{w}_i^2 \mid A(\tau')\right] &= \Pr[i \in S \mid A(\tau')] \times \mathbf{E}\left[\hat{w}_i^2 \mid i \in S \wedge A(\tau')\right] \\
&= \min\{1, w_i/\tau'\} \times (\max\{w_i, \tau'\})^2 \\
&= \max\{w_i^2, w_i\tau'\}.
\end{aligned}$$

Variance of subset sum

Lemma

If $k \geq 2$ *for any* $i \neq j$, $\mathbf{E}[\hat{w}_i \hat{w}_j] = w_i w_j$.

More generally

Lemma

Fix any set $C \subset [n]$. $\mathbf{E}\left[\prod_{i \in C} \hat{w}_i\right] = \prod_{i \in C} w_i$ *if* $|C| \leq k$ *and is* $\mathbf{0}$ *if* $|C| > k$.

Variance of subset sum

Lemma

If $k \geq 2$ *for any* $i \neq j$, $\mathbf{E}[\hat{w}_i \hat{w}_j] = w_i w_j$.

More generally

Lemma

Fix any set $C \subset [n]$. $\mathbf{E}\left[\prod_{i \in C} \hat{w}_i\right] = \prod_{i \in C} w_i$ *if* $|C| \leq k$ *and is* $\mathbf{0}$ *if* $|C| > k$.

Requires a proof by induction. See notes

Variance of subset sum

Lemma

If $k \geq 2$ *for any* $i \neq j$, $\mathbf{E}[\hat{w}_i \hat{w}_j] = w_i w_j$.

More generally

Lemma

Fix any set $C \subset [n]$. $\mathbf{E}\left[\prod_{i \in C} \hat{w}_i\right] = \prod_{i \in C} w_i$ *if* $|C| \leq k$ *and is* $\mathbf{0}$ *if* $|C| > k$.

Requires a proof by induction. See notes

Why is this interesting/non-obvious? In vanilla importance sampling the variables $\hat{w}_i$ are independent. However, here the variables are correlated because we choose exactly k. Nevertheless, they exhibit properties similar to independence.

Variance of subset sum

Lemma

If $k \geq 2$ for any $i \neq j$, $\mathbf{E}[\hat{w}_i \hat{w}_j] = w_i w_j$.

Consequence:

- Fix C. Unbiased estimator of $w(C \cap X)$ is $\hat{w}(C \cap S)$.
- Can we know the variance of the estimate to know if we are doing ok?
- $Var[\hat{w}(C \cap X)] = \sum_{i \in C \cap S} Var[\hat{w}_i] = \sum_{i \in C \cap S} \mathbf{E}[v_i]$. Hence, storing τ and $\hat{w}_i$ values suffices to estimate the variance of the estimate.