EE769 Introduction to Machine Learning (Jan 2024 edition)

Electrical Engineering, Indian Institute of Technology Bombay

Programming Assignment – 3 : Unsupervised Learning

Instructions:

a) Only submit ipython notebooks and link to video demo (10 minute). 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. Extract ResNet-18 features for Ants vs. Bees training data from https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html . Visualize and pre-process the data as appropriate. [1]
    b. Train k-means, and find the appropriate number of k using the silhouette method. [1]
    c. Visualize a few samples in each cluster to see what is common in images from a cluster. [0.5]
    d. Examine cluster purity in terms of whether each cluster is either predominantly ants or bees, or a mix of both? [0.5]
    e. Train DBSCAN, and see if by varying MinPts and ε, you can get the same number of clusters as k-means. [1]
    f. Find out if there is a good metric for DBSCAN clustering, and if so, then use it. [1]
    g. Repeat part c and comment on which one is more appropriate clustering method for this data. [0.5]
    h. Examine cluster purity in terms of whether each cluster is either predominantly ants or bees, or a mix of both? [0.5]
    i. Using the ants or bees label to color the samples and visualize the t-sne embedding in 2D. [1]
2. PCA:
    a. Train PCA on the data from the previous question. [1]
    b. Plot the variance explained versus PCA dimensions. [1]
    c. Reconstruct the data with various numbers of PCA dimensions, and compute the NRMSE. [1]
3. Non-linear dimension reduction:
    a. Train KPCA on the data from Q1. [1]
    b. Plot the variance explained versus KPCA dimensions for up to 10 dimensions. [1]