EPFL

Exercise Set X, Algorithms II 2024

These exercises are for your own benefit. Feel free to collaborate and share your answers with other students. Solve as many problems as you can and ask for help if you get stuck for too long. Problems marked * are more difficult but also more fun :).

These problems are taken from various sources at EPFL and on the Internet, too numerous to cite individually.

Locality Sensitive Hashing

1 *(Final exam question from a previous year)* **LSH for Jaccard similarity.**

Recall the Jaccard index that we saw in Exercises: Suppose we have a universe U. For non-empty sets $A, B \subseteq U$, the Jaccard index is defined as

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Design a locality sensitive hash (LSH) family $\mathcal{H}$ of functions $h : 2^U \to [0,1]$ such that for any non-empty sets $A, B \subseteq U$,

$$\Pr_{h\sim\mathcal{H}}[h(A) \neq h(B)] \begin{cases} \leq 0.01 & \text{if } J(A,B) \geq 0.99, \\ \geq 0.1 & \text{if } J(A,B) \leq 0.9. \end{cases}$$

(In this problem you are asked to explain the hash family and argue that it satisfies the above properties. Recall that you are allowed to refer to material covered in the course.)

Solution: Let us describe $\mathcal{H}$ by giving a procedure to sample an element $h \in \mathcal{H}$:

- for each $u \in U$, sample h_u uniformly at random from $[0,1]$.
- set $h(A) = \min_{u\in A} h_u$ for any non-empty $A \subseteq U$ (i.e., MinHashing).

In Exercise Set 7, we showed that $\Pr[h(A) = h(B)] = J(A,B)$. So $\Pr[h(A) \neq h(B)] = 1 - J(A,B)$ and the claimed bounds follow immediately.

2 In this problem we design an LSH for points in $\mathbb{R}^d$ with the ℓ_1 distance, i.e.

$$d(p,q) = \sum_{i=1}^{d} |p_i - q_i|.$$

Define a class of hash functions as follows: Fix a positive number w. Each hash function is defined via a choice of d independently selected random real numbers $s_1, s_2, \ldots, s_d$, each uniform in $[0, w)$. The hash function associated with this random set of choices is

$$h(x_1, \ldots, x_d) = \left(\left\lfloor \frac{x_1 - s_1}{w} \right\rfloor, \left\lfloor \frac{x_2 - s_2}{w} \right\rfloor, \ldots, \left\lfloor \frac{x_d - s_d}{w} \right\rfloor \right).$$

Let $\alpha_i = |p_i - q_i|$. What is the probability that $h(p) = h(q)$, in terms of the α_i values? It may be easier to first think of the case when $w = 1$. Try also to simplify your expression if w is much larger than α_i's, using that $(1 - x) \approx e^{-x}$ for small values of $x \geq 0$.

Solution: Let us try to picture what the hashing function does. On the i-th coordinate, it partitions $\mathbb{R}$ into buckets of the form ..., $[s_i - w, s_i), [s_i, s_i + w), [s_i + w, s_i + 2w)$, ..., each of length w, with a random "offset". Given two numbers p_i and q_i, the probability that they fall into the same bucket is $1 - \frac{|p_i - q_i|}{w}$ (unless they are farther away than w, in which case it is 0).[1] Therefore:

- if $|p_i - q_i| > w$ for some i, then $\Pr[h(p) = h(q)] = 0$,
- otherwise

$$\Pr[h(p) = h(q)] = \prod_{i=1}^{d} \left(1 - \frac{|p_i - q_i|}{w}\right) \approx \prod_{i=1}^{d} e^{-\frac{|p_i - q_i|}{w}} = e^{-\frac{\sum_{i=1}^{d} |p_i - q_i|}{w}} = e^{-\frac{\|p-q\|_1}{w}}.$$

3 *(*)* A certain ex-president's "university" has experienced a lot of cheating and you have been contacted to fix the problem. In particular, you should design a system for detecting plagiarism. To your help you have downloaded all the recent theses from the web. Let n be the number of theses. To simplify matters, you use the "bag of words" representation which just represents each thesis i by a vector v_i. The vector has an entry for each word in the English language that equals the number of times that word appears in the thesis. For example, in the ex-president's thesis, say thesis i, the word *rigorous* appears only once, whereas *tremendous* appears 1000 times, and so

$$v_i(\text{"rigorous"}) = 1 \qquad \text{and} \qquad v_i(\text{"tremendous"}) = 1000\,.$$

