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

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 Please do not cite this note as a reliable source.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

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

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

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

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:

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

   Or in a another convenient form:

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

5. A probability distribution ψ of z is such that:

$$\forall\, c_1, c_2 \in \mathbb{R} \quad \Pr[c_1 \leq z \leq c_2] = \int_{c_1}^{c_2} \psi(t)dt$$

6. For a continuous variable z we have that:

$$\mathbb{E}[z] = \int_{-\infty}^{\infty} f(t)\psi(t)dt \qquad \mathrm{Var}[z] = \int_{-\infty}^{\infty} f^2(t)\psi(t)dt - (\mathbb{E}[z])^2$$

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 $x \in [0, 1]$.

2 Integrating blackbox functions

setup

Here, we will try to write an algorithm for approximately integrating blackbox functions. Given a function f, the algorithm must produce an approximation for the integral of f over a given range $[a, b]$. Alas, while it can evaluate $f(t)$ for any value of t, it does not have any notion of the inner workings of f. More precisely, the algorithm is given a range $[a, b] \in \mathbb{R}$ two parameters $\varepsilon, \delta > 0$ and a function f. It is required to produce a value $A = ALG(f, a, b, \varepsilon, \delta)$ such that with probability at least $1 - \delta$:

$$(1 - \varepsilon) \int_a^b f(t)dt \leq A \leq (1 + \varepsilon) \int_a^b f(t)dt \ .$$

To make things simpler, the function f is bounded both from below and from above, $\forall\ x\ 1 \leq f(x) \leq 2$. The questions will lead you through constructing this algorithm.

questions

1. Consider the variable x taking values uniformly at random over the range $[a, b]$. Write the equation for the probability distribution function ψ of x.
2. Prove that $\int_a^b f(t)dt = (b - a)\mathbb{E}[f(x)]$.
3. Show that $\mathrm{Var}[f(x)] \leq 3(\mathbb{E}[f(x)])^2$. Hint: remember that $f(x) \in [1, 2]$.
4. For an integer s, define $Y = \frac{1}{s} \sum_{i=1}^{s} f(x_i)$ where x_i are all chosen uniformly and i.i.d. from $[a, b]$. Compute $\mathbb{E}[Y]$ and show that $\mathrm{Var}[Y] \leq 3\mathbb{E}[Y]/s$.
5. Compute a value for s which guaranties that

   $$\Pr[|Y - \mathbb{E}[Y]| \geq \varepsilon\mathbb{E}[Y]] \leq \delta \ .$$

   Describe the resulting algorithm $ALG(f, a, b, \varepsilon, \delta)$ and argue that it meets the required conditions.

3 Matrix Sampling

setup

Consider an $m \times n$, $\{1, -1\}$ matrix A. More formally, $A \in \mathbb{R}^{m \times n}$ and $\forall\ i \in [m], j \in [n]\ A_{i,j} \in \{1, -1\}$. In this question we will try to compute an approximation for AA^T efficiently by sampling columns from A. Define a n i.i.d. random variables $q_1, \ldots, q_n$ such that:

$$q_i = \begin{cases} 1/\sqrt{p} & \text{w.p. } p \\ 0 & \text{otherwise} \end{cases}$$

for some fixed value $p \in [0, 1]$. The sampled matrix B is such that $B_{i,j} = A_{i,j} q_j$

questions

1. What is the expected number of non zero entries in the matrix B?
2. Let A_i denote the i'th row of A and similarly B_i. Argue that
$$\mathbb{E}[\langle B_{i_1}, B_{i_2} \rangle] = \langle A_{i_1}, A_{i_2} \rangle\ .$$
3. Use Chernoff's inequality to bound from above the following probability:
$$\Pr[|\langle B_{i_1}, B_{i_1} \rangle - \langle A_{i_1}, A_{i_1} \rangle| \geq \varepsilon n]$$
for a fixed $\varepsilon \in [0, 1]$. Note that $\langle A_{i_1}, A_{i_1} \rangle = n$.
4. Bound from above the following probability:
$$\Pr[|\langle B_{i_1}, B_{i_2} \rangle - \langle A_{i_1}, A_{i_2} \rangle| \geq \varepsilon n]$$
Hint: it is convenient to consider the sets $J^+ = \{j \mid A_{i_1,j} A_{i_2,j} = 1\}$ and $J^- = \{j \mid A_{i_1,j} A_{i_2,j} = -1\}$ and setting $n^+ = |J^+|$ and $n^- = |J^-|$.
5. Using the union bound, compute a value for p which guaranties that with probability at least $1 - \delta$ we have that:
$$\forall\ \ i_1, i_2 \in [m]\ \ |(BB^T)_{i_1,i_2} - (AA^T)_{i_1,i_2}| \leq \varepsilon n\ .$$

4 2-Means Clustering

setup

You are given n points $x_1 \ldots, x_n \in \mathbb{R}^d$ which naturally fall into two clusters. There exist two points $y_1, y_2 \in \mathbb{R}^d$ such that the distance between y_1 and y_2 is ℓ (that is $\|y_1 - y_2\|_2 = \ell$). There are $n/2$ points around y_1 such that $\|x_i - y_1\|_2 \leq 1$. The other $n/2$ points are around y_2 and $\|x_i - y_2\|_2 \leq 1$. Note that the points y_1 and y_2 are not known to you. Reminder: the cost of k-means clustering is $\min_{c_1,\ldots,c_k \in \mathbb{R}^d} \sum_{i=1}^{n} \min_{j \in [k]} \|x_i - c_j\|^2$.

questions

1. What is the cost of 2-means clustering when the two chosen cluster centers are $c_1 = y_1$ and $c_2 = y_2$?
2. Argue that if we pick as centers c_1, c_2 two points, one from each cluster, then the cost is at most $4n$.
3. Argue that if we pick as centers c_1, c_2 two points from the same cluster then the cost is at least $n/2(\ell - 2)^2$.
4. Assume that $\ell > 5$. Describe an algorithm for finding a clustering assignment whose cost is at most $2n$ with probability at least $1 - \delta$. Your algorithm's running time dependence on the number of points n must be linear.
5. Given the algorithm in the previous question, describe an algorithm for finding the *optimal cluster centers* with probability $1 - \delta$ and prove its correctness. (note: you are not asked to recover y_1 and y_2)