Spectral Clustering of Mathematics Exam Questions

Jiguang Li

The University of Chicago - Sendhil's Lab Presentation

Feb 19th, 2021

Problem Statement

- **Problem:** The Department of Elementary and Secondary Education (DESE) in MA classifies Grade 10 mathematics questions into 4 types:
  - Algebra (A)
  - Geometry (G)
  - Number and Quantity (N)
  - Probability and Statistics (P)

Problem Statement

- **Problem:** The Department of Elementary and Secondary Education (DESE) in MA classifies Grade 10 mathematics questions into 4 types:
  - Algebra (A)
  - Geometry (G)
  - Number and Quantity (N)
  - Probability and Statistics (P)

Does this classification make sense? Can we come up with a new cluster that is "better" than DESE's?

Evaluation Plan

- **Data:** $N = 70000$ students, $Q_a = 32$ multiple choice questions - imagine a matrix of size N by 32.

Evaluation Plan

- **Data:** $N = 70000$ students, $Q_a = 32$ multiple choice questions - imagine a matrix of size N by 32.
- **Evaluation:** Randomly pick 6 questions as holdout sample Q_h. Let DESE's cluster be $\{C_i\}_{i=1}^4$ and our cluster be $\{C'_j\}_{j=1}^m$. For each student s, compute:
  - True average score in 6 hold out sample: $\bar{y}_s = \frac{1}{|Q_h|} \sum_Q d_{sq}$
  - Baseline: $\bar{y}'_s = \frac{1}{|Q_h|} \sum_{Q_h} \sum_{i=1}^4 a_{si} \mathbb{1}_{q \in C_i}$
  - Our Prediction: $\tilde{y}_s = \frac{1}{|Q_h|} \sum_{Q_h} \sum_{j=1}^m a_{sj'} \mathbb{1}_{q \in c'_j}$

Evaluation Plan

- **Data:** $N = 70000$ students, $Q_a = 32$ multiple choice questions - imagine a matrix of size N by 32.
- **Evaluation:** Randomly pick 6 questions as holdout sample Q_h. Let DESE's cluster be $\{C_i\}_{i=1}^4$ and our cluster be $\{C_j'\}_{j=1}^m$.

For each student s, compute:

- True average score in 6 hold out sample: $\bar{y}_s = \frac{1}{|Q_h|}\sum_Q d_{sq}$
- Baseline: $\bar{y}_s' = \frac{1}{|Q_h|}\sum_{Q_h}\sum_{i=1}^4 a_{si}\mathbb{1}_{q\in C_i}$
- Our Prediction: $\tilde{y}_s = \frac{1}{|Q_h|}\sum_{Q_h}\sum_{j=1}^m a_{sj'}\mathbb{1}_{q\in c_j'}$

Finally, we compare the MSE:

- $\text{MSE}_1 = \mathbb{E}(\bar{y}_s - \bar{y}_s')^2$
- $\text{MSE}_2 = \mathbb{E}(\bar{y}_s - \tilde{y}_s')^2$

if $\text{MSE}_2 < \text{MSE}_1$, we win!

Why Spectral Clustering Algorithm?

- Doesn't Make any Strong Assumption of The Data

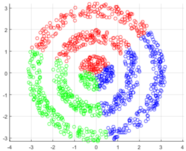


Figure: Kmeans Nightmare

Why Spectral Clustering Algorithm?

- Doesn't Make any Strong Assumption of The Data

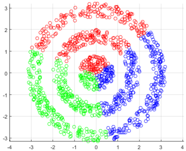


Figure: Kmeans Nightmare

- Works for High Dimension - just imagine plot 26 points in a $70,000$ dimension

Why Spectral Clustering Algorithm?

- Doesn't Make any Strong Assumption of The Data

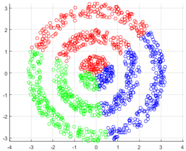


Figure: Kmeans Nightmare

- Works for High Dimension - just imagine plot 26 points in a $70,000$ dimension
- Easy to Implement

Why Spectral Clustering Algorithm?

