0368-3248-01-Algorithms in Data Mining **Fall 2013**

Lecture 4: Home Assignment, Due Dec 3rd

Lecturer: Edo Liberty

Warning: *This note may contain typos and other inaccuracies which are usually discussed during class. Please do not cite this note as a reliable source. If you find mistakes, please inform me.*

---

1 Probabilistic inequalities

setup

In this question you will be asked to derive the three most used probabilistic inequalities for a specific random variable. Let $x_1, \ldots, x_n$ be independent $\{-1, 1\}$ valued random variables. Each x_i takes the value 1 with probability 1/2 and -1 else. Let $X = \sum_{i=1}^{n} x_i$.

questions

1. Let the random variable Y be defined as $Y = |X|$. Prove that Markov's inequality holds for Y. Hint: note that Y takes integer values. Also, there is no need to compute $\Pr[Y = i]$.

2. Prove Chebyshev's inequality for the above random variable X. You can use the fact that Markov's inequality holds for any positive variable regardless of your success (or lack of if) in the previous question. Hint: $\mathrm{Var}[X] = E[(X - E[X])^2]$.

3. Argue that
$$\Pr[X > a] = \Pr[\Pi_{i=1}^{n} e^{\lambda x_i} > e^{\lambda a}] \leq \frac{E[\Pi_{i=1}^{n} e^{\lambda x_i}]}{e^{\lambda a}}$$
for any $\lambda \in [0, 1]$. Explain each transition.

4. Argue that:
$$\frac{E[\Pi_{i=1}^{n} e^{\lambda x_i}]}{e^{\lambda a}} = \frac{\Pi_{i=1}^{n} E[e^{\lambda x_i}]}{e^{\lambda a}} = \frac{(E[e^{\lambda x_1}])^n}{e^{\lambda a}}$$
What properties of the random variables x_i did you use in each transition?

5. Conclude that $\Pr[X > a] \leq e^{-\frac{a^2}{2n}}$ by showing that:
$$\exists\ \lambda \in [0, 1]\ \ s.t.\ \ \frac{(E[e^{\lambda x_1}])^n}{e^{\lambda a}} \leq e^{-\frac{a^2}{2n}}$$
Hint: For the hyperbolic cosine function we have $\cosh(x) = \frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2}$ for $x \in [0, 1]$.

answers

1.

$$\begin{aligned}
E[Y] &= \sum_{i=0}^{n} \Pr[Y=i]\cdot i \\
&= \sum_{i=0}^{t} \Pr[Y=i]\cdot i + \sum_{i=t+1}^{n} \Pr[Y=i]\cdot i \\
&\geq \sum_{i=t+1}^{n} \Pr[Y=i]\cdot i \\
&\geq \sum_{i=t+1}^{n} \Pr[Y=i]\cdot t \\
&= t\cdot \Pr[Y>t]
\end{aligned}$$

Therefore, $E[Y] \geq t\cdot \Pr[Y>t]$ which is Markov's inequality.

2. This is identical to the general proof of Chebyshev's inequality. We define $Z = (X - E[X])^2$. Since Z is positive we can use Markov's inequality for it and get:

$$\Pr[|X-E[X]|>t] = \Pr[Z>t^2] \leq \frac{E[Z]}{t^2} = \frac{\mathrm{Var}[X]}{t^2}$$

Here we used that $E[Z] = E[(X-E[X])^2] = \mathrm{Var}[X]$.

3. First transition:

$$\Pr[X>a] = \Pr[\lambda X > \lambda a] = \Pr[e^{\lambda X} > e^{\lambda a}] = \Pr[e^{\lambda \sum x_i} > e^{\lambda a}] = \Pr[\Pi_{i=1}^{n} e^{\lambda x_i} > e^{\lambda a}]$$

These hold due to the monotonicity of multiplication by a positive constant and exponentiation. Now, using Markov's inequality on the last inequality we get:

$$\Pr[\Pi_{i=1}^{n} e^{\lambda x_i} > e^{\lambda a}] \leq \frac{E[\Pi_{i=1}^{n} e^{\lambda x_i}]}{e^{\lambda a}}$$

4. The first transition is true due to the independence of the variables x_i. This means that $e^{\lambda x_i}$ are independent. The second transition is due to all expectations of $e^{\lambda x_i}$ being equal which stems from x_i being identically distributed.

5. First, we compute the expectation of $e^{\lambda x_i}$

$$E[e^{\lambda x_i}] = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} = \cosh(\lambda) \leq e^{\lambda^2/2}$$

From the above we have that $\Pr[X>a] \leq e^{n\lambda^2/2-\lambda a}$. Setting $\lambda = a/n$ we get $e^{n\lambda^2/2-\lambda a} = e^{-\frac{a^2}{2n}}$ which concludes the proof.

