EE769 Introduction to Machine Learning (July 2024 edition)

Electrical Engineering, Indian Institute of Technology Bombay

Programming Assignment – 4 : 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'.
d) Submit a link to a viewable 10 minute video walk through of your code and insights

Problem statements:

Data: https://www.kaggle.com/datasets/alirezachahardoli/customer-data-clustring

Objective: Derive customer insights based on their credit card use features

1. Data preprocessing: [2]
    a. Visualize and pre-process the data as appropriate. You might have to use a power, an exponential, or a log transformation.
    b. You may find and drop some of the highly correlated or inappropriate variables, or encode discrete variables as appropriate
2. Clustering: Try to find meaningful customer segments using clustering [4]
    a. Train k-means, and find the appropriate number of k.
    b. Train DBSCAN, and see if by varying MinPts and ε, you can get the same number of clusters as k-means.
    c. Using the cluster assignment as the label, visualize the t-sne embedding.
    d. Try to give each cluster a name, such as "reckless spenders"
3. PCA: Try to find if there are only a few components/directions that explain most of the variance in the data. [3]
    a. First, normalize each variable independently. Then Train PCA on appropriate variables.
    b. Plot the variance explained versus PCA dimensions.
    c. Reconstruct the data with various numbers of PCA dimensions, and compute the MSE.