ID:71330708 **Empirical validation**

The JLLemma introduces a concept allowing the transformation of high-dimensional datasets into lower-dimensional Euclidean spaces while controlling the distortion in pairwise distances. The theoretical findings of this lemma lead to two key conclusions,

- The minimal number of dimensions (n_components) increases logarithmically with the growth of the number of samples (n_samples) to ensure an epsilon (eps) embedding.
- The minimal number of dimensions (n_components) for a given number of samples significantly decreases with an increase in admissible distortion.

To empirically validate the theoretical bounds mentioned earlier, experiments were conducted on both the 20 newsgroups text document dataset and the digits dataset. A sparse random matrix was employed to project 500 documents, each with 100k features. This was done using various values for the target number of dimensions (n_components). Similarly, for the digits dataset, data from 500 handwritten digit pictures, each containing 8x8 gray level pixels, was randomly projected to spaces with different larger numbers of dimensions (n_components). These experiments aimed to validate the empirical implications of the JL lemma, demonstrating the practical application of sparse random matrix projections in reducing dimensionality for diverse datasets.

1. **Data Transformation:** The SparseRandomProjection function is used to reduce the dimensionality by projecting the original one into a sparse random matrix. This is another alternative to use gaussian random projection.
2. **Matrix Creation:** The fit_transform function is used to fit the data and transform it. It returns a transformed version of data. And then the random matrix is created and the size of the random matrix is displayed.
3. **Euclidean Distances:** Evaluating Euclidean distances using the euclidean_distances function.
4. **Graph Plotting:** Plotting graphs based on min and max distances and calculating mean distance rates. The figure is plotted by using the following functions
    1. plt.xlabel(), used to label the x-axis.
    2. plt.ylabel(),used to label the y-axis.
    3. plt.title(),used to set the title of the graph.
    4. plt.figure(),used to plot the entire figure with the above specifications.
5. **The mean distance:** The mean distance rate was calculated by determining the ratio of projected distances to original distances. The function np.mean(rates) is an inbuilt function which calculates the mean of rates.

The results revealed that for lower values, distortion is more pronounced, while for larger values, distortion is controlled, and distances are effectively preserved. This process was repeated for different random projections (300, 1000, and 10000). The results revealed that for lower values, distortion is more pronounced, while for larger values, distortion is controlled, and distances are effectively preserved. Therefore, the random projections on digits dataset has no dimensionality reduction whereas in 20 newsgroups the dimensionality is decreased with preserving pairwise distances. Regardless of no.of features as per JLLemma Projecting 500 samples requires at least several 1000 dimensions.