**0368-3248-01-Algorithms in Data Mining** **Fall 2012**

# Lecture 14: Algorithms in Data Mining - Exam Answers

*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 Please do not cite this note as a reliable source.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 part of a question 7.
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 $X$:

$$\Pr[X > t] \leq E[X]/t.$$

3. **Chebyshev's inequality:** For any random variable $X$:

$$\Pr[|X - E[X]| > t] \leq \text{Var}[X]/t^2.$$

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:

$$\begin{aligned} \Pr[X > (1 + \varepsilon)\mu] &\leq e^{-\mu\varepsilon^2/4} \\ \Pr[X < (1 - \varepsilon)\mu] &\leq e^{-\mu\varepsilon^2/2} \end{aligned}$$

Or in a another convenient form:

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

5. $sin(0.005°) \approx 9 \cdot 10^{-4}$, $cos(0.005°) \approx 1 - 4 \cdot 10^{-7}$

# 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 Number of stars in the sky

## setup

An enthusiast astronomer decides to count the number of stars in the sky (which are visible using her telescope). She is rather low-tech so exact counting is out of the question. Alas, she knows that the visual angle of her telescope is $0.01°$. She figures out that this information should suffice in order to estimate the correct answer. For simplicity, we assume she can point her telescope in any direction on the sphere, as if she is floating with her telescope in space.[1]

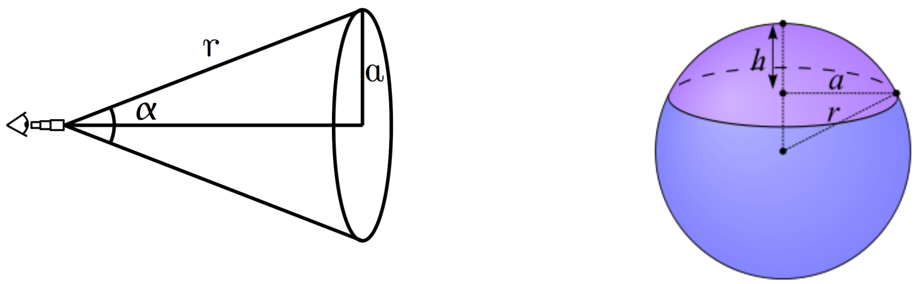


Figure 1: On the left: an illustration of the visual angle $\alpha$ of a telescope ($sin(\alpha/2) = a/r$). On the right: a spherical cap of base $a$ and hight $h$ on a sphere of radius $r$. The area of the spherical cap is $2\pi rh$. The surface area of the entire sphere is $4\pi r^2$.

## questions

1. What portion, $q$, of the sky does the telescope cover?
2. Let $x_i = 1$ if star $i$ is in the field of vision and zero else. If the telescope is pointed in a direction uniformly at random, what is the probability that $x_i = 1$. You can use the quantity $q$ above even if you did not mange to solve the first question.
3. Denote by $N$ the number of visible stars. Let $z$ denote the number of visible stars, compute $\mathbb{E}[z]$.
4. Assume that one can never see more than $100Nq$ stars simultaneously through the telescope. Show that $\text{Var}[z] \leq 100N^2q^2$
5. Define $Y = \frac{1}{s}\sum_{i=1}^{s}\frac{z_i}{q}$ where $z_i$ are independent star counts, each in an independent random direction. Compute a value for $s$ such that $\Pr[|Y - N| \geq \varepsilon N] \leq \delta$. The value of $s$ should depend on both $\varepsilon$ and $\delta$.

[1]This can actually be simulated by waiting for different parts of the day or year (you cannot look thorough the sun) but this discussion is irrelevant in this context.

## answers

1. The portion of the sky the telescope covers is exactly the ratio between the area of the sphere $4\pi r^2$ and the area of the spherical cap $2\pi rh$ defined by the telescope. To compute $h$ we use simple trigonometry and get $h = r - r \cdot cos(0.005) = r \cdot 4 \cdot 10^{-7}$ (according to the attached useful information). Finally $q = (2\pi rh)/(4\pi r^2) = 2 \cdot 10^{-7}$.

2. The probability that a star begin visible here is identical to the probability of it being visible in the scenario where the telescope is fixed and the star is located uniformly at random on the sphere. In this scenario it is clear that the probability is exactly $q = 2 \cdot 10^{-7}$.

3. Note that $z = \sum_{i=1}^{n} x_i$. Computing the expectation:

$$\mathbb{E}[z] = \mathbb{E}[\sum_{i=1}^{N} x_i] = \sum_{i=1}^{N} \mathbb{E}[x_i] = \sum_{i=1}^{N} q = Nq$$

## 3 Finding the number of users on facebook

### setup

Facebook does not publish exactly how many members it has. While the network does release official figures once in while there are good reasons to verify their reports. One thing Facebook does offer is a web interface which allows external applications to receive information about specific users. More specifically, given a user id the service returns some user data or not-a-user (if no such user exists). This question will lead you through estimating the correct number of users using only this fact.

### questions

1. Assume that the number of users is $N$ and that user ids are 32 bit integers. If one picks a 32 bit integer uniformly at random, what is the probability that it matches an existing user id. In other words, the returned response is not not-a-user.

2. Denote be $u(x)$ a function that takes the value $2^{32}$ if $x$ matches and existing user id and zero else. Let $x_i$ be an integer drawn uniformly at random from $1, \ldots, 2^{32}$. Compute the expected value of

$$Z = \frac{1}{s} \sum_{i=1}^{s} u(x_i)$$

3. Show that the random variable $Y = 2^{-32} sZ$ is a sum of independent indicator $\{0, 1\}$ variables.

4. Using the Chernoff bound, find a value for $s$ such that:

$$\Pr[|Z - N| \geq \varepsilon N] \leq \delta$$

   for given $\varepsilon, \delta > 0$.

5. Based on the fact that Facebook has more than $2^{29}$ users, would you consider this approach reasonable? Lately, facebook changed their user id format to be a 64 bit integer. Is this approach still reasonable?

## 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 $L$.