Authors: Wojciech Ptaś (7042843), Mhd Jawad Al Rahwanji (7038980)

Problem 2 - Different flavors of embeddings

1. PCA is the process of finding the principal components which we can use to either visualize the data or derive alternative predictors for regression. The PCs are orthogonal axes that feature the most variance in the data. PCA captures the pattern in the data and preserves the global linear structure.

   MDS aims to map data points to a low-dimensional manifold. The mappings are found by minimizing a stress function S using gradient descent. It only requires similarities or dissimilarities between observations. It preserves the pairwise distances (dissimilarities) between data points in the mappings/representation.

   tSNE is an SNE variant that is easier to optimize (new cost function), produces better results and avoids the crowding problem by using the t-distribution for low-dimensional representations. It preserves the neighborhood graph; it optimizes low-dimensional coordinates such that their neighborhood is as similar as possible to its high-dimensional counterpart leveraging GD and KL divergence (distribution distance measure).
2. PCA is highly sensitive to variables' scales. Since PCs are linear combinations of the original variables that maximize variance, a variable with values in the thousands or millions will overshadow other variables in terms of variance, hence, results will greatly differ between scaled and unscaled data.

   MDS, unlike PCA, isn't robust to new data entries. If new data is to be added then MDS must be recomputed.

   tSNE is only good for data visualization, meaning as long as we're mapping to 2 or 3 dimensions. When our purpose is dimensionality reduction ($d > 3$) then this approach will cease to preserve the local structure of the data as well.

References

❖ Hinton GE, Roweis S. Stochastic neighbor embedding. Advances in neural information processing systems. 2002;15.

❖ Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008 Nov 1;9(11).