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

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 Please do not cite this note as a reliable source.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 35 points and each question part is worth 7 points.
3. Since the maximal grade is 105, grades will potentially be rounded down to 100.
4. If you solved more than three questions, please indicate which three you would like to be checked.
5. The exam's duration is 3 hours. If you need more time please ask the attending professor.
6. 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:

$$\Pr[|X - \mu| > \varepsilon\mu] \le 2e^{-\mu\varepsilon^2/4}$$

5. For $z \in [0, 1]$ we have $e^z < 1 + z + z^2$ and that $1 + z^2 < e^{z^2}$

1 Probabilistic inequalities

setup

In this question you will be asked to derive a version of the Chernoff bound for sums of independent mean zero random variables. Let $X_1, \ldots, X_n$ be a independent random variables such that

$$|X_i| \leq R \text{ and } \mathbb{E}[X_i] = 0 \text{ and } \mathbb{E}[X_i^2] = \sigma_i^2$$

Let $\sigma^2 = \sum_{i=1}^n \sigma_i^2$ and $X = \sum_{i=1}^n X_i$.

questions

1. Prove that $\mathbb{E}[X] = 0$ and that $\mathbb{E}[X^2] = \sigma^2$

2. Argue that for any positive parameter $\lambda > 0$ we have:

$$\Pr[X > t] \leq \frac{\Pi_{i=1}^n \mathbb{E}[e^{\lambda X_i}]}{e^{\lambda t}}$$

   Explain each step in the derivation and indicate what properties of the random variables X_i are used.

3. Argue that for any $\lambda \in (0, 1/R]$

$$\mathbb{E}[e^{\lambda X_i}] \leq e^{\lambda^2 \sigma_i^2}$$

   Explain each step in the derivation and indicate what properties of the random variables X_i are used. **Hint:** you should use the fact that for any $z \leq 1$ we have $e^z < 1 + z + z^2$ and that $1 + z^2 < e^{z^2}$.

4. Conclude that for $\lambda = \min\{t/2\sigma^2, 1/R\}$ we obtain:

$$\begin{aligned} \Pr[X > t] &\leq e^{-t^2/4\sigma^2} \text{ if } t \leq 2\sigma^2/R & (1) \\ \Pr[X > t] &\leq e^{-t/2R} \text{ if } t \geq 2\sigma^2/R & (2) \end{aligned}$$

5. Let $Z_1, \ldots, Z_n$ be random indicator variables such that $Z_i = 1$ with probability p_i and $Z_i = 0$ else. Let $Z = \sum_{i=1}^n Z_i$ and $\mu = \mathbb{E}[Z] = \sum_{i=1}^n p_i$. Using the above, conclude that for $\varepsilon \in [0, 1]$:

$$\Pr[Z - \mu > \varepsilon\mu] \leq e^{-\mu\varepsilon^2/4}$$

   Hint: notice that you cannot apply the above directly because $\mathbb{E}[Z_i] = p_i \neq 0$. However the fact that $\mathbb{E}[Z_i - p_i] = 0$ might assist you.

2 Approximate item frequencies

setup

We are given a stream of n integers $a_1, \ldots, a_n \in [m]$. We define the frequency of item i to be the number of times i appeared in the stream. That is, $f_i = \sum_{j \in [n]} 1_{a_j == i}$. Assume we also hold an array b of ℓ counters, initially set to zero. Finally we assume a perfect hash function $h : [m] \to [\ell]$. The question will discuss the result of the following algorithm.