2 Approximating the size of a graph

setup

In this question we will try to approximate the size of a graph. A graph $G(V, E)$ is a set of nodes $|V| = n$ and a set of edges $|E| = m$. Each edge $e \in V \times V$ is a set of two nodes which support it. We assume the graph is simple which means there are no duplicate edges and no self loops (i.e. an edge $e = (u, u)$). The degree of a node, $\deg(u)$, is the number of edges which it supports. More formally $\deg(u) = |\{e \in E | u \in e\}|$. The degree of each node in the graph is at least 1. The question refers to the following sampling procedure:

1. $e = (u, v) \leftarrow$ an edge uniformly at random from E.
2. with probability 1/2
3.   return u
4. else
5.   return v

Throughout this question we assume that *i*) we can sample edges uniformly from the graph *ii*) that the number of edges m in known *iii*) that given a node u we can easily compute $\deg(u)$. The value of n, however, is unknown.

questions

1. Let $p(u)$ denote the probability that the sampling procedure returns a specific node, u. Compute $p(u)$ as a function of $\deg(u)$ and m. (Note: $\sum_{u \in V} \deg(u) = 2m$)

2. Let $f(u) = \frac{2m}{\deg(u)}$. Compute:
$$E_{x \sim smp}[f(x)]$$
where $x \sim smp$ denotes that x is chosen according to the distribution on the nodes generated by the above sampling procedure.

3. We say that a graph is d-degree-bounded if $\max_{u \in V} \deg(u) \leq d$. Show that for a d-degree-bounded graph:
$$\text{Var}_{x \sim smp}[f(x)] \leq dn^2$$

4. Let $Y = \frac{1}{s} \sum_{i=1}^{s} f(x_i)$ where x_i are nodes chosen independently from the graph according to the above sampling procedure. Compute $E[Y]$ **and** show that $\text{Var}[Y] \leq dn^2/s$.

5. Use Chebyshev's inequality to find a value for s such that for any d-degree-bounded graph and any two constants $\varepsilon \in [0, 1]$ and $\delta \in [0, 1]$:
$$\Pr[|Y - n| > \varepsilon n] < \delta.$$
s should be a function of d, ε and δ.

answers

1. A node is chosen only if an edge it is adjacent to is picked with probability and then it is the node picked between the two. The first event happens with probability $deg(u)/m$ since the edges a re chosen uniformly at random. The second event happens with probability 1/2 independently of the first event. This gives $p(u) = \frac{deg(u)}{m}\frac{deg(2)}{2} = \frac{deg(u)}{2m}$.

2. By the definition to the expectation:

$$E_{x\sim smp}[f(x)] = \sum_{u\in V} p(u)f(u) = \sum_{u\in V} \frac{deg(u)}{2m}\frac{2m}{\deg(u)} = \sum_{u\in V} 1 = n$$

3. We say that a graph is d-degree-bounded if $\max_{u\in V}\deg(u) \leq d$. Show that for a d-degree-bounded graph:

$$\mathrm{Var}_{x\sim smp}[f(x)] \leq E_{x\sim smp}[f^2(x)] = \sum_{u\in V} \frac{deg(u)}{2m}(\frac{2m}{\deg(u)})^2 = \sum_{u\in V} \frac{2m}{\deg(u)}$$

   Since $deg(u) \geq 1$ then $\sum_{u\in V} \frac{2m}{\deg(u)} \leq \sum_{u\in V} \frac{2m}{1} = 2mn$. Also, since the graph is d-degree-bounded $2m = \sum_{u\in V} deg(u) \leq nd$ thus $2mn \leq dn^2$.

4. Y is the average of s independent copies of $f(x)$ and therefore, by linearity of the expectation, we have that $E[Y] = E[f] = n$. Moreover, Since the nodes x_i are chosen independently we have that $\mathrm{Var}[Y] = \frac{1}{s^2}\sum_{i=1}^{s}\mathrm{Var}[f(x_i)]$. Since $f(x_i)$ distribute identically and substituting $\mathrm{Var}(x) \leq dn^2$ we get $\frac{1}{s^2}\sum_{i=1}^{s}\mathrm{Var}[f(x_i)] \leq \frac{s}{s^2}dn^2 = dn^2/s$.

