TMDC PARAMETERS

**Data preprocessing parameters**

**imputation**: String. The method of imputating the missing values in the data file. (Default: “zeros”).

“zeros”: Impute the missing values with zeros.

“median”: Impute the missing values with the median value.

“knn”: Impute the missing values using k-Nearest Neighbors.

“none”: No imputation of the missing values.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**scaling**: String. The method of scaling the values in the data file. (Default: “minmax”).

“minmax”: Transform features by scaling each feature to a given range.

“standard”: Standardizes features by removing the mean and scaling to unit variance.

Advanced:

* The user can specify the range for the “minmax” scaling (Default: [0-1]) by specifying the range in the line ~143 in main.py.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**dim\_red**: String. The options to apply or not, a dimensionality reduction technique on the data before it is fed into TMDC’s training. (Default: “pca\_auto”)

“pca\_auto”: Use of Principal Component Analysis (PCA) with 2 principal components.

“pca\_elbow”: TMDC automatically chooses the optimal no. of principal components based on the elbow rule on the normalized PCA explained variance plot. The selection is made based on the elbow point of 45 degrees to the x axis.

“t-sne”: Use of the t-distributed Stochastic Neighbor Embedding (t-SNE).

“umap”: Use of the Uniform Manifold Approximation and Projection (UMAP).

“ica”: Use of the Independent Component Analysis(ICA).

“none”: No dimensionality reduction.

Tips

* Applying dimensionality reduction hugely improves training time.

Advanced

* With the “pca\_auto” parameter value, the user can also manually select, if they so choose to, the no. of principal components, by changing the “2” in the source code (line ~83 in dim\_red.py) with the desired number.
* The user can plot the PCA explained variance plot through the source code and obtain more visual information by setting the parameter “show\_pca\_plot” to True (line ~21 in dim\_red.py).
* The user can modify each technique’s parameters as desired in the dim\_red.py

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Neural network training parameters**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**distance**: String. The mathematical distance to be used for the distance calculations. (Default: “euclidean”).

“euclidean”: Use of the Euclidean distance.

“manhattan” Use of the Manhattan distance.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**neighbors**: Boolean. The option to also update or not the weights of the neighbors of the BMU neuron. (Default: True)

True: Updates the weights of the neighbors of the BMU neuron.

False: Only the BMU weights will be updated.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**epochs**: Positive integer. The number of epochs of the neural network training. It determines how many times the whole dataset will be seen by the network. (Default: *5000*)

Tips

* Literature suggests at least 500 iterations for every neuron [Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*(1-3), 1-6].

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**lr**: Positive float. The value of the initial learning rate of the neural network training. It adjusts how much the neuron weights will be altered to mimic the data points. (Default: 0.3)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**n\_neurons:** The number of neurons of the neural network. It determines the number of clusters. (Default: -1)

Positive (>=2) integer: User specified number of neurons.

-1: Automatic selection of the no. of neurons based on the elbow rule on the normalized Sum of Squared Errors (SSE) plot. The selection is made based on the elbow point of 45 degrees to the x axis. The number of neurons up to which the training will run is determined by the “max\_n\_neurons” parameter.

Advanced

* The user can plot the SSE plot through the source code and obtain more visual information by setting the “show\_sse\_plot” to True (line ~47 in auto\_neuron\_number\_selection.py).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**max\_n\_neurons:** Positive (>2) integer.The number of neurons up to which the automatic neuron selection will run. (Default: 8)

Tips

* This value cannot be greater than the number of sample class labels in the dataset.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**neuron\_init**: String. The neuron initialization technique. (Default: “points”)

“random”: The neuron weights are initialized randomly from a uniform distribution based on the min and max values in the fed data.

“points”: Every neuron is initialized as a randomly selected existing data point.

No same data point can be selected for two neurons.

By selecting “points”, TMDC randomly selects data points equal to the no. of neurons and calculates their average in-between Euclidean distance. The combination with the highest average distance is selected as the chosen data points to initialize the neurons. This way, TMDC initializes the neurons by trying to utilize the largest possible span in the time-related, high dimensionality space. The number of different combinations to be calculated is dictated by the “depth” parameter.

Tips

* Due to stochasticity, a good practice is to redo the training a couple of times for whichever neuron initialization technique.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**depth:** Positive integer.The number of different combinations for TMDC to calculate the distance dispersion in order to initialize the neurons. (Default: 10000)

Positive integer: The number of different combinations.

“auto”: It automatically selects and calculates every possible combination.

Tips

* Very high values (> 500000) can result in very high running times. This depends on the user’s hardware and the number of neurons.
* High values of this parameter while also using n\_neurons = -1, can result in extremely high running times.
* Higher values result in a higher chance that the algorithm will choose the optimal data points for neuron initialization (depending on the size of the dataset).

Advanced

* Being able to select and calculate the dispersion of every possible combination of data points, means the exact same initialization for every time TMDC is used. This eliminates the stochastic nature of neuron initialization and results in better consistency.
* The number of possible combinations without repetitions of a given dataset with samples and number of neurons is  . For a given dataset, if this number does not exceed the computational threshold, depending on the hardware (e.g., 500000), it is advised to use the “auto” parameter value for the best consistency.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**t1**: Positive integer. Constant value, which controls the exponential decrease of the learning rate. (Default: =epochs / 2)

Tips

* A good practice is to solve the *exp* decrease function (equation 2 in “TMDC Equations”) for different t1 and i (current iteration) values and assess the value of the learning rate at that iteration.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**t2**: Positive integer. Constant value, which controls the exponential decrease of the neighborhood function. (Default: =epochs)

Tips

* Higher t2 corresponds to higher and more lasting cooperation between neurons. Solving the neighborhood function (equation 3 in “TMDC Equations”) can give insight to the constant’s desired value.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**verbose**: Positive integer. Determines the way the output will be presented to the user. (Default: 2)

0: No console output.

1: Prints out a list of the cluster labels for all data points.

2: Prints out the clusters along with every label member.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**random\_state**: Positive integer (<=10000). Determines the random state of the algorithm. Choosing a value eliminates the stochastic nature of the “points” neuron initialization and chooses the same combinations every time TMDC is run as a whole. (Default: a random value each time TMDC is run)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_