To deal with theses of different lengths, we normalize the vectors to have unit length. Let $u_1, \ldots, u_n$ denote the normalized vectors, i.e., $u_i = v_i / \|v_i\|_2$ for $i = 1, \ldots, n$. A good distance measure of similarity between two normalized vectors p and q (corresponding to two different theses) is the cosine similarity (or angular similarity) defined as follows:

$$\text{dist}(p, q) = \text{"the angle between } p \text{ and } q\text{"} = \cos^{-1}(\langle p, q \rangle)\,.$$

To detect plagiarism, it seems reasonable to inspect similar theses. That is, if the normalized vector q corresponding to a newly submitted thesis satisfies $\text{dist}(q, u_i) \leq 1^\circ$ for some $i = 1, \ldots, n$, then we would like to inspect u_i for plagiarism. To find such close u_i's, we wish to design an efficient (approximate) nearest neighbor search data structure. To do so, you need to design a family of locally sensitive hash (LSH) functions $\mathcal{H}$ that map normalized "bag of word" vectors to $\{0, 1\}$ such that for some $p_1 > p_2$:

- If $\text{dist}(q, u_i) \leq 1^\circ$ then $\Pr_{h \sim \mathcal{H}}[h(q) = h(u_i)] \geq p_1$. (we need to inspect u_i)
- If $\text{dist}(q, u_i) \geq 10^\circ$ then $\Pr_{h \sim \mathcal{H}}[h(q) = h(u_i)] < p_2$. (it is safe to ignore u_i)

What values of p_1 and p_2 do you get?

[1] To see this, assume wlog that $p_i < q_i < p_i + w$; there will be exactly one bucket-beginning in the interval $(p_i, p_i + w]$, the position of that bucket-beginning is distributed uniformly on that interval, and p_i and q_i will go into different buckets if and only if that bucket-beginning falls into $(p_i, q_i]$. The probability of this happening is $\frac{|p_i - q_i|}{w}$.

Hint: Randomly cut the sphere into two halves.

Solution:
We use the random hyperplane rounding algorithm. We choose a random vector r from a unit hyper sphere. Then we assign $h(u_i) \leftarrow 0$ if $r \cdot u_i \leq 0$ and $h(u_i) \leftarrow 1$ otherwise. The probability that we separate two vectors is exactly the degree between these two vectors divided by 180. Therefore, the values that we get for p_1 and p_2 are $1 - 1/180$ and $1 - 10/180$, respectively.

Submodular Functions (after Tuesday's lecture)

4 There is a set $[n] = \{1, \ldots, n\}$ of n different ice cream tastes. Maggie Simpson's total happiness goes up by $v_i \geq 0$ if she eats ice cream $i \in [n]$. However, as her stomach has bounded size, her happiness can never exceed some certain value $B > 0$. In other words, if Maggie eats $S \subseteq [n]$, her total happiness is

$$f(S) = \min\left(\sum_{i \in S} v_i, B\right).$$

Show that f is a submodular function.

Solution:
Let X and Y be two arbitrary subsets of $[n]$ such that $X \subseteq Y \subsetneq [n]$. We will prove that f satisfies the diminishing returns property: that is, for any $x \in [n] \setminus Y$,

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y). \tag{1}$$

Notice that due to the non-negativity of v_i's, $f(X) \leq f(Y)$. We consider the following cases:
Case 1: $\sum_{i \in X} v_i \geq B$.
In this case, since $X \subseteq Y$ and $v_i \geq 0$, we have $\sum_{i \in Y} v_i \geq B$. Consequently, all $f(X), f(Y), f(X \cup \{x\})$, and $f(Y \cup \{x\})$ evaluates to B and Inequality 1 trivially holds.
Case 2: $\sum_{i \in X} v_i < B$ and $v_x + \sum_{i \in X} v_i \geq B$.
Since $X \cup \{x\}$ is subset of $Y \cup \{x\}$ and v_i's are non-negative, we have both $f(X \cup \{x\}) = B$ and $f(Y \cup \{x\}) = B$ in this case. Therefore, $LHS = B - f(X)$ and $RHS = B - f(Y)$, and since $f(X) \leq f(Y)$, we have $LHS \geq RHS$.
Case 3: $v_x + \sum_{i \in X} v_i < B$.
For this case, $LHS = v_x$ and RHS can be at most v_x which again give $LHS \geq RHS$.

5 Let $f : 2^N \to \mathbb{R}$ be a submodular function. Show that the following functions are also submodular:

- $g(S) = f(S \cup A)$ where A is a fixed set.
- $g(S) = f(S \cap A)$ where A is a fixed set.
- $g(S) = f(N \setminus S)$.

