CSE 8803 Homework 2

Jingyi Feng

November 12, 2023

1 Linear Programming

$$f(x) = min_x \|Ax - b\|_p \quad A \in \mathbb{R}^{m \times n} \quad b \in \mathbb{R}^{1 \times m} \tag{1}$$

1.1

Problem how the full details of how the above problem can be solved as a linear programming problem when $p = 1$ (the derivation was mostly given in the lecture)

Answer $f(x) = \|Ax - b\|_1 = \sum_{i=1}^{m} |a_i^T x - b_i|$ where a_i^T is the ith row of A. Let $|a_i^T x - b_i| = \delta_i$, then our goal becomes minimizing $\sum_{i=1}^{m} \delta_i$ where $-\delta_i < a_i^T x - b_i < \delta_i$. Consider the following matrix representation

$$\begin{pmatrix} -I_m & A \\ -I_m & -A \end{pmatrix} \begin{pmatrix} \delta \\ x \end{pmatrix} \leq \begin{pmatrix} b \\ -b \end{pmatrix} \quad \delta = \begin{pmatrix} \delta_1 \\ ... \\ \delta_m \end{pmatrix}$$

From above we can get two inequalities $Ax - \delta \leq b$ and $Ax - b \geq -\delta$ that satisfy $-\delta_i < a_i^T x - b_i < \delta_i$. Let

$$B = \begin{pmatrix} -I_m & A \\ -I_m & -A \end{pmatrix} \quad w = \begin{pmatrix} \delta \\ x \end{pmatrix} \quad c = \begin{pmatrix} b \\ -b \end{pmatrix}$$

We have the inequality in linear programming $B^T w \leq c$. Next, we need to let the goal of minimizing $\sum_{i=1}^m \delta_i$ be represented in the form of $max_w\, s^T w$ for some s. Minimizing $\sum_{i=1}^m \delta_i$ is actually maximizing $-\sum_{i=1}^m \delta_i$. Hence, we can let

$$-s = \begin{pmatrix} 1 \\ ... \\ 1 \\ 0 \\ .. \\ 0 \end{pmatrix}$$

We can get $argmin_w\, s^T w$ by solving $-argmax_w\, (-s)^T w$. By far, we have make L_1 norm minimization problem in a Dual LP formulation. We can also convert this into Primal formulation by solving $-min_t\, c^T y$ such that $By = -s$ for some $y \geq 0$.

1.2

Problem how the full derivations of how the above problem can be solved as a linear programming problem when $p = \infty$)

Answer Similar to the above question. $f(x) = \|Ax - b\|_\infty = max_{1 \leq i \leq m} |a_i^T x - b_i|$ where a_i^T is the ith row of A. Let $max_{1 \leq i \leq m} |a_i^T x - b_i| = \delta$, then our goal becomes minimizing δ

where $-\delta < a_i^T x - b_i < \delta$. Consider the following matrix representation

$$\begin{pmatrix} -I_m & A \\ -I_m & -A \end{pmatrix} \begin{pmatrix} \delta \\ x \end{pmatrix} \leq \begin{pmatrix} b \\ -b \end{pmatrix} \quad \delta = \begin{pmatrix} \delta \\ ... \\ \delta \end{pmatrix}$$

From above we can get two inequalities $Ax - \delta \leq b$ and $Ax - b \geq -\delta$ that satisfy $-\delta < a_i^T x - b_i < \delta$. Let

$$B = \begin{pmatrix} -I_m & A \\ -I_m & -A \end{pmatrix} \quad w = \begin{pmatrix} \delta \\ x \end{pmatrix} \quad c = \begin{pmatrix} b \\ -b \end{pmatrix}$$

We have the inequality in linear programming $B^T w \leq c$. Next, we need to let the goal of minimizing δ be represented in the form of $max_w\, s^T w$ for some s. Minimizing δ is actually maximizing $-\delta$. Hence, we can let

$$-s = \begin{pmatrix} 1 \\ 0 \\ .. \\ 0 \end{pmatrix}$$

We can get $argmin_w\, s^T w$ by solving $-argmax_w\, (-s)^T w$. By far, we have make L_1 norm minimization problem in a Dual LP formulation. We can also convert this into Primal formulation by solving $-min_t\, c^T y$ such that $By = -s$ for some $y \geq 0$.