$b \leftarrow$ empty counters array of size ℓ initialized to 0.
$h \leftarrow$ perfect hash function from $1, \ldots, m$ to $1, \ldots, \ell$.
for $i \in a_1, \ldots, a_n$ **do**
  $b[h(i)] \leftarrow b[h(i)] + 1$
end for

Particularly, we will examine the relation between f_i and $b[h(i)]$. Throughout the question we will denote by $h^{-1}(i)$ the set of indexes that collide with i using the hash function h, namely, $h^{-1}(i) = \{j | h(i) = h(j)\} \setminus \{i\}$.

questions

1. What is the probability that $j \in h^{-1}(i)$ for $j \neq i$.
2. Show that by the end of the stream $b[h(i)] = f_i + \sum_{j \in h^{-1}(i)} f_j$.
3. Show that $\mathbb{E}[b[h(i)]] = f_i + (n - f_i)/\ell$.
4. Show that $\mathrm{Var}[b[h(i)]] \leq \sum_{j \in [m]} f_j^2 / \ell$.
5. Use Chebyshev's inequality and your results for 3 and 4 to find a value of ℓ such that

$$\Pr[b[h(i)] > f_i + \varepsilon n] \leq \delta$$

   It might help to notice that $n = \sum_{i \in [m]} f_i \geq \sqrt{\sum_{i \in [m]} f_i^2}$.

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 that 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 $r(a) = n/2$. An α-approximate-median is a number a such that:

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

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 for some value of k.

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:

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

(Hint: this is only an auxiliary step that is supposed to help you relate $k/2$ and $\mathbb{E}[Z]$ in a form similar to Chernoff's bound)

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

5. What value of k will guarantee that $|X_{>\alpha}| < k/2$ **and** $|X_{<\alpha}| < k/2$ simultaneously with probability at least $1 - \delta$?

4 k-means clustering with equal cluster sizes

setup

You are given n vectors $X = \{x_1, \ldots, x_n\}$ and $x_i \in \mathbb{R}^d$. The k-means cost function for X and a set of k centers $M = \{\mu_1, \ldots, \mu_k\}$ is defined as:

$$f(X, M) = \sum_{i=1}^{n} \min_j \|x_i - \mu_j\|^2 = \sum_{j=1}^{k} \sum_{i \in S_j} \|x_i - \mu_j\|^2$$

where $i \in S_j$ if the center closest to x_i is μ_j. From here on, we will denote by $\mathrm{OPT}(X, k)$ the lowest possible value of f using k clusters for the points X. Namely, $\mathrm{OPT}(X, k) = f(X, M^*)$ where $M^* = \arg\min_{|M|=k} f(X, M)$. Moreover, we denote by $\mathrm{ALG}(X, k)$ the cost of $f(X, M)$ where each center μ_j is picked uniformly at random (with replacement) from X. In other words, for every i and j we have $\Pr[\mu_j = x_i] = 1/n$. Note that $\mathrm{ALG}(X, k)$ is a random variable which depends on the choice of centers.

To make things simpler, we assume that the best solution is obtained with equal cluster sizes . That it $|S_j^*| = n/k$ for all j and n/k is an integer.

questions

1. We start with the case of $k = 1$. That is, there is only one center. In this case the optimal center has a closed form solution which is $\mu_1^* = \frac{1}{n} \sum_{i=1}^{n} x_i$. Show that

$$\mathrm{OPT}(X, 1) = \sum_{i=1}^{n} x_i^T x_i - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} x_i^T x_j$$

2. Compute $\mathbb{E}[ALG(X, 1)]$. Show that $\mathbb{E}[ALG(X, 1)] = 2\mathrm{OPT}(X, 1)$.

3. We now turn to the more interesting case where $k > 1$. Define E_{cover} to be the event that the algorithm picks exactly one point from each optimal cluster S_j^*. Show that $\Pr(E_{cover}) \geq e^{-k}$. **Hint:** you might find Stirling's formula useful: $\log(k!) = k \log(k) - k + O(\log(k))$.

4. Argue that given that E_{cover} happens the expected cost of the algorithm is low. That is

$$\mathbb{E}[\mathrm{ALG}(X, k) | E_{cover}] \leq 2\mathrm{OPT}(X, k).$$

5. Given the observation above, describe an algorithm whose running time is $O(e^k \log(1/\delta) dkn)$ that produced a set of centers M such that $f(X, M) \leq 4\mathrm{OPT}(X, k)$ with probability at least $1 - \delta$.