Algorithmic Foundations of Data Science: Assignment #7

Khaled AlHosani (kah579), Myunggun Seo (ms9144)

New York University Abu Dhabi
April 1, 2018

Problem 1

Create a table **D** of size $n \times k$. In the position (i, j) of D save the minimum loss (the total sum of squares of distance between each datapoint and the corresponding cluster center) of the i lowest data points with j clusters. Initialize the table with $D[i, j] = 0$ for all $i \leq j$ Since if we have more clusters than the points, we can always get 0 loss.

Create a table **d** of size n^2. In the position (i, j) of d, we will save the sum of squares of distance between each datapoint $m \in \{i...j\}$ and their mean $(i \leq j)$. For a fixed i, we can iteratively compute $d[i, j] = d[i, j-1] + \frac{i-j-2}{i-j-1}(x_j - \mu)^2)$ and update $\mu = \frac{i-j-2}{i-j-1}(\mu + x_j)$.

The value of $D[i, j]$ can be computed by breaking the i points into all possible clusterings of $1 + (j-1)$ clusters. We need to consider the cases where we have $j \leq m < i$ where m is the index of the greatest point on the $(j-1)$th cluster. (We don't need to consider the case when $i \leq j$ since the loss will be 0.)

We can find the $D[i, j] = \min_{j \leq m < i}(D[m, j-1] + d[j, i])$ where $d[j, i]$ is the sum of squares of distance between each datapoint in $j, ..., i$ and their mean. Since $d[j, i]$ can be retrieved in linear time, computing each $D[i, j]$ requires $O(n)$ time, which is order required for looping with $m \in [n]$. This job needs to be done for half of the $O(nk)$ table and thus requires $O(n^2k)$ time overall.

We read the following paper as reference:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5148156/

Problem 2

Please refer to `Assignment #7 (ms9144, kah579) Problem 2` for the code.

Problem 3

Please refer to `Assignment #7 (ms9144, kah579) Problem 3` for the code.

Problem 4

We need to show that the subspace of the eigenvectors of L with eigenvalue 0 is spanned by the vectors $\{\mathbb{1}_{A_i} : i \in [k]\}$.

We have that for any $x \in \mathbb{R}^n$, $x^T Lx = \sum_{\{i,j\} \in E}(x_i - x_j)^2$. If x is an eigenvector corresponding to eigenvalue 0, the left hand side is equal to 0. In such a case, the right hand side equals 0 iff $x_i = x_j$ for all $\{i, j\} \in E$. This means that unless E is an empty set, in which case $L = 0$ and x can take any value, the vertices of any and all connected components have the same x_i for each 0-eigenvector whereas two disconnected vertices must have different x_i values for each 0-eigenvector.

If for each 0-eigenvector v_i ($i = 1..k$) each connected component has a corresponding value c_j for $j \in [k]$, then $v_i = \sum_j c_j \mathbb{1}_{A_j}$. Which means that v_i is a linear combination of all A_j's. Thus the subspace spanned by v_i's are also spanned by A_j's.