1.3

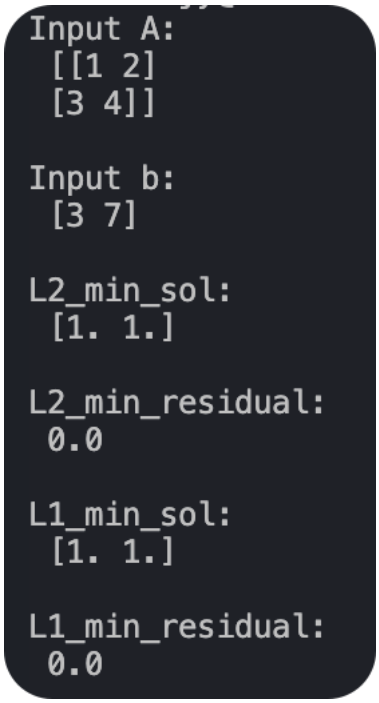
Problem 1 Programming: $xls = LS2(A, b)$

Answer Please see codes in file named *Linear_Programming*.

Problem 2 Programming: $xone = LS1(A, b)$

Answer Please see codes in file named *Linear_Programming*.

Here I use one linear system with exact solution to implicitly illustrate the correctness of the algorithm.



1.4

Problem Explore the differences between the solutions xls and xone by testing LS2 and LS1on various matrices. Come up with an input matrix A with outliers and b, and show that one of LS2 and LS1 is less sensitive to the outliers than the other. Present your findings along with some explanations.

Answer In this problem, to intuitively see the difference between $L1$ norm and $L2$ norm, we use 2D matrix A to form a linear regression problem as an example. We generated 20 data samples with linearly relationship, and added noise. To add a outlier, we make one entry of b be extremely large, as show in the following plot.

From above, we can see obviously the fitted straight line shifted above when we use $L2$ norm while it remains pretty much the same when we use $L1$ norm. This may because $L1 = \sqrt{\sum_{i=1}^{n} |a_i|}$ and $L2 = \sqrt{\sum_{i=1}^{n} |a_i|^2}$ where $a_i = Ax - b$. $L2$ can be interpreted as

