# Assignment: Assignment 4

Pass the exercise
You received 3 reviews and 0 likes
(/learn/approximation-algorithms-part-2/peer/E03rO/assignment-4/submit)

Review 3 classmates
2/3 reviews completed
(/learn/approximation-algorithms-part-2/peer/E03rO/assignment-4/give-feedback)

Instructions (/learn/approximation-algorithms-part-2/peer/E03rO/assignment-4)

My submission (/learn/approximation-algorithms-part-2/peer/E03rO/assignment-4/submit)

**Review classmates (/learn/approximation-algorithms-part-2/peer/E03rO/assignment-4/give-feedback)**

Discussions (/learn/approximation-algorithms-part-2/peer/E03rO/assignment-4/discussions)

## Correlation Clustering Assignment

by Ioannis Panagiotas
February 28, 2016

like Flag this submission

**An SDP based randomized algorithm for the Correlation Clustering problem**

The objective of this exercise is to design an algorithm for the {\it correlation clustering problem}. Given an undirected graph $G = (V, E)$ without loops, for each edge $e = \{i, j\} \in E$ there are two non-negative numbers $w_e^+, w_e^- \geq 0$ representing how similar and dissimilar are the nodes $i$ and $j$, respectively. For $S \subseteq V$, let $E(S)$ be the set of edges with both endpoints in $S$, that is, $E(S) = \{\{i, j\} \in E; i, j \in S\}$. The goal is to find a partition $\mathcal{S}$ of $V$ in order to maximize

$f(\mathcal{S}) = \sum_{S \in \mathcal{S}: e \in E(S)} w_e^+ + \sum_{e \in E \setminus \cup E(S)} w_e^-$.

In words, the objective is to find a partition that maximizes the total similarity inside each set of the partition plus the dissimilarity between nodes in different sets of the partition.

Consider the following simple algorithm:

**Algorithm 1**

*Let $\mathcal{S}_1 = \{\{i\} : i \in V\}$ the partition that considers each vertex as a single cluster, and $\mathcal{S}_2 = \{V\}$, that is every vertex in the same cluster. Compute the values $f(\mathcal{S}_1)$ and $f(\mathcal{S}_2)$ of this two partitions, and output the best among this two.*

Question 1. Compute the values $f(\mathcal{S}_1)$, $f(\mathcal{S}_2)$ in terms of the weights $w^-$ and $w^+$.

Question 2. Conclude that previous algorithm is a 1/2-approximation.

Let $B = \{e_\ell : \ell \in \{1, 2, \ldots, n\}\}$ be the canonical basis in $\mathbb{R}^n$, where $n = |V|$. For every vertex $i \in V$ there is a vector $x_i$ that is equal to $e_k$ if node $i$ is assigned to cluster $k$. Consider the following program:

$$\max\Big\{\sum_{\{i,j\} \in E} \Big(w^+_{\{i,j\}} x_i \cdot x_j + w^-_{\{i,j\}} (1 - x_i \cdot x_j)\Big) : x_i \in B \text{ for all } i \in V\Big\}.$$

Question 3. Explain why this program is a formulation of the correlation clustering problem.

The formulation is relaxed to obtain the following vector program:

$$\max \sum_{\{i,j\} \in E} \Big(w^+_{\{i,j\}} x_i \cdot x_j + w^-_{\{i,j\}} (1 - x_i \cdot x_j)\Big)$$

subject to $v_i \cdot v_i = 1 \text{ for all } i \in V,$

$v_i \cdot v_j \geq 0 \text{ for all } i, j \in V,$

$v_i \in \mathbb{R}^n \text{ for all } i \in V.$

Consider the following algorithm:

**Algorithm SDP**

Help Center

*Solve the the previous relaxation to obtain vectors* $\{v_i : i \in V\}$*, with objective value equal to* $Z$*. Draw independently two random hyperplanes with normals* $r_1$ *and* $r_2$*. This determines four regions,*

$R_1 = \{i \in V : r_1 \cdot v_i \geq 0 \text{ and } r_2 \cdot v_i \geq 0\},$

$R_2 = \{i \in V : r_1 \cdot v_i \geq 0 \text{ and } r_2 \cdot v_i < 0\},$

$R_3 = \{i \in V : r_1 \cdot v_i < 0 \text{ and } r_2 \cdot v_i \geq 0\},$

$R_4 = \{i \in V : r_1 \cdot v_i < 0 \text{ and } r_2 \cdot v_i < 0\},$

*and output the partition* $\mathcal{R} = \{R_1, R_2, R_3, R_4\}$.

In the following, the goal is to analyse this algorithm, and to prove that it is a 3/4-approximation.

Question 4. Let $X_{\{i,j\}}$ be the random variable that is equal to 1 if the vectors $v_i$ and $v_j$ lie in the same side of the two random hyperplanes, and zero otherwise. Using an argument similar to the one used for Max-Cut, prove that $\text{Prob}(X_{\{i,j\}} = 1) = (1 - \frac{1}{\pi}\theta_{\{i,j\}})^2$, where $\theta_{\{i,j\}} = \arccos(v_i \cdot v_j)$ is the angle between vectors $v_i$ and $v_j$.