Solution:

- Consider the marginal value of some item $u \in N$. We have

$$g(u \mid S) = f((\{u\} \cup S) \cup A) - f(S \cup A) = f(\{u\} \cup (S \cup A)) - f(S \cup A) = f(u \mid (S \cup A)),$$

for every set S. Let $C \subseteq D$. Then, $C \cup A \subseteq D \cup A$. Hence:

$$g(u \mid C) = f(u \mid (C \cup A)) \geq f(u \mid (D \cup A)) = g(u \mid D).$$

- Consider the marginal value of some item $u \in N$. We have

$$g(u \mid S) = f((\{u\} \cup S) \cap A) - f(S \cap A) = f((\{u\} \cap A) \cup (S \cap A)) - f(S \cap A)$$

Note that if $u \notin A$ we have $g(u \mid S) = 0$ for all S. Now suppose that $u \in A$. Then,

$$g(u \mid S) = f(\{u\} \cup (S \cap A)) - f(S \cap A) = f(u \mid (S \cap A))$$

for all sets S. Consider any sets $C \subseteq D$. Then, observe that $C \cap A \subseteq D \cap A$, so:

$$g(u \mid C) = f(u \mid (C \cap A)) \geq f(u \mid (D \cap A)) = g(u \mid D).$$

- Consider the marginal value of some item $u \in N$. Let $C \subseteq D$. Then, $N \setminus (D \cup \{u\}) \subseteq N \setminus (C \cup \{u\})$, and so:

$$\begin{aligned} g(u \mid C) &= f(N \setminus (C \cup \{u\})) - f(N \setminus C) \\ &= -f(u \mid (N \setminus (C \cup \{u\}))) \\ &\geq -f(u \mid (N \setminus (D \cup \{u\}))) \\ &= f(N \setminus (D \cup \{u\})) - f(N \setminus D) \\ &= g(u \mid D). \end{aligned}$$

6 *(half *)* Let $f : 2^N \to \mathbb{R}$ be a submodular function. Let $A \subseteq N$ and suppose that $A(p)$ is a random subset of A where each element $u \in A$ appears in $A(p)$ with probability p. Show that:

$$\mathbb{E}[f(A(p))] \geq (1-p)f(\emptyset) + pf(A).$$

To prove this, use that the Lovàsz extension $\hat{f}$ of f is the convex closure. That is, for input $z \in [0,1]^n$, we have that $\hat{f}(z)$ equals the minimum

$$\mathbb{E}_{S \sim \mu}[f(S)]$$

over all distributions μ of subsets satisfying the marginal probabilities: $\Pr_{S \sim \mu}[i \in S] = z_i$ for all $i \in \{1, 2, \ldots, n\}$.

Solution: Consider the vector $x = p\mathbf{1}_A$. Observe that the marginal probabilities of $A(p)$ agree with x. It follows then that:

$$\hat{f}(x) \leq \mathbb{E}[f(A(p))],$$

since $\hat{f}(x)$ is the smallest expected value attained by *any* distribution whose marginal probabilities agree with x. We also have $\hat{f}(x) = \mathbb{E}_{\theta \in \mathcal{U}(0,1)} f(T_\theta)$. But, $T_\theta = A$ whenever $\theta \leq p$ (which happens with probability p) and $T_\theta = \emptyset$ whenever $\theta > p$ (which happens with probability $1-p$). Thus:

$$\hat{f}(x) = pf(A) + (1-p)f(\emptyset).$$

7 Consider a directed $G = (V, E)$ and define the set function $f : 2^V \to \mathbb{R}$ by

$$f(S) = |\{(u, v) \in E : u \in S, v \notin S\}| \qquad \text{for every } S \subseteq V\,.$$

That is, $f(S)$ equals the number of arcs that exits the set S.

7a Show that f is a (non-monotone) submodular function

Solution: Let $A \subseteq B \subseteq V$ and let $u \in V \setminus B$. We show that $f(u \,|\, A) \geq f(u \,|\, B)$. We have that

$$f(u \,|\, A) = f(\{u\} \cup A) - f(A) = (\#\text{ of arcs from } u \text{ to } V \setminus A\ ) - (\#\text{ of arcs from } A \text{ to } u\ ),$$

and similarly,

$$f(u \,|\, B) = (\#\text{ of arcs from } u \text{ to } V \setminus B\ ) - (\#\text{ of arcs from } B \text{ to } u\ ).$$

But, because $A \subseteq B$, we have that

$$(\#\text{ of arcs from } u \text{ to } V \setminus B\ ) \leq (\#\text{ of arcs from } u \text{ to } V \setminus A\ )$$

and

$$(\#\text{ of arcs from } B \text{ to } u\ ) \geq (\#\text{ of arcs from } A \text{ to } u\ ),$$

which implies

$$\begin{aligned} f(u \,|\, B) &= (\#\text{ of arcs from } u \text{ to } V \setminus B\ ) - (\#\text{ of arcs from } B \text{ to } u\ ) \\ &\leq (\#\text{ of arcs from } u \text{ to } V \setminus A\ ) - (\#\text{ of arcs from } A \text{ to } u\ ) \\ &= f(u \,|\, A). \end{aligned}$$

Hence f is submodular.

