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

Lecture 13: 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 Please do not cite this note as a reliable source.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 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 $x \in [0, 1]$.

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 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 $\mathrm{Var}[Y] \leq 3(\mathbb{E}[Y])^2/s$.
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.

answers

1. The function $\phi(x)$ is a piecewise constant function. $\phi(x) = 1/(b-a)$ for $a \leq x \leq b$ and zero else.

2. To show this we simply compute the expectation of $f(x)$:

$$\mathbb{E}[f(x)] = \int_{-\infty}^{\infty} f(t)\psi(t)dt = \int_a^b f(t)\frac{1}{b-a}dt$$

   which gives $\int_a^b f(t)dt = (b-a)\mathbb{E}[f(x)]$

3. We compute the variance of $f(x)$.

$$\mathrm{Var}[f(x)] = \int_{-\infty}^{\infty} f^2(t)\psi(t)dt - (\mathbb{E}[f(x)])^2 \leq \int_a^b 4\frac{1}{b-a}dt - (\mathbb{E}[f(x)])^2 = 4 - (\mathbb{E}[f(x)])^2$$

   Now, since $f(x) \in [1,2]$ we also have that $\mathbb{E}[f(x)] \geq 1$. This means that $4 \leq 4(\mathbb{E}[f(x)])^2$ and gives that $\mathrm{Var}[f(x)] \leq 3(\mathbb{E}[f(x)])^2$.

4. First, from linearity of expectation and the fact that x_i are i.i.d. we have that:

$$\mathbb{E}[Y] = \mathbb{E}[\frac{1}{s}\sum_{i=1}^{s} f(x_i)] = \frac{1}{s}\sum_{i=1}^{s}\mathbb{E}[f(x_i)] = \frac{1}{s}\sum_{i=1}^{s}\mathbb{E}[f(x)] = \mathbb{E}[f(x)] \ .$$

   Since x_i are independante we have that the variance of the sum is the sum of variances.

$$\mathrm{Var}[Y] = \frac{1}{s^2}\sum_{i=1}^{s}\mathrm{Var}[f(x_i)] \leq \frac{1}{s^2}\sum_{i=1}^{s} 3(\mathbb{E}[Y])^2 = 3(\mathbb{E}[Y])^2/s \ .$$

5. Using Chebyshev's inequality we that that

$$\Pr[|Y - \mathbb{E}[Y]| \geq \varepsilon\mathbb{E}[Y]] \leq \frac{\mathrm{Var}[Y]}{(\varepsilon\mathbb{E}[Y])^2} \leq \frac{3}{\varepsilon^2 s} \ .$$

   The condition that $\frac{3}{\varepsilon^2 s} \leq \delta$ is met by $s \geq 3/\varepsilon^2\delta$. The resulting algorithm $ALG(f, a, b, \varepsilon, \delta)$ is trivial. Sample $s = 3/\varepsilon^2\delta$ different values $x_1, \ldots, x_s$ uniformly at random from the interval $[a, b]$ and evaluate $f(x_i)$ for each. Return the average of the evaluations $Y = \frac{1}{s}\sum_{i=1}^{s} f(x_i)$.

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 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:
$$\forall\ \ i_1, i_2 \in [m]\ \ |(BB^T)_{i_1,i_2} - (AA^T)_{i_1,i_2}| \leq \varepsilon n\ .$$

answers

1. The expected number of non zero variables q_i is np. Since for each of those there are m non zeros in B the answer is mnp.

2. Here we compute the expectation

$$\begin{aligned}\mathbb{E}[\langle B_{i_1}, B_{i_2}\rangle] &= \mathbb{E}[\sum_{j=1}^{n} B_{i_1}(j)B_{i_2}(j)] = \mathbb{E}[\sum_{j=1}^{n} A_{i_1}(j)A_{i_2}(j)q_j^2] \\ &= \sum_{j=1}^{n} A_{i_1}(j)A_{i_2}(j)\mathbb{E}[q_j^2] = \sum_{j=1}^{n} A_{i_1}(j)A_{i_2}(j) \\ &= \langle A_{i_1}, A_{i_2}\rangle\end{aligned}$$

3. Note that $\langle B_{i_1}, B_{i_1}\rangle = \sum_{j=1}^{n} A_{i_1}(j)A_{i_1}(j)q_j^2 = \frac{1}{p}\sum_{j=1}^{n} b_i$ where $b_i = 1$ w.p. p and zero else.

$$\Pr[|\frac{1}{p}\sum_{j=1}^{n} b_i - n| \geq \varepsilon n] = \Pr[|\sum_{j=1}^{n} b_i - pn| \geq p\varepsilon n]$$

   Since b_i are independent indicator variables we use Chernoff's bound.

