DD2434 - Machine Learning, Advanced Course
Assignment 2B

Tristan Perrot
tristanp@kth.se

December 2023

Contents

1 Multidimensional Scaling (MDS) and Isomap

1.1 Question 1

The intuitive reason that the "double centering" trick works is that, as for the PCA, we want to center and then here we want to center by subtracting the overall mean and therefore the mean for the columns and the rows. This is why subtracting the mean twice works.

1.2 Question 2

While the double centering method center the data around the origin, the "first point" trick center the data around the first point of the dataset. Therefore the solution will be different but in MDS we are only interested in the relative position of the points and not their absolute position. Therefore the solution will be the same up to a translation.

1.3 Question 3

As stated, the classical MDS algorithm when Y is known is based on the eigen-decomposition of $S = Y^T Y$. Then, the singular values of Y are the nonnegative square roots of the eigenvalues of S. PCA on Y is based on the singular value decomposition of $Y = U\Sigma V^T$. Then, the singular values of Y are the diagonal entries of Σ. Therefore the methods are equivalent because they both calculate the singular values of Y to reduce the dimension by getting the directions with the highest variance. When n is much larger than p (many observations, few dimensions) PCA will tend to be more efficient because it avoids the need to calculate and store a large distance matrix.

1.4 Question 4

During the Isomap method, the process to obtain the neighborhood graph may yield a disconnected graph. For example, if we have two clusters of points that are well separated and a small p, each point will be connected to there p nearest neighbors and therefore the graph will be disconnected.

1.5 Question 5

Imagine as stated above a dataset with two clusters well separated that yield a disconnected graph. Now, we could search the two closest points in the separated clusters and connect them by an weighted edge of there distance in the dataset. We could repeat this if there is more than 2 clusters until the graph is connected. This method is based on the fact that we need to have a fully connected graph but we still want to well describe the distance between the points. Therefore, it is expected to work well in practice.

2 Success probability in the Johnson-Lindenstrauss lemma

2.1 Question 6

Let us denote the probability of success of an trial p and the probability of failure $q = 1 - p$ and the number of trials n. We know that $p \geq \frac{1}{n}$. We want to have $q^k \leq 0.05$. Which means:

$$\begin{aligned} q^k \leq 0.05 &\Leftrightarrow \ln q^k \leq \ln 0.05 \\ &\Leftrightarrow k \geq \frac{\ln 0.05}{\ln q} \\ &\Leftrightarrow k \geq \frac{\ln 0.05}{\ln(1-p)} \\ &\Leftrightarrow k \geq \frac{\ln 0.05}{\ln(1-\frac{1}{n})} \end{aligned}$$

With a high n we have $\ln(1 - \frac{1}{n}) \approx -\frac{1}{n}$ and therefore:

$$k \geq \frac{\ln 0.05}{\ln(1 - \frac{1}{n})} \Leftrightarrow k \geq \frac{\ln 0.05}{-\frac{1}{n}}$$
$$\Leftrightarrow k \geq -n \ln 0.05$$

Therefore, a $\mathcal{O}(n)$ independent trials are sufficient to ensure that the probability of success is at least 95%.

3 Node similarity for representation learning

3.1 Question 7

The matrix $P = D^{-1}A$ can be thought of as a transition probability matrix in a random walk on the graph, where D is the degree matrix and A is the adjacency matrix. The element P_{ij} gives the probability of moving from node i to node j in a single step of the random walk. Therefore, the element $(P^k)_{ij}$ gives the probability of reaching node j from node i in exactly k steps. And then, the factor α^k discounts the influence of longer paths in the similarity measure. Since $0 < \alpha < 1$, the longer the path (i.e., the larger the value of k), the less it contributes to the overall similarity. By summing over all powers of k, the definition of S_{ij} considers paths of all lengths, but with diminishing weights for longer paths. This infinite series converges because $\alpha < 1$ and $\|P\| \leq 1$, given that P is a probability matrix. The resulting similarity measure S_{ij} captures not just the direct connections (as given by the adjacency matrix A) but also the global structure of the graph by incorporating the effect of paths of all lengths.

3.2 Question 8

TODO

4 Spectral graph analysis

4.1 Question 9

TODO

4.2 Question 10

TODO

4.3 Question 11

TODO

5 Programming task

5.1 Question 12

TODO

A Appendix