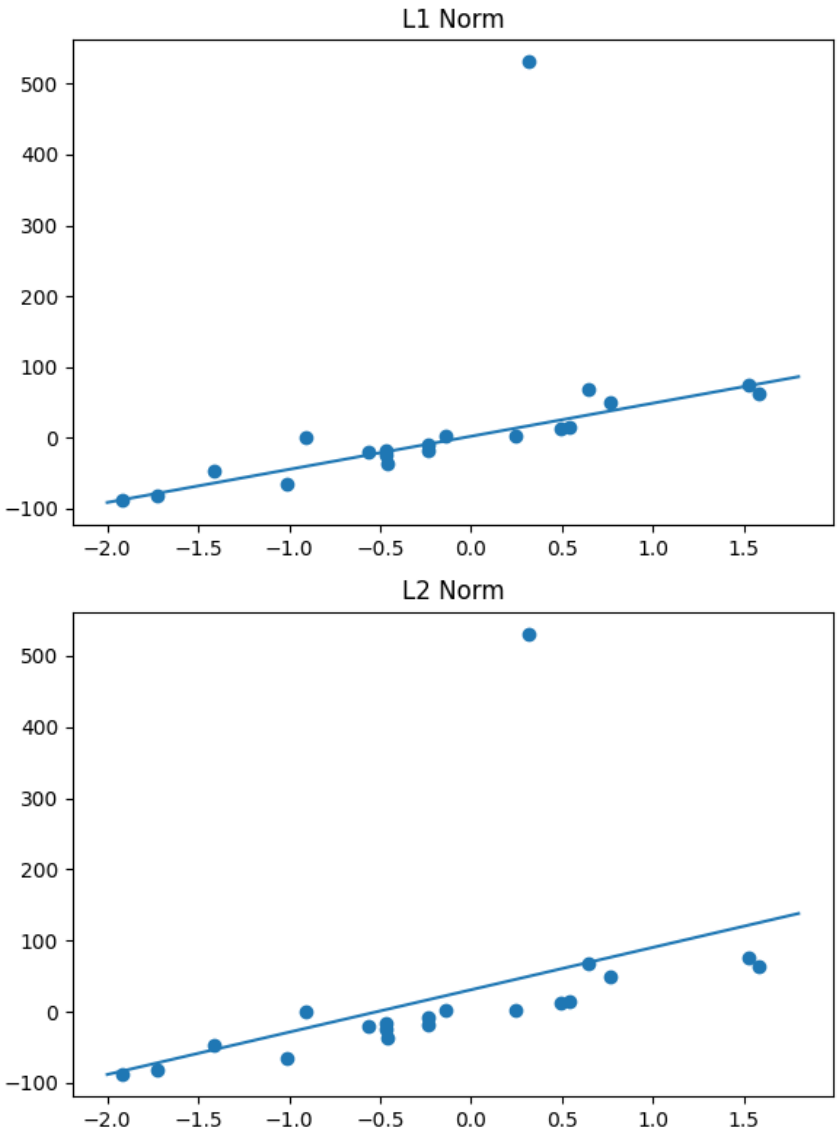


Figure 1: regression example

the direct distance from sample points to the line, while $L1$ can be seen as the Manhattan distance from sample points to the line. It will take less efforts to balance the residuals from each point to the line with $L2$ norm. However, when we use $L1$ norm, the residual reduced of the outlier will be larger than the increased residuals of all other sample points. That may be the reason that $L1$ norm is less sensitive to the outlier.

2 Spectral Clustering

2.1 Problem

Problem Present the full derivation/justification and description of the spectral clustering algorithm for binary clustering of a given undirected graph that minimizes the normalized cut with a relaxation.

Answer To find the binary clustering is actually same to minimize the normalized Ncut of the undirected graph, where $Ncut(A, \bar{A}) = \frac{1}{2}\sum_{i=1}^{2} \frac{W(A_i, \bar{A}_i)}{vol(A_i)}$, $vol(A_i) = \sum_{v_i \in A_i} d_{ii}$.

Given a subset $A \in V$, define $f = (f_1, f_2, ..., f_n)^T \in \mathbb{R}^n \quad n = |V|$ with entries

$$f_i = \begin{cases} \sqrt{vol(\bar{A})/vol(A)} & v_i \in A \\ -\sqrt{vol(A))/vol(\bar{A})} & v_i \in \bar{A} \end{cases} \tag{2}$$

In the special case of $k = 2$, we know that $\frac{1}{2}\sum_{v_i \in A, v_j \in \bar{A}} w_{ij} = cut(A, \bar{A}) = cut(A, \bar{A}) = \frac{1}{2}\sum_{v_i \in \bar{A}, v_j \in A} w_{ij}$. Hence, the following equation holds.

