Kungliga Tekniska högskolan

Machine Learning advanced course

Assignment 2B

Student :
Yohan Pellerin

Course coordinator and teachers :
Aristides Gionis

January 16, 2024

1 Multidimensional Scaling (MDS) and Isomap

Question 1

In "double centering", the similarity matrix (S) and the distance matrix (D) both contain the same information the distance between two points. Both preserve the relative distances between points in the high-dimensional space and seeks to represent these relationships in a lower-dimensional space. This is why the double tricks is possible. Moreover, The process of subtracting the row means and column means from each entry in D helps remove biases introduced by the differences in row and column scales. By doing this, you're effectively centering the distances and focusing on the relative differences between points rather than their absolute values.

Question 2

The “first point” trick also remove biases introduced be the differences in row and column scale. It captures the relative positioning of points with respect to the reference point (first point) in the dataset. Then, by removing it, the biases are removed.

The "first point" trick solution emphasizes the relationships between distances from the first point and pairwise distances, while the "double centering" trick focuses on centering and preserving the relative positioning of points based on the original distance matrix. Both solutions are correct in the sense that they attempt to capture and represent the relationships encoded in the distances, but they approach the problem from different perspectives, leading to distinct but equally valid configurations in a lower-dimensional space.

Question 3

In the PCA method, we wrote Y with the singular value decomposition as $Y = U\Sigma V^T$ from the singular value decomposition we can write $Y^TY = V\Sigma^2V^T$ or in MDS we have $S = Y^TY = V\Lambda^2V^T$ then we can say if we choose the right V for the eigendecomposition we have $\Sigma = \Lambda^{1/2}$ Moreover, in the PCA method, we map data to k-dimensional space by $X = U_k^TY$, by $X = I_{k*n}\Lambda^{1/2}V^T$ in the MDS. Then, as U_k is the k columns of U that correspond to the k largest singular values if the eigen value are well ordered as $U^TU = I$, we have $U_k^TU = I_{k*n}$.
Finally, we can write : $X_{PCA} = U_k^TY = U_k^TU\Sigma V^T = I_{k*n}\Lambda^{1/2}V^T = X_{MDS}$

To find the more efficient method, we focus on the time complexity of the different method. For the PCA using SVD method, the time complexity is the one of the SVD which is $O(n^3) + O(d * n^2)$
For the MDS method, computing S requires $O(n^2 * d)$ operations then it is just a eigen value decomposition, at most $O(n^3)$. The two methods seems similar in term of time complexity. However, the MDS can also be used when Y is not known.

Question 4

In the Isomap method, the process to obtain the neighborhood graph G involves constructing a graph where each data point is connected to its nearest neighbors.
Imagine a dataset where the points are arranged in multiple clusters that are quite distant from each other, with no points connecting these clusters within the nearest neighbor distance. In this scenario, all points within each cluster are close enough to be connected in the neighborhood graph, but there are no points close enough between the two clusters to form connections according to the nearest neighbor criterion.
So, if the Isomap method strictly adheres to the nearest neighbor criterion for graph construction, it might yield a disconnected graph in such cases.

Question 5

One possible heuristic to patch the problem is to instead of strictly considering only the nearest neighbors within a fixed distance, expand the neighborhood radius to connect points that might be within a certain distance threshold, even if they are not the absolute nearest neighbors. This allows for connections between clusters that might otherwise remain disconnected. Then, with this heuristic, if we choose a good neighborhood radius the different clusters will be linked.

2 Success probability in the Johnson-Lindenstrauss lemma

Question 6

Let's note S the probability of success for one trial. We are looking the probability that one of the trail succeed. To do so, let's calculate the probability that all of them fail $P\left(\bigcap_{i=1}^{n}\overline{S_i}\right)$. By the Johnson-Lindenstrauss lemma, we know that $P\left(S_i\right) \geq 1/n$.

$$P\left(\bigcap_{i=1}^{p}\overline{S_i}\right) = \bigcap_{i=1}^{p} P\left(\overline{S_i}\right) \leq \prod_{i=1}^{p}(1-1/n) = (1-1/n)^p \tag{1}$$

Then, the probability of having one success is $1 - P\left(\bigcap_{i=1}^{p}\overline{S_i}\right) \geq 1-(1-1/n)^p$ And $1-(1-1/n)^p \geq 95/100$ if and only $(1-1/n)^p \leq 5/100$ if and only $p \geq \frac{\log(5/100)}{\log(1-1/n)}$

Or $\log(1-1/n) \sim -1/n$ then $\frac{\log(5/100)}{\log(1-1/n)} = O(n)$ Then O(n) independent trials are sufficient for the probability of success to be at least 95%.

3 Node similarity for representation learning

Question 7

The matrix P represents the probabilities of transitioning from one node to another in the graph within a single step. Each entry P_{ij} signifies the probability of moving from node i to node j in one step following the graph edges.

Raising the matrix P to higher powers (P^k) computes the probabilities of reaching node j from node i in k steps. Each P_{ij}^k element in the resultant matrix gives the probability of transitioning from node i to node j in exactly k steps.

The parameter α is a real number between 0 and 1. It serves as a discount factor, diminishing the influence of probabilities for higher powers of P (P^k) as k increases. This reflects a a preference for shorter paths in the graph, as paths of greater length (higher k) have decreasing weight in the similarity measure.

Two nodes are considered similar if there is a high probability of transitioning from one to another. Additionally, the closer this transition occurs, the higher the perceived similarity between the nodes.

Question 8

4 Spectral graph analysis

Question 9

$$x^T L x = x^T x - \frac{1}{d} x^T A x = \frac{1}{d}(\sum_{i}^{\|V\|} d * x_i^2 - \sum_{i=1}^{\|V\|} \sum_{j=1}^{\|V\|} x_i A_{ij} x_j) \tag{2}$$

or, as each node as d edges, $\sum_{i=1}^{\|V\|} d * x_i^2 = \sum_{(i,j) \in E} x_i^2 + x_j^2$

for each i, $A_i j = 1$ if and only (i,j) $\in E$ and there is d j like this then, $\sum_{j=1}^{\|V\|} x_i A_{ij} x_j = \sum_{k=1}^{d} x_i x_{i_k}$

Then, $\sum_{i=1}^{\|V\|} \sum_{j=1}^{\|V\|} x_i A_{ij} x_j = \sum_{i=1}^{\|V\|} \sum_{k=1}^{d} x_i x_{i_k}$ However, this is like counting each edge twice. So, $\sum_{i=1}^{\|V\|} \sum_{j=1}^{\|V\|} x_i A_{ij} x_j = \sum_{(i,j) \in E} x_i * x_j$.

Finally, putting all together we get : $x^T L x = \frac{1}{d} \sum_{(i,j) \in E} (x_i - x_j)^2$

Question 10

As $x^T L x = \frac{1}{d} \sum_{(i,j) \in E} (x_i - x_j)^2 \geq 0$

L (the normalized Laplacian) is a positive semi-definite matrix.

Question 11

Non-trivial means different the vector is not zero. If x_* exists and is not zero, it means that for each $(i, j) \in E$, $x_{*i} = x_{*j}$. So, each connected vertex as the same value, then each vertex connected is very similar,then it is a meaningful embedding.

Question 12