- Doesn't Make any Strong Assumption of The Data

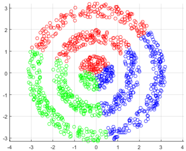


Figure: Kmeans Nightmare

- Works for High Dimension - just imagine plot 26 points in a $70,000$ dimension
- Easy to Implement
- Intriguing Visualizations - everybody likes graphs

Why Spectral Clustering Algorithm?

- Doesn't Make any Strong Assumption of The Data

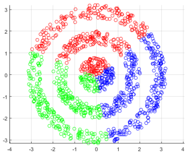
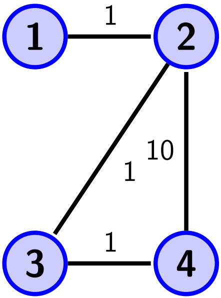


Figure: Kmeans Nightmare

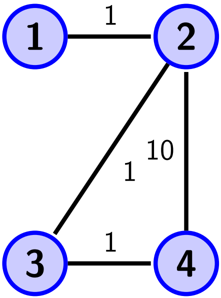
- Works for High Dimension - just imagine plot 26 points in a $70,000$ dimension
- Easy to Implement
- Intriguing Visualizations - everybody likes graphs
- It draws a beautiful connection between graph theory, linear algebra, and ML.



Degree Matrix:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 11 \end{bmatrix}$$

Notationless Graph Theory 101



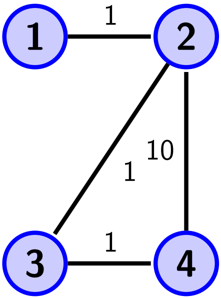
- **Weighted Adjacency Matrix**:

$$W = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 10 \\ 0 & 1 & 0 & 1 \\ 0 & 10 & 1 & 0 \end{bmatrix}$$

Degree Matrix:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 11 \end{bmatrix}$$

Notationless Graph Theory 101



Degree Matrix:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 11 \end{bmatrix}$$

- **Weighted Adjacency Matrix**:

$$W = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 10 \\ 0 & 1 & 0 & 1 \\ 0 & 10 & 1 & 0 \end{bmatrix}$$

- **Laplacian Matrix**:

$$L = D - W$$

One last Concept: Similarity Graphs/Matrix

Definition: Suppose we have data points $x_1, \cdots, x_n$, and some notion of similarity $s_{ij} \geq 0$: connect x_i and x_j with edge weights s_{ij}. The adjacency matrix of this graph is called **similarity matrix**.

Two Popular Types of Similarity Graphs

- ϵ - neighborhood graph: we only connect vertices v_i and v_j if their similarity score is higher than some threshold ϵ
- KNN Graph: connect each vertex v_i with its K nearest neighbors in terms of similarity scores.

One last Concept: Similarity Graphs/Matrix

Definition: Suppose we have data points $x_1, \cdots, x_n$, and some notion of similarity $s_{ij} \geq 0$: connect x_i and x_j with edge weights s_{ij}. The adjacency matrix of this graph is called **similarity matrix**.

Two Popular Types of Similarity Graphs

- ϵ - neighborhood graph: we only connect vertices v_i and v_j if their similarity score is higher than some threshold ϵ
- KNN Graph: connect each vertex v_i with its K nearest neighbors in terms of similarity scores.

We want to find a partition of the graph such that the edges between different groups have very low weights and the edges within a group have high weights.

Unnormalized Spectral Clustering Algorithm

Algorithm 1 UNNORMALIZED SPECTRAL CLUSTERING

Input : Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.

