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 $S = Y^TY$. 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:

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

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

The matrix formula of S is $S = \sum_{k=1}^{\infty} \alpha^k P^k$. Then, we have:

$$\begin{aligned} S &= \alpha P + \alpha^2 P^2 + \alpha^3 P^3 + \alpha^4 P^4 + \dots \\ \alpha S P &= \alpha^2 P^2 + \alpha^3 P^3 + \alpha^4 P^4 + \dots \\ S - \alpha S P &= \alpha P \\ S(I - \alpha P) &= \alpha P \\ S &= \alpha P (I - \alpha P)^{-1} \end{aligned}$$

4 Spectral graph analysis

4.1 Question 9

We have G a undirected d-regular graph therefore every vertex contributes to exactly d edges. Let $x \in \mathbb{R}^{|V|}$, we have:

$$\begin{aligned} x^T L x &= x^T x - \frac{1}{d} x^T A x \\ &= \sum_{i=1}^{|V|} x_i^2 - \frac{1}{d} \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} x_i A_{ij} x_j \\ &= \sum_{i=1}^{|V|} x_i^2 - \frac{2}{d} \sum_{(u,v)\in E} x_u x_v \\ &= \frac{1}{d} \sum_{(u,v)\in E} (x_u^2 + x_v^2) - \frac{2}{d} \sum_{(u,v)\in E} x_u x_v \\ &= \frac{1}{d} \sum_{(u,v)\in E} (x_u - x_v)^2 \end{aligned}$$

4.2 Question 10

Since we have for any vector $x \in \mathbb{R}^{|V|}$ that $x^T L x = \frac{1}{d} \sum_{(u,v) \in E} (x_u - x_v)^2 \geq 0$. Therefore, we have that L is positive semi-definite by definition. We also have that L is symmetric because $L = I - \frac{1}{d} A$ and A is symmetric.

4.3 Question 11

A trivial vector x_* is a vector that satisfies a trivial solution. Here, a vector x_* composed of only the same number is a trivial vector because the value will always be 0 since $x_u = x_v$ for all $(u, v) \in E$. Therefore, a trivial vector is a vector that obviously minimize by 0 the expression $x^T L x$.

If x_* minimizes the expression $x^T L x$, this means that x_* corresponds to the eigenvector of the normalized Laplacian L associated with its smallest non-zero eigenvalue (since the smallest eigenvalue is always zero for any graph with at least one edge and its corresponding eigenvector is the constant vector). Therefore the eigenvector corresponding to the second smallest eigenvalue of the Laplacian (after the trivial eigenvalue of zero) provides a meaningful embedding: it captures the most significant ”cut” of the graph. If you were to split the graph into two parts based on the sign of the components of this vector, it would represent a cut that minimizes the number of edges between the two parts (i.e., a sparse cut). Using Equation (1) from question 2B.9, we can justify that x_* as an embedding respects the graph structure since $x^T L x$ is minimized. This indicates that the sum of the squared differences $(x_u - x_v)^2$ across edges is as small as possible, which implies that connected nodes have similar values in the embedding. Hence, x_* embeds nodes in a way that reflects their connectivity in the graph, with the embedding values for strongly connected nodes being close to each other and those for weakly connected nodes being farther apart.

5 Programming task

5.1 Question 12

TODO

A Appendix