EE769 Introduction to Machine Learning (Jan 2021 edition)

Electrical Engineering, Indian Institute of Technology Bombay

Programming Assignment – 3 : Unsupervised Learning

Instructions:

a) Only submit ipython notebooks. The notebook should be a complete code plus report with copious comments, references and URLs, outputs, critical observations, and your reasoning to choose next steps.
b) Use good coding practices such as avoiding hard-coding, using self-explanatory variable names, using functions (if applicable). This will also be graded.
c) Cite your sources if you use code from the internet. Also clarify what you have modified. Ensure that the code has a permissive license or it can be assumed that academic purposes fall under 'fair use'.

Problem statements:

1. Clustering:
    a. Visualize and pre-process the data as appropriate from the file DataClustering.csv. You might have to use a power, an exponential, or a log transformation.
    b. Train k-means, and find the appropriate number of k.
    c. Train DBSCAN, and see if by varying MinPts and ε, you can get the same number of clusters as k-means.
    d. Using the cluster assignment as the label, visualize the t-sne embedding.
2. PCA:
    a. Visualize the data from the file DataPCA.csv.
    b. Train PCA.
    c. Plot the variance explained versus PCA dimensions.
    d. Reconstruct the data with various numbers of PCA dimensions, and compute the MSE.
3. Non-linear dimension reduction:
    a. Visualize the data from the file DataKPCA.csv.
    b. Train KPCA.
    c. Plot the variance explained versus KPCA dimensions for up to 10 dimensions.