- Construct a similarity graph from S (e.g. KNN graph). Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L
- Compute the first k eigenvectors $u_1, \cdots, u_k$
- Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, \cdots, u_k$ as columns
- For $i = 1 \cdots n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U
- Cluster the points $(y_i)_{i=1,\cdots,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \cdots, C_k$

Output: Clusters $A_i, \cdots, A_k$ with $A_i = \{j \mid y_j \in C_i\}$

Back to Original Question

Main idea: we treat each of the 26 non-holdout question as a node in a graph, we form edges between v_i and v_j with edge weight s_{ij}.

Back to Original Question

Main idea: we treat each of the 26 non-holdout question as a node in a graph, we form edges between v_i and v_j with edge weight s_{ij}.

Two candidates of Similarity Scores

- Correlation Matrix: Easy to implement

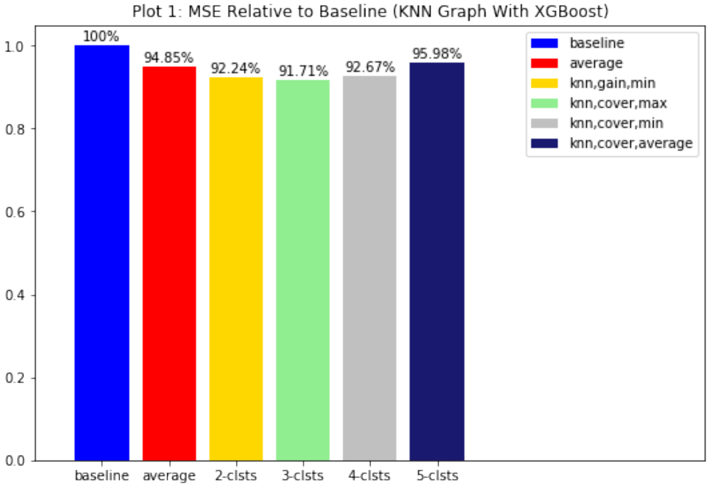
Back to Original Question

Main idea: we treat each of the 26 non-holdout question as a node in a graph, we form edges between v_i and v_j with edge weight s_{ij}.

Two candidates of Similarity Scores

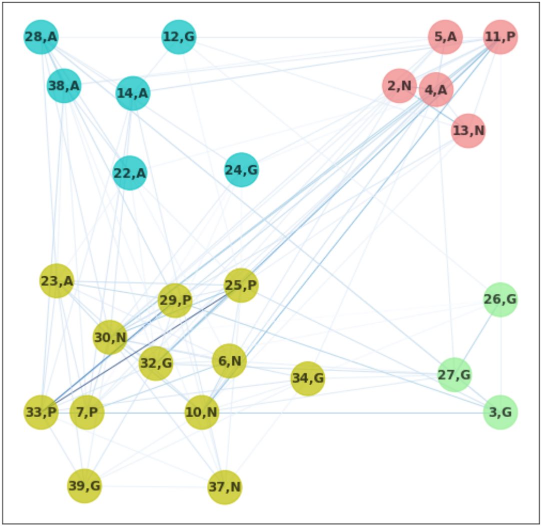
- Correlation Matrix: Easy to implement
- XGBoost Feature Importance Scores : be careful for symmetry issues

Result: Bar Chart



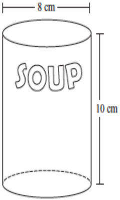
Result: Graph Visualization

KNN cluster XGBoost: 4-Cluster Case



Result: Discrepancy In Geometric Problems

3 A container of soup is in the shape of a right circular cylinder. The container and its dimensions are shown below.



What is the volume, in cubic centimeters, of the container?

A. 200π

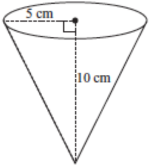
B. 160π

C. 80π

D. 40π

(a) Pure Geometry

26 The diagram below shows a cup in the shape of a right circular cone and some of its measurements.



Which of the following is closest to the volume, in cubic centimeters, of the cup?