Notice that f is not monotone unless $E = \emptyset$. It is easy to see that both $f(\emptyset) = 0$ and $f(V) = 0$, but if there is at least one arc $(u, v) \in E$ then $f(\{u\}) > 0$.

7b Let S be a random subset of vertices obtained by including each vertex with probability $1/2$ independently of other vertices. Show that

$$\mathbb{E}[f(S)] = |E|/4 \geq \mathrm{OPT}/4\,,$$

where $\mathrm{OPT} = \max_{T \subseteq V} f(T)$.

Also give an example of a graph where $|E| = \mathrm{OPT}$ and thus it shows that the analysis is tight with respect to OPT.

Solution: Let O be an optimal solution and let $E' = \{(u, v) \in E : u \in O, v \notin O\}$. Then $\mathrm{OPT} = f(O) = |E'|$. For each $(u, v) \in E'$, let $X_{u,v}$ be the random variable such that, $X_{u,v} = 1$ if $u \in S$ and $v \notin S$, and $X_{u,v} = 0$ otherwise. Then we have that $f(S) \geq \sum_{(u,v) \in E'} X_{u,v}$. By the choice of S, we have that $\mathbb{E}[X_{u,v}] = \Pr[X_{u,v} = 1] = \Pr[u \in S] \cdot \Pr[v \notin S] = \frac{1}{2} \cdot \frac{1}{2} = 1/4$ for all $(u, v) \in E'$. Therefore, by linearity of expectation, we have that

$$\mathbb{E}[f(S)] \geq \sum_{(u,v) \in E'} \mathbb{E}[X_{u,v}] = \frac{1}{4}|E'| = \frac{\mathrm{OPT}'}{4}.$$

To see this is tight, consider the graph $G = (V = [2n], E = \{1, 3, \ldots, 2n-1\} \times \{2, 4, \ldots, 2n\})$. It has $2n$ vertices and all arcs goes from an odd vertex to an even vertex so that $OPT = |E|$ and the set $O = \{1, 3, \ldots, 2n-1\}$ of all odd vertices is an optimal set. In this case, we can easily verify that $f(S) = \sum_{(u,v) \in E} X_{u,v}$ and $\mathbb{E}[f(S)] = |E|/4 = \mathrm{OPT}/4$.

7c (*) Consider any submodular function f that is

- non-negative: $f(T) \geq 0$ for all T.

Let S be a random subset of elements obtained by including each element with probability $1/2$ independently of other elements. Then

$$\mathbb{E}[f(S)] \geq \mathrm{OPT}/4\,,$$

where $\mathrm{OPT} = \max_T f(T)$.

This shows that the simple randomized algorithm actually gives a good approximation to any (even non-monotone) submodular function assuming it is non-negative.

Solution: Let O be an optimal set so that $\mathrm{OPT} = f(O)$. Let $O(1/2)$ be a random subset of O (i.e., each element $o \in O$ is selected independently with probability $1/2$). Recall from Problem 6 that $\mathbb{E}[f(O(1/2))] \geq (1/2)f(O) + (1-1/2)f(\emptyset) \geq (1/2)\mathrm{OPT}$. Let $\bar{O} = N \setminus O$ and let $\bar{O}(1/2)$ be a random subset of $\bar{O}$. Let $O' \subset O$. The function $g_{O'}(T) = f(T \cup O')$ is also a submodular function. By the same argument, we have that $\mathbb{E}[g_{O'}(\bar{O}(1/2))] \geq (1-1/2)g_{O'}(\bar{O}) + (1/2)g_{O'}(\emptyset) \geq (1/2)f(\emptyset \cup O') = (1/2)f(O')$. Thus if $N(1/2)$ is a random subset of N, we have the following:

$$\begin{aligned}
\mathbb{E}[f(N(1/2))] &= \sum_{O' \subseteq O} \Pr[O(1/2) = O']\mathbb{E}[f(N(1/2)) \,|\, O(1/2) = O'] \\
&= \sum_{O' \subseteq O} \Pr[O(1/2) = O']\mathbb{E}[g_{O'}(\bar{O}(1/2))] \\
&\geq \frac{1}{2} \sum_{O' \subseteq O} \Pr[O(1/2) = O']f(O') \\
&= \frac{1}{2}\mathbb{E}[f(O(1/2))] \geq \frac{1}{4}\mathrm{OPT}.
\end{aligned}$$