0368-3248-01-Algorithms in Data Mining **Fall 2011**

Lecture 13: Algorithms In Data Mining - Exam

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 Please do not cite this note as a reliable source.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.*

General Info

1. Solve 3 out of 4 questions.
2. Each correct answer is worth 33.3 points.
3. If you have solved more than three questions, please indicate which three you would like to be checked.
4. The exam's duration is 3 hours. If you need more time please ask the attending professor.
5. Good luck!

Useful facts

1. For any vector $x \in \mathbb{R}^d$ we define the p-norm of x as follows:
$$||x||_p = [\sum_{i=1}^{d} (x(i))^p]^{1/p}$$
2. **Markov's inequality:** For any *non-negative* random variable X:
$$\Pr[X > t] \le E[X]/t.$$
3. **Chebyshev's inequality:** For any random variable X:
$$\Pr[|X - E[X]| > t] \le \mathrm{Var}[X]/t^2.$$
4. **Chernoff's inequality:** Let $x_1, \ldots, x_n$ be independent $\{0, 1\}$ valued random variables. Each x_i takes the value 1 with probability p_i and 0 else. Let $X = \sum_{i=1}^{n} x_i$ and let $\mu = E[X] = \sum_{i=1}^{n} p_i$. Then:
$$\Pr[X > (1+\varepsilon)\mu] \le e^{-\mu\varepsilon^2/4}$$
$$\Pr[X < (1-\varepsilon)\mu] \le e^{-\mu\varepsilon^2/2}$$
Or in a another convenient form:
$$\Pr[|X - \mu| > \varepsilon\mu] \le 2e^{-\mu\varepsilon^2/4}$$
5. **Hoeffding's inequality:** Let $x_1, \ldots, x_n$ be independent random variables taking values in $\{+1, -1\}$ each with probability $1/2$, then:
$$\Pr[|\sum_{i=1}^{n} x_i a_i| > t] \le 2e^{-\frac{t^2}{\sum_{i=1}^{n} a_i^2}}.$$

6. For any $x \geq 2$ we have:
$$e^{-1} \geq (1 - \frac{1}{x})^x \geq \frac{2}{3}e^{-1}$$

7. For convenience:
$$\frac{3}{5} \leq 1 - e^{-1} \approx 0.632 \leq \frac{2}{3} \quad \text{and} \quad \frac{3}{4} \leq 1 - \frac{2}{3}e^{-1} \approx 0.754 \leq \frac{4}{5}$$

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: $\text{Var}[X] = E[(X - E[X])^2]$.

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 $x \in [0, 1]$.

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 value of m in known *iii*) that given a node u we can 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:
$$\mathrm{Var}_{x \sim smp}[f(x)] \leq dn^2$$
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 $\mathrm{Var}[Y] \leq dn^2/s$.
5. Use Chebyshev's inequality to find a value for s such that for 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.

3 Approximate median

setup

Given a list A of n numbers $a_1, \dots, a_n$, 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 $\{x_1, \dots, x_k\} = X$. 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 $1 - \delta$.

4 Soulmate search

setup

In this question you will be asked to derive a search algorithm for a unique nearest neighbor given a local sensitive hash function family. We assume a universe of n objects $x_1, \ldots, x_n$ and a distance function d. For any pair of points $0 \leq d(x_i, x_j) \leq 1$. Moreover, each point x_i has exactly one soulmate point x_j such that $d(x_i, x_j) \leq r$, r is known constant. For all other points in the universe $d(x_i, x_{j'}) > 2r$. You are given a family H of hash functions such that $\Pr_{h \sim H}[h(x_i) = h(x_j)] = \frac{1}{1+d(x_i,x_j)}$ for any pair ($h \sim H$ means that h is chosen uniformly from H). We also define a bucketing hash function g which accepts an element x and returns a list of hash values.

$$g(x) = [h_1(x), \ldots, h_k(x)]$$

where each of the hush functions $h_1, \ldots, h_k$ was chosen uniformly and independently from the family H. We say that x_i and x_j are in love if $g(x_i) = g(x_j)$.

questions

1. What is the probability of two points, whose distance is $d(x_i, x_j)$, falling in love?
2. Compute a value for k for which the probability that x_i and x_j who are not soulmates ($d(x_i, x_j) \geq 2r$) of falling in love is at most $1/n$. Or, find k for which the following holds:

$$\Pr[g(x_i) = g(x_j) \mid d(x_i, x_j) \geq 2r] \leq 1/n$$

3. For this value of k, what is the probability that x_i falls in love with her soulmate? That means $\Pr[g(x_i) = g(x_j) \mid d(x_i, x_j) \leq r]$. Help: you can use the approximation $\frac{\log(1+r)}{\log(1+2r)} \approx \frac{1}{2}$.
4. We now create m independent copies of g, $g_1, \ldots, g_m$. We say that x_i finds x_j if $g_\ell(x_i) = g_\ell(x_j)$ for at least one function g_ℓ. Give a bound on the value of m which insures that **all** x_i find their soulmates with probability at least $1 - \delta$?
5. Given the above value for m, bound from above the **expected** number of points x_j that x_i fell in love with which were not her soulmates.