5. Since $E[Y] = n$ we get that the above holds if

$$\Pr[|Y - E[n]| > \varepsilon n] < \frac{\mathrm{Var}[Y]}{\varepsilon^2 n^2} \leq \frac{dn^2/s}{\varepsilon^2 n^2} = \frac{d}{s\varepsilon^2}$$

   The condition that $\frac{d}{s\varepsilon^2} \leq \delta$ holds for $s \geq \frac{d}{\delta\varepsilon^2}$

3 Approximate median

setup

Given a list A of n numbers $a_1, \ldots, a_n$, we define the rank of an element $r(a_i)$ as the number of elements which are smaller than it. For example, the smallest number has rank zero and the largest has rank $n-1$. Equal elements are ordered arbitrarily. The median of A is an element a such that $r(a) = n/2$ (rounded either up or down). An α-approximate-median is a number a such that:

$$n(1/2 - \alpha) \le r(a) \le n(1/2 + \alpha)$$

In this question we sample k elements uniformly at random *with replacement* from the list A. Let the samples be $\{x_1, \ldots, x_k\} = X$. You will be asked to show that the median of X is an α-approximate-median of A.

questions

1. What is the probability the a randomly chosen element x is such that:

$$r(x) > n(1/2 + \alpha)$$

2. Let us define $X_{>\alpha}$ as the set of samples whose rank is greater than $n(1/2 + \alpha)$. More precisely, $X_{>\alpha} = \{x_i \in X | r(x_i) > n(1/2 + \alpha)\}$. Similarly we define $X_{<\alpha} = \{x_i \in X | r(x_i) < n(1/2 - \alpha)\}$. Prove that if $|X_{>\alpha}| < k/2$ and $|X_{<\alpha}| < k/2$ then the median of X is an α-approximate-median of A.

3. Let $Z = |X_{>\alpha}|$. Find t for which:

$$\Pr[Z \ge k/2] = \Pr[Z \ge (1+t)E[Z]]$$

4. Bound from above the probability that $Z \ge k/2$ as tightly as possible. If you do so using a probabilistic inequality, justify your choice.

5. Compute the minimal value for k which will guarantee that $|X_{>\alpha}| < k/2$ **and** $|X_{<\alpha}| < k/2$ with probability at least $1 - \delta$.

answers

1. There are $n(1/2-\alpha)$ elements for which $r(x) > n(1/2+\alpha)$. Since the element is chosen uniformly, the probability of that happening is $(1/2-\alpha)$.

2. First we note that the median of X cannot be either in $X_{>\alpha}$ or in $X_{<\alpha}$. This is simply because each of them includes less than half of the elements in X. Moreover, by the definitions of $X_{>\alpha}$ and $X_{<\alpha}$ we have:
$$n(1/2-\alpha) \leq r(median(X)) \quad \text{and} \quad r(median(X)) \leq n(1/2+\alpha)$$
which means that $median(X)$ is an α-approximate-median of A.

3. Since the probability of a sample being in $X_{>\alpha}$ is exactly $1/2-\alpha$ and since we have k independent samples, $E[Z] = E[|X_{>\alpha}|] = k(1/2-\alpha)$. Solving for t we get
$$(1+t)E[Z] = k/2 \;\rightarrow\; (1+t)(1/2-\alpha) = 1/2 \quad \rightarrow \quad t = \frac{2\alpha}{1-2\alpha}$$

4. Since the value of Z is the sum of independent indicator variables we can apply Chernoff's inequality. Denoting $\mu = E[Z] = k(1/2-\alpha)$ and $t = \frac{2\alpha}{1-2\alpha}$ we have:
$$\Pr[Z \geq k/2] = \Pr[Z \geq (1+t)\mu] \leq e^{-\mu t^2/4}$$

5. Similarly to the the above we can argue that
$$\Pr[|X_{<\alpha}| \geq k/2] \leq e^{-\mu t^2/4}$$
From the union bound we have that the probability of the event that $|X_{<\alpha}| \geq k/2$ or that $|X_{>\alpha}| \geq k/2$ is at most the sum of their probabilities.
$$\Pr\left[|X_{<\alpha}| \geq k/2 \cup |X_{>\alpha}| \geq k/2\right] \leq \Pr\left[|X_{<\alpha}| \geq k/2\right] + \Pr[|X_{>\alpha}| \geq k/2] \leq 2e^{-\mu t^2/4}$$
Demanding that this failure probability is less than δ we guarantee success with probability at least $1-\delta$. Substituting $\mu = k(1/2-\alpha)$ and $t = \frac{2\alpha}{1-2\alpha}$ this is achieved for
$$2e^{-\mu t^2/4} < \delta \quad \rightarrow \quad k > \frac{4\log(2/\delta)(1/2-\alpha)}{\alpha^2}$$

4 Simple high capacity hashing

setup

In this question we try to evaluate the capacity of a special hash table. For simplicity, we assume that the hashed elements are a subset of $[N]$ ($[N]$ denots the set $\{1, \ldots, N\}$). The hash table consists of an array A of length n and L perfect hash functions $h_\ell : [N] \to [n]$. Throughout the exercise we assume the existence of perfect hash functions. That is, $\Pr[h(x) = i] = 1/n$ for all $x \in [N]$ and $i \in [n]$ independently of the values $h(x')$. For convenience we also assume that the entries in A are initialized to the value 0.