$$
\begin{aligned}
f^T L f &= \sum_{(v_i,v_j)\in E} w_{ij}(f_i - f_j)^2 \\
&= \frac{1}{2}\Bigg(\sum_{v_i\in A, v_j\in \bar{A}} w_{ij}\left(\sqrt{\frac{vol(\bar{A})}{vol(A)}} + \sqrt{\frac{vol(A)}{vol(\bar{A})}}\right)^2 \\
&\quad + \sum_{v_i\in \bar{A}, v_j\in A} w_{ij}\left(-\sqrt{\frac{vol(A)}{vol(\bar{A})}} - \sqrt{\frac{vol(\bar{A})}{vol(A)}}\right)^2\Bigg) \\
&= \left(\sqrt{\frac{vol(\bar{A})}{vol(A)}} + \sqrt{\frac{vol(A)}{vol(\bar{A})}}\right)^2 \cdot \quad \frac{1}{2}\sum_{v_i\in A, v_j\in \bar{A}} w_{ij} \\
&= cut(A,\bar{A})\left(\sqrt{\frac{vol(\bar{A})}{vol(A)}} + \sqrt{\frac{vol(A)}{vol(\bar{A})}}\right)^2 \\
&= cut(A,\bar{A})\left(\frac{vol(\bar{A})}{vol(A)} + 2 + \frac{vol(A)}{vol(\bar{A})}\right) \\
&= cut(A,\bar{A})\left(\frac{vol(\bar{A}) + vol(A)}{vol(A)} + \frac{vol(A) + vol(\bar{A})}{vol(\bar{A})}\right) \\
&= \left(\frac{cut(A,\bar{A})}{vol(A)} + \frac{cut(\bar{A},A)}{vol(\bar{A})}\right)\Big(vol(\bar{A}) + vol(A)\Big) \\
&= Ncut(A,\bar{A})\cdot vol(V)
\end{aligned}
\tag{3}
$$

Notice, for the vector f, we have the following two equations

$$
\begin{aligned}
(Df)^T\mathbb{1} &= \sum_{i=1}^{n} d_{ii} f_i \\
&= \sum_{v_i\in A} d_{ii}\sqrt{\frac{vol(\bar{A})}{vol(A)}} - \sum_{v_i\in \bar{A}} d_{ii}\sqrt{\frac{vol(A)}{vol(\bar{A})}} \\
&= vol(A)\sqrt{\frac{vol(\bar{A})}{vol(A)}} - vol(\bar{A})\sqrt{\frac{vol(A)}{vol(\bar{A})}} \\
&= \sqrt{vol(A)vol(\bar{A})} - \sqrt{vol(\bar{A})vol(A)} \\
&= 0
\end{aligned}
\tag{4}
$$

$$
\begin{aligned}
f^T Df &= \sum_{i=1}^{n} d_{ii} f_i^2 \\
&= \sum_{v_i \in A} d_{ii} \frac{vol(\bar{A})}{vol(A)} + \sum_{v_i \in \bar{A}} d_{ii} \frac{vol(A)}{vol(\bar{A})} \\
&= vol(A) \frac{vol(\bar{A})}{vol(A)} + vol(\bar{A}) \frac{vol(A)}{vol(\bar{A})} \\
&= vol(\bar{A}) + vol(A) \\
&= vol(V)
\end{aligned} \tag{5}
$$

Since $vol(V)$ is constant, to minimize Ncut is converted to

$$min_A f^T L f \quad where \quad Df \perp \mathbb{1}, \quad \|f^T Df\| = vol(V)$$

We relax the problem by let f be any vector in $\mathbb{R}^n$. Hence the goal becomes

$$min_{f \in \mathbb{R}^n} f^T L f \quad where \quad Df \perp \mathbb{1}, \quad \|f^T Df\| = vol(V)$$

Let $g = D^{\frac{1}{2}} f$. Then our minimization goal becomes

$$min_{g \in \mathbb{R}^n} g^T D^{-\frac{1}{2}} L D^{-\frac{1}{2}} g \quad where \quad D^{\frac{1}{2}} g \perp \mathbb{1}, \quad \|g^2\| = vol(V)$$

which is same as

$$min_{g \in \mathbb{R}^n} g^T L_{sym} g \quad where \quad D^{\frac{1}{2}} g \perp \mathbb{1}, \quad \|g^2\| = vol(V)$$

By Rayleigh-Ritz theorem, we know g is the second smallest eigenvector of L_{sym}. Notice, since $g = D^{\frac{1}{2}} f$, we know f is the second smallest eigenvector of L_{rw}. After computing f

using D_{rw}, we can find the cluster A and $\bar{A}$ from f. By the setting of f_i, we know

$$\begin{cases} v_i \in A & f_i \geq 0 \\ v_i \in \bar{A} & f_i < 0 \end{cases} \tag{6}$$

Similar for g. Since $g_{ii} = \sqrt{d_{ii}} f_{ii}$ and $\sqrt{d_{ii}}$ is always positive, we have

$$\begin{cases} v_i \in A & g_i \geq 0 \\ v_i \in \bar{A} & g_i < 0 \end{cases} \tag{7}$$

Therefore, we can separate the clusters by recognizing the sign of entries of either f or g.

