ELL888 Minor

Ayush Singh

February 22, 2022

1 Problem 1 : Missing Values

1.1 Idea

Since we have missing values, we try multiple methods: either we impute missing values or we change the distance(dissimilarity) function.

1.2 Basics

In this question and in the following ones as well, I have used EEG dataset with 14 attributes and 2 output classes. I have used KNN graph construction method to construct connectivity matrix($E_{ij} = 1$ or 0) and then Gaussian Weights construction using the above connectivity matrix.

To evaluate the efficiency of a constructed graph, I have used two metrics, (i) Neighbourhood Efficiency of a node: Fraction of neighbours of the node which belong to the class of the neighbour.

(ii) Classification Efficiency of a node: If one was to classify this node using the majority of the classes of its neighbours, then if the classification is correct it gets efficiency 1, otherwise 0.

So, as a whole, the efficiencies of graphs as defined as:

(i) Neighbourhood Efficiency : Average Neighbourhood efficiency over all nodes

(ii) Classification Efficiency : Average Classification efficiency over all

1.3 Method 1: new dissimilarity function

We define a new dissimilarity function for attributes with missing values: For two data points $x_i, x_j \in R^d$, with $N_{ij} = \{m \in \{1, 2, .., d\} | x_{im} \text{and} x_{jm}$ are both not missing$\}$

Then, special Distance between i and j is, $d_s(i, j) = (\sum_{m \in N_{ij}} (x_{im} - x_{jm})^2)^{1/2}$