Algorithm 1 *Add*(x)

```
for ℓ ∈ [L] do
  if A[h_ℓ(x)] == 0 or A[h_ℓ(x)] == x then
    A[h_ℓ(x)] = x
    Return Success
  end if
end for
Return Fail
```

Algorithm 2 *Query*(x)

```
for ℓ ∈ [L] do
  if A[h_ℓ(x)] == x then
    Return True
  else if A[h_ℓ(x)] == 0 then
    Return False
  end if
end for
Return False
```

questions

1. Argue the correctness of the hashing scheme. a) If an element was **successfully** added to the table by $Add(x)$ it will be found by $Query(x)$. b) If an element was not added to the table by $Add(x)$ it will not be found by $Query(x)$.

2. Assume that exactly m cells in the array are occupied. That is, m cells contain values $A[j] > 0$ and for the rest $A[j] = 0$. Given a new element x which is in not stored in the hash table. What is the probability that location $h_1(x)$ in A is occupied.

3. What is the probability that procedure $Add(x)$ fails for an element x not in the hash table? (here we still assume there are exactly m elements already in the table)

4. Assume we start with an empty hash table and insert m elements one after the other. Use the union bound to get a value for L for which $Add(x)$ succeeds in **all** m element insertions with probability at least $1 - \delta$

5. Argue that the **expected** running time of both $Add(x)$ and $Query(x)$ is $O(1)$. That is, it does not depend on L.

answers

1. If $Add(x)$ returned "success" then for some ℓ we have $A[h_\ell(x)] = x$ and for any $\ell' < \ell$ it holds that $A[h_{\ell'}(x)] \notin \{0, x\}$. Therefore it will be found by $Query(x)$. Also, if x was not added than it cannot be found by $Query$ since it returns "True" only if $A[h_\ell(x)] = x$ for some ℓ.

2. Since x is was not added and since h_1 is a perfect hash function then $\Pr[h_1(x) = i] = 1/n$ for all $i \in [n]$. Since there are m occupied cells this sums to $\Pr[A[h_1(x)] > 0] = m/n$.

3. Add fails only if for each to the $\ell \in [L]$ hash functions $A[h_\ell(x)] > 0$. Since they are chosen independently of each other we have
$$\Pr[Add(x) \text{ fails}] = (m/n)^L$$

4. Using the union bound we have that $\Pr[fail] \le \sum_{i \in [m]}((i-1)/n)^L$. This is because there are at most $i-1$ elements in the hash table when we insert the i'th one. Computing this sum can be made simpler by bounding it with an integral.
$$\sum_{i \in [m]}((i-1)/n)^L \le \int_{t=1}^{m+1}((t-1)/n)^L dt = \int_{t=0}^{m}(t/n)^L dt = \frac{1}{L+1}(m/n)^{L+1}$$
That said, even a bound as simple as $m(m/n)^L$ would have sufficed. For the sake of simplicity let us use the latter. We obtain that the failure probability is $m(m/n)^L \le \delta$ if $L \ge \log(m/\delta)/\log(n/m)$. Note that the hash can contain millions of items and be at $\sim 80\%$ capacity and still $L \sim 100$.

5. Let us start with the expected running time of Add. Denote by $\ell^* = \min_\ell A[h_\ell(x)] = 0$. Clearly, the running is $O(\ell^*)$ since each lookup requires $O(1)$ time.
$$\mathbb{E}[\ell^*] = \sum_{\ell=1}^{L} \ell \Pr[\ell^* = \ell] \le \sum_{\ell=1}^{\infty} \ell (\frac{m}{n})^{\ell-1}(1 - \frac{m}{n}) = O(1)$$
This assumes the ratio between m and n is fixed. Regardless, this does not depend on L.

   Now we argue the same about $Query$. If x has been added then $Query(x)$ takes the same amount of time that $Add(x)$ did at the time of insertion. If x has not been added then $Query$ returns $False$ in the same amount of time it would have taken to run $Add(x)$. If both both cases it reduces the calculation above.