2.2 Problem

Problem Write programs to compute two clusters for a given graph using spectral clustering method based on the unnormalized Laplacian.

Answer Please see codes in file named *Spectral_Clustering*. Below is the result of a random generated example.

2.3 Problem

Problem Write programs to compute two clusters for a given graph using spectral clustering method based on the symmetric normalized Laplacian

Answer Please see codes in file named *Spectral_Clustering*. Below is the result of the example above.

```
Cluster by Ncut(L_sym):
 [1 0 0 0 1 0]
```

2.4 Problem

Problem Present the binary clustering results obtained from your programs for the cockroach graph (Fig. 2 [von2007tutorial]). Present your findings along with some explanations/discussions.

Answer The results of applying unnormalized RatioCut using L and normalized Ncut using L_{sym} and L_{rw} on 12 nodes [von2007tutorial] are show as following.

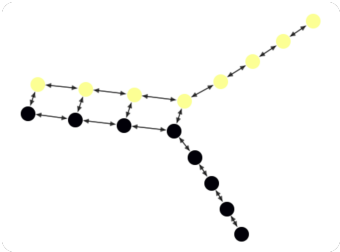


Figure 2: RatioCut

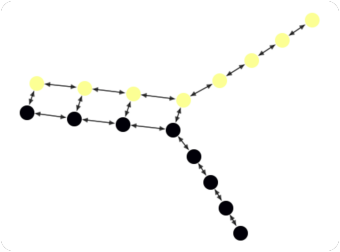


Figure 3: Ncut with L_{sym}

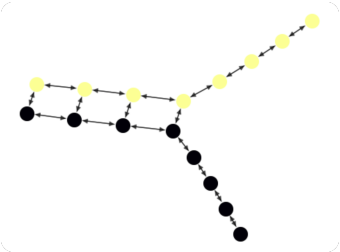


Figure 4: Ncut with L_{rw}

We noticed that the results of normalized and unnormalized cut of [von2007tutorial] graph are the same. However, none of them gave the optimal cut. If we cut the graph in the middle (i.e. cut between (v_k, v_{k+1}) and (v_{3k}, v_{3k+1})), we have

$$RatioCut = \frac{2}{8} + \frac{2}{8} = \frac{1}{2}, \qquad Ncut = \frac{2}{22} + \frac{2}{14} = \frac{18}{77}$$

However, for current cut, we have

$$RatioCut = \frac{4}{8} + \frac{4}{8} = 1, \qquad Ncut = \frac{4}{18} + \frac{4}{18} = \frac{4}{9}$$

We can find cut better than current cut, thus the current cut is not optimal. Just by observing the graph, we may assume the most reasonable cut should be the cut between (v_k, v_{k+1}) and (v_{3k}, v_{3k+1}), since the clusters will separate vertices based edge number. The assumed optimal cut should be like Figure5.