A. 83

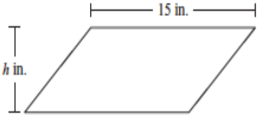
B. 105

C. 262

D. 524

(b) Pure Geometry

12 A parallelogram and some of its dimensions are shown below.



The area of the parallelogram is 90 square inches. What is h, the height in inches of the parallelogram?

A. 6

B. 8

C. 10

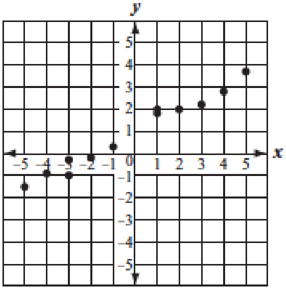
D. 12

(c) Algebra + Geometry

Figure: Discrepancy in Geometric Problems

Result: A Probability Question in Disguise

11 A set of data is shown in the scatterplot below.



Which of the following equations best represents the line of best fit for the data in the scatterplot?

A. $y = -\frac{1}{2}x - 2$

B. $y = -\frac{1}{2}x + 1$

C. $y = \frac{1}{2}x - 2$

Conclusion: Some Implementation Advice

- KNN graphs tend to perform better than ϵ graph in practice.
- Be careful about the assumptions similarity scores: symmetric and positive
- Use Cross-validation to find optimal k and ϵ. Don't disconnect the graph for small samples.
- Give spectral clustering a try next time when you want to use KMeans

Conclusion: Some Implementation Advice

- KNN graphs tend to perform better than ϵ graph in practice.
- Be careful about the assumptions similarity scores: symmetric and positive
- Use Cross-validation to find optimal k and ϵ. Don't disconnect the graph for small samples.
- Give spectral clustering a try next time when you want to use KMeans

Properties of Graph Laplacian Matrix

Proposition

1. For every vector $f \in \mathbb{R}^n$, we have

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2$$

Properties of Graph Laplacian Matrix

Proposition

1. *For every vector $f \in \mathbb{R}^n$, we have*

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2$$

2. *L is symmetric and positive semidefinite.*

Properties of Graph Laplacian Matrix

Proposition

1. For every vector $f \in \mathbb{R}^n$, we have

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2$$

2. L is symmetric and positive semidefinite.
3. The smallest eigenvalue of L is 0, the corresponding eigenvector is all one vector $\mathbb{1}$

Properties of Graph Laplacian Matrix

Proposition

1. For every vector $f \in \mathbb{R}^n$, we have

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$$

2. L is symmetric and positive semidefinite.
3. The smallest eigenvalue of L is 0, the corresponding eigenvector is all one vector $\mathbb{1}$
4. L has a non-negative, real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

RatioCut of View

Intuitively, suppose we want to form K clusters $A_1, \cdots, A_k$, we want to minimize

$$\text{Cut}(A_1, \cdots, A_k) = \frac{1}{2} \sum_{i=1}^{k} W(A_i, \bar{A}_i)$$

RatioCut of View

Intuitively, suppose we want to form K clusters $A_1, \cdots, A_k$, we want to minimize

$$\mathrm{Cut}(A_1, \cdots, A_k) = \frac{1}{2}\sum_{i=1}^{k} W(A_i, \bar{A}_i)$$

However, the solution often separates one individual vertex from the rest of the graph. Instead we can minimize

$$\mathrm{RatioCut}(A_1, \cdots, A_k) = \sum_{i=1}^{k} \frac{\mathrm{Cut}(A_i, \bar{A}_i)}{|A_i|}$$

Intuition: when $k = 2$

We want to find the best subset A, such that

$$\min_{A \subset V} \mathrm{RatioCut}(A, \bar{A})$$

Intuition: when $k = 2$

We want to find the best subset A, such that

$$\min_{A \subset V} \mathrm{RatioCut}(A, \bar{A})$$

Define

$$f_i = \begin{cases} \sqrt{\frac{|\bar{A}|}{|A|}} & v_i \in A \\ -\sqrt{\frac{|A|}{|\bar{A}|}} & v_i \in \bar{A} \end{cases}$$