Question 5. Let $f(\mathcal{R}) = \sum_{\{i,j\} \in E} \left( w^+_{\{i,j\}} X_{\{i,j\}} + w^-_{\{i,j\}} (1 - X_{\{i,j\}}) \right)$ the value of the partition $\mathcal{R}$, and denote $g(\theta) = (1 - \frac{1}{\pi}\theta)^2$ the probability function computed before. Prove that the expected value of $f(\mathcal{R})$, denoted by $E(f(\mathcal{R}))$, is

$\sum_{\{i,j\} \in E} \left( w^+_{\{i,j\}} g(\theta_{\{i,j\}}) + w^-_{\{i,j\}} (1 - g(\theta_{\{i,j\}})) \right).$

The following lemma will be helpful to conclude the analysis (*You don't need to prove it.)*

**Lemma.** For $\theta \in [0, \pi/2]$, $g(\theta) \geq \frac{3}{4}\cos(\theta)$ and $1 - g(\theta) \geq \frac{3}{4}(1 - \cos(\theta))$.

Question 6. Using the lemma conclude that $E(f(\mathcal{R})) \geq \frac{3}{4} \cdot Z$, and that the algorithm is a 3/4-approximation.

You can find my assignment here (https://drive.google.com/file/d/0B6vNxu30yUYVQk1jdURyVGFmT0U/view?usp=sharing)
Thank you!

The answer to Question 1 is of the following form:

The partition $\mathcal{S}_1$ satisfies that $E(\{v\}) = \emptyset$ for all $v \in V$ (the graph has no loops) and then $f(\mathcal{S}_1) = \sum_{e \in E} w^-_e$. On the other hand, partition $\mathcal{S}_2$ satisfies that all edges are internal, and then $f(\mathcal{S}_2) = \sum_{e \in E} w^+_e$.

- 2 pts
  Yes
- 0 pts
  No

The answer to Question 2 is of the following form:

An upper bound on the value of opt is the total sum of all weights, that is, $\text{opt} \leq \sum_{e \in E} (w^-_e + w^+_e)$. Then, $\text{OPT} \leq f(\mathcal{S}_1) + f(\mathcal{S}_2)$, which implies that $\max\{f(\mathcal{S}_1), f(\mathcal{S}_2)\} \geq \frac{1}{2} \cdot \text{opt}.$

- 3 pts
  Yes
- 0 pts
  No

The answer to Question 3 is of the following form:

The product $x_i \cdot x_j = 1$ if and only if nodes $i, j$ belong to the same cluster. In this case the edge is internal and then it contributes $w^+_{\{i,j\}} = w^+_{\{i,j\}} x_i \cdot x_j$ to the objective value. On the other hand, $x_i \cdot x_j = 0$ and the the nodes $i, j$ belong to different sets of the partition. The edge $\{i, j\}$ contributes $w^-_{\{i,j\}} = w^-_{\{i,j\}}(1 - x_i \cdot x_j)$ to the objective value.

- 3 pts
  Yes
- 0 pts
  No

The answer to Question 4 is of the following form:

The probability that vectors $v_i, v_j$ belong to different sides of a random hyperplane is equal to $\theta_{\{i,j\}}/\pi$. Therefore, the probability that $v_i, v_j$ belong to the same side of the random hyperplane $r_1$ is equal to $1 - \theta_{\{i,j\}}/\pi$, and the same holds for the random hyperplane $r_2$. Since both are drawn independently, it follows that $\text{Prob}(X_{\{i,j\}} = 1) = (1 - \frac{1}{\pi}\theta_{\{i,j\}})^2$.

4 pts
Yes

0 pts
No

The answer to Question 5 is of the following form:

$$E(f(\mathcal{R})) = \sum_{\{i,j\}\in E} \left( w^+_{\{i,j\}} E(X_{\{i,j\}}) + w^-_{\{i,j\}} (1 - E(X_{\{i,j\}})) \right)$$

$$= \sum_{\{i,j\}\in E} \left( w^+_{\{i,j\}} \mathrm{Prob}(X_{\{i,j\}} = 1) + w^-_{\{i,j\}} (1 - \mathrm{Prob}(X_{\{i,j\}} = 1)) \right)$$

$$= \sum_{\{i,j\}\in E} \left( w^+_{\{i,j\}} g(\theta_{\{i,j\}}) + w^-_{\{i,j\}} (1 - g(\theta_{\{i,j\}})) \right).$$

2 pts
Yes

0 pts
No

The answer to Question 6 is of the following form:

To conclude we use the following facts: i) $\theta_{\{i,j\}} = \arccos(v_i \cdot v_j)$, ii) the lemma, and iii) $Z \geq \mathrm{opt}$. Therefore,

$$E(f(\mathcal{R})) = \sum_{\{i,j\}\in E} \left( w^+_{\{i,j\}} g(\theta_{\{i,j\}}) + w^-_{\{i,j\}} (1 - g(\theta_{\{i,j\}})) \right)$$

$$\geq \frac{3}{4} \sum_{\{i,j\}\in E} \left( w^+_{\{i,j\}} \cos(\theta_{\{i,j\}}) + w^-_{\{i,j\}} (1 - \cos(\theta_{\{i,j\}})) \right)$$

$$= \frac{3}{4} \sum_{\{i,j\}\in E} \left( w^+_{\{i,j\}} v_i \cdot v_j + w^-_{\{i,j\}} (1 - v_i \cdot v_j)) \right)$$

$$= \frac{3}{4} Z$$

$$\geq \frac{3}{4} \mathrm{opt}.$$

4 pts
Yes

0 pts
No

Submit Review

Comments
Visible to classmates

share your thoughts...