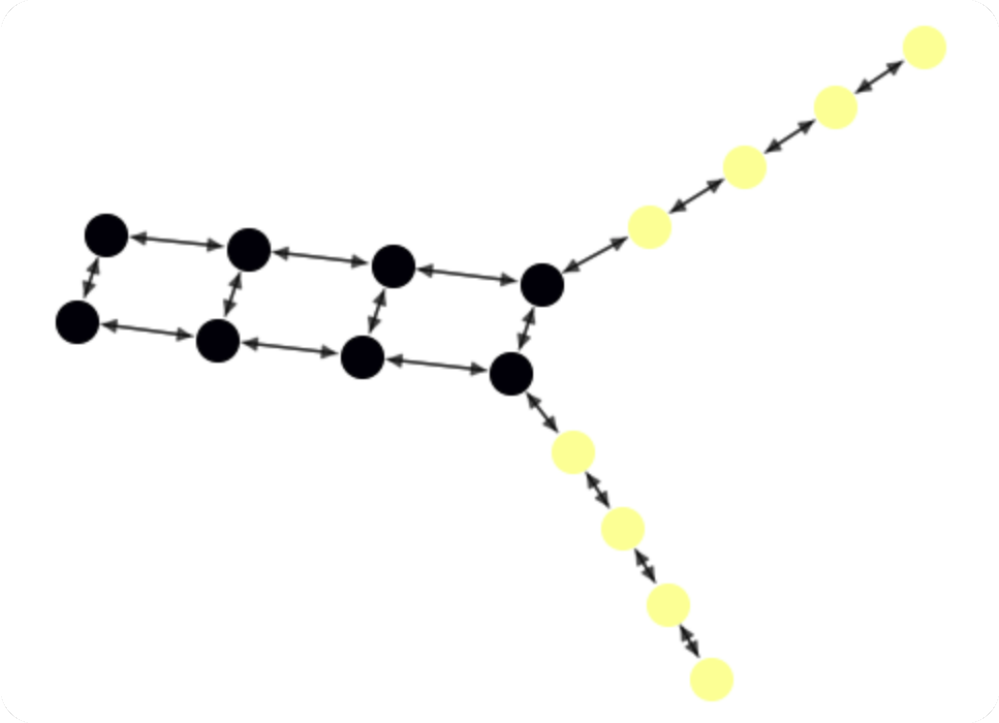


Figure 5: assumed optimal cut

2.5 Problem

Problem Based on the inspection of the cockroach graph (i.e., by manually figuring out the optimal binary clusters), can you verify that the computed binary clustering results are the actual optimal solutions for the corresponding minimization criteria (Sec. 4 [von2007tutorial]) ? Discuss the quality of the computed clustering results.

Answer In the case of [von2007tutorial] graph, the optimal cut should be in the middle (i.e. cut between (v_k, v_{k+1}) and (v_{3k}, v_{3k+1})).

First, we verify for $RatioCut = \frac{cut(A,\bar{A})}{|A|} + \frac{cut(\bar{A},A)}{|\bar{A}|} = 2 \cdot \frac{cut(A,\bar{A})}{|A|}$. The smaller cut is the smallest RatioCut is. To keep same cluster size, the only way to do that is to cut in middle. Hence, we need to cut between (v_k, v_{k+1}) and (v_{3k}, v_{3k+1}) to get optimal cut.

Then, we verify

$$\begin{aligned} Ncut =& \frac{cut(A,\bar{A})}{vol(A)} + \frac{cut(\bar{A},A)}{vol(\bar{A})} \\ &= cut(A,\bar{A}) \cdot (\frac{1}{vol(A)} + \frac{1}{vol(\bar{A})}) \\ &= cut(A,\bar{A}) \cdot (\frac{1}{vol(A)} + \frac{1}{vol(V) - vol(A)}) \\ &= cut(A,\bar{A}) \cdot vol(V) \cdot \frac{1}{vol(A)(vol(V) - vol(A))} \end{aligned} \tag{8}$$

By observation, $cut(A,\bar{A})$ should be as small as possible and $\frac{1}{vol(A)(vol(V)-vol(A))}$ should be as large as possible. We have possible cuts of 2 cuts, 3 cuts, 4 cuts or even 5 cuts situations. In each situation, the smaller gap between $vol(A)$ and $vol(\bar{A})$, the smaller $Ncut$ is. Hence, the best $Ncut$ for each situation is $\frac{18}{77}, \frac{12}{35}, \frac{4}{9}$... We noticed that $Ncut$ increases as cut number increases, because the $cut(A,\bar{A})$ is more dominate in $cut(A,\bar{A}) \cdot vol(V) \cdot \frac{1}{vol(A)(vol(V)-vol(A))}$. Hence, we know the optimal cut is to cut between (v_k, v_{k+1}) and (v_{3k}, v_{3k+1}).