$$\Pr[|\sum_{j=1}^{n} b_i - pn| \geq \varepsilon pn] \leq 2e^{-\frac{pn\varepsilon^2}{4}}$$

4. Consider the sets $J^+ = \{j \mid A_{i_1,j}A_{i_2,j} = 1\}$ and $J^- = \{j \mid A_{i_1,j}A_{i_2,j} = -1\}$. We have that:

$$\langle B_{i_1}, B_{i_2}\rangle = \sum_{j\in J^+} q_j^2 - \sum_{j\in J^-} q_j^2$$

   Moreover, setting $n^+ = |J^+|$ and $n^- = |J^-|$ we have $\langle A_{i_1}, A_{i_2}\rangle = n^+ - n^-$. Therefore,

$$\langle B_{i_1}, B_{i_2}\rangle - \langle A_{i_1}, A_{i_2}\rangle = (\sum_{j\in J^+} q_j^2 - n^+) - (\sum_{j\in J^-} q_j^2 - n^-)$$

   Thus, to have that $|\langle B_{i_1}, B_{i_2}\rangle - \langle A_{i_1}, A_{i_2}\rangle| \leq \varepsilon n$ it suffices to have $|\sum_{j\in J^+} q_j^2 - n^+| \leq \varepsilon n/2$ and $|\sum_{j\in J^-} q_j^2 - n^-| \leq \varepsilon n/2$. For each one of the above we can apply the Chernoff bound as above.

$$\Pr[|\sum_{j\in J^+} q_j^2 - n^+| \geq \varepsilon n/2] = \Pr[|\sum_{j=1}^{n^+} b_j - n^+p| \geq \varepsilon np/2] \leq 2e^{-\frac{(\varepsilon np/2)^2}{4n^+p}} \leq 2e^{-\frac{\varepsilon^2 np}{16}}$$

   which uses the fact that $n^+ \leq n$. Repeating the same exercise for n^- and using the Union bound we gets:

$$\Pr[|\langle B_{i_1}, B_{i_2}\rangle - \langle A_{i_1}, A_{i_2}\rangle| \geq \varepsilon n] \leq 4e^{-\frac{\varepsilon^2 np}{16}}$$

5. Using the union bound on all $\binom{m}{2} \leq m^2$ options of choosing i_1 and i_2 we require that:

$$m^2 4e^{-\frac{\varepsilon^2 np}{16}} \leq \delta \ .$$

   This is satisfied by $p \geq \frac{16}{4\varepsilon^2}\log(\frac{4m^2}{\delta})$

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 $4n$ 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)

answers

1. Each point x_i is of distance 1 from its center so the total cost is n.

2. Each point in the data is in the same ball (of radius 1) with either c_1 or c_2. Say x_i is in the same cluster as c_1. From the triangle inequality, we get that $\|x_i - c_1\| \leq \|x_i - y_1\| + \|y_i - c_1\| \leq 2$. Thus, $\|x_i - c_1\|^4 \leq 4$. Summing over all points we get that the total cost is at most $4n$.

3. If both c_1 and c_2 are in the same cluster than all the $n/2$ points on the other cluster (the one without centers) are at distance at least $\ell - 2$ from either c_1 or c_2. Summing over their costs we get $n/2(\ell-2)^2$.

4. Since $\ell > 5$ we have that $n/2(\ell-2)^2 \geq 4n$. This means that any clustering whose centers are split is better than any clustering whose centers are not split. Note that is we pick two points at random from the data as centers c_1 and c_2 with probability $1/2$ we pick them such that they are split. Moreover, after we pick centers we can compute the k-means cost and make sure that the cost is at most $4n$. If we fail, we can repeat this process until we succeed. It is easy to see that after at most $\log(1/\delta)$ we succeed in one of the rounds with probability at least $1 - \delta$.

5. First we describe the algorithm and then prove its correctness. We perform the previous procedure and obtain two cluster centers c_1 and c_2. Then we set s_1 and s_2 to be the sets of points whose closest center is c_1 and c_2 respectively. The algorithm returns $c_1^* = \frac{2}{n}\sum_{i \in s_1} x_i$ and similarly $c_2^* = \frac{2}{n}\sum_{i \in s_2} x_i$ In other words, the algorithm performs one iteration of Lloyd's algorithm initialized at centers c_1, c_2.

   The correctness stems from the following facts. First, the clusters s_1 and s_2 are optimal. This is because s_1 contains the $n/2$ points of distance at most 1 to y_1 and s_2 all the rest. It can be verified that no other clustering is better. Assume by contradiction that such a clustering exists. Note that it can be improved by moving one of the points between the clusters which contradicts optimality. Second, the optimal placement of centers for a given set of points is at their average. This was shown in class.