Intuition: when $k = 2$

We want to find the best subset A, such that

$$\min_{A \subset V} \mathrm{RatioCut}(A, \bar{A})$$

Define

$$f_i = \begin{cases} \sqrt{\frac{|\bar{A}|}{|A|}} & v_i \in A \\ -\sqrt{\frac{|A|}{|\bar{A}|}} & v_i \in \bar{A} \end{cases}$$

We can show

- $f^T L f = |V| \, \mathrm{RatioCut}(A, \bar{A})$
- $\sum_{i=1}^{n} f_i = 0$
- $\|f\|^2 = n$

Intuition: when $k = 2$, continue

Now, the previous optimization problem can be transformed to

$$\min_{A \subset V} f^T L f \textbf{ subject to } f \perp \mathbb{1},\ \|f\| = \sqrt{n}$$

Intuition: when $k = 2$, continue

Now, the previous optimization problem can be transformed to

$$\min_{A \subset V} f^T L f \textbf{ subject to } f \perp \mathbb{1}, \; \|f\| = \sqrt{n}$$

However, this is discrete optimization problem as the entries of the solution vector f are restricted. we want to relax it to

$$min_{f \in \mathbb{R}^n} f^T L f \textbf{ subject to } f \perp \mathbb{1}, \; \|f\| = \sqrt{n}$$

Intuition: when $k = 2$, continue

Now, the previous optimization problem can be transformed to

$$\min_{A \subset V} f^T L f \textbf{ subject to } f \perp \mathbb{1}, \; \|f\| = \sqrt{n}$$

However, this is discrete optimization problem as the entries of the solution vector f are restricted. we want to relax it to

$$min_{f \in \mathbb{R}^n} f^T L f \textbf{ subject to } f \perp \mathbb{1}, \; \|f\| = \sqrt{n}$$

By Rayleigh-Rietz Theorem, the solution is the second smallest eigenvalue of L.

Intuition: larger k

Suppose we want to partition V into k sets $A_1, \cdots, A_k$, we define k indicator vectors $h_j = [h_{1,j}, \cdots, h_{n,j}]^T$ by

$$h_{i,j} = \begin{cases} \frac{1}{\sqrt{|A_j|}} & v_i \in A_j \\ 0 & \text{otherwise} \end{cases}$$

Intuition: larger k

Suppose we want to partition V into k sets $A_1, \cdots, A_k$, we define k indicator vectors $h_j = [h_{1,j}, \cdots, h_{n,j}]^T$ by

$$h_{i,j} = \begin{cases} \frac{1}{\sqrt{|A_j|}} & v_i \in A_j \\ 0 & \text{otherwise} \end{cases}$$

Let $H \in \mathbb{R}^{n \times k}$ as the matrix containing these k indicator vectors as columns. We can show:

- $H^T H = I$
- $h_i^T L h_i = \frac{\text{Cut}(A_i, \bar{A}_i)}{|A_i|}$
- $h_i^T L h_i = (H^T L H)_{ii}$
- $\text{RatioCut}(A_1, \cdots, A_k) = \sum_{i=1}^{k} h_i^T l h_i = \text{Tr}(H' L H)$

Intuition: larger k, continue

We can rewrite our previous problem as

$$\min_{A_1 \cdots A_k} \mathrm{Tr}(H^T L H) \textbf{subject to } H^T H = I$$

We have to relax it for numerical solvability:

$$\min_{H \in \mathbb{R}^{n \times k}} \mathrm{Tr}(H^T L H) \textbf{subject to } H^T H = I$$

By Rayleigh-Rietz, the solution is given by the first k eigenvectors of L

Refernces

- Von Luxburg, U. A tutorial on spectral clustering. Stat Comput 17, 395–416 (2007). https://doi.org/10.1007/s11222-007-9033-z