Problem Come up with another graph for which the clustering results from the normalized and unnormalized methods are different. Discuss the quality of the computed clustering results

Answer In this case, we consider the graph in Figure6.

We get clusters by applying both unnormalized RatioCut and normalized Ncut. Results are show as following.

Intuitively, the cut in Figure 8 performs better than Figure 7. We can prove the obser-

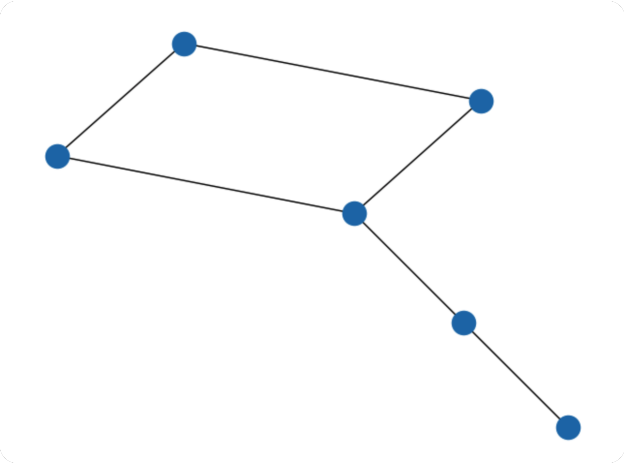


Figure 6: 6 nodes graph

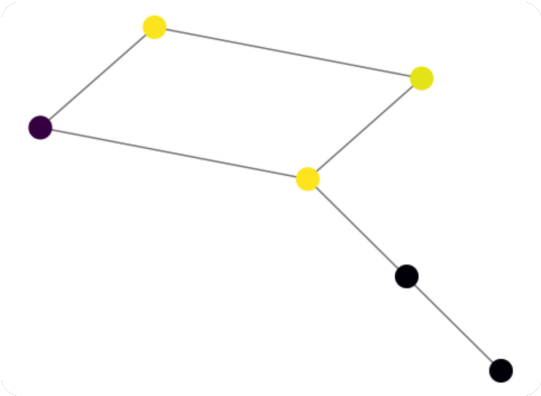


Figure 7: unnormalized RatioCut

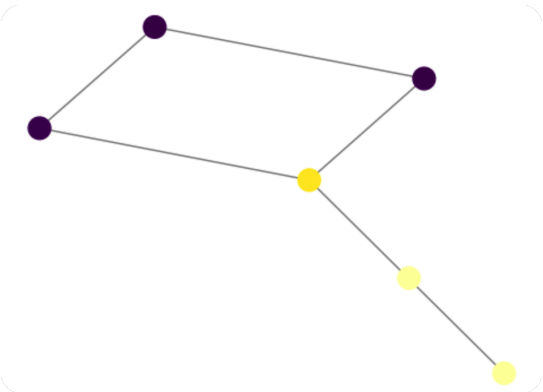


Figure 8: Normalized Ncut

vation by calculating the actual value of RatioCut and Ncut.

RatioCut and Ncut of Figure 7:

$$RatioCut = \sum_{i=1}^{2} \frac{cut(A_i, \bar{A}_i)}{|A_i|} = \frac{3}{3} + \frac{3}{3} = 2$$

$$Ncut = \sum_{i=1}^{2} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)} = \frac{3}{7} + \frac{3}{5} = \frac{36}{35}$$

RatioCut and Ncut of Figure 8:

$$RatioCut = \sum_{i=1}^{2} \frac{cut(A_i, \bar{A}_i)}{|A_i|} = \frac{2}{3} + \frac{2}{3} = \frac{4}{3}$$

$$Ncut = \sum_{i=1}^{2} \frac{cut(A_i, \bar{A}_i)}{vol(A_i)} = \frac{2}{6} + \frac{2}{6} = \frac{2}{3}$$

From above, we can see that normalized Ncut performs better in both criterions.