Basic Machine Learning Bootcamp - Week 13

Q1: Load the MNIST dataset and split it into a training set and a test set (take the first 60,000 instances for training, and the remaining 10,000 for testing). Train a random forest classifier on the dataset and time how long it takes (use %time), then evaluate the resulting model on the test set. Next, use PCA to reduce the dataset's dimensionality, with an explained variance ratio of 95%. Train a new random forest classifier on the reduced dataset and see how long it takes. Was training much faster? Next, evaluate the classifier on the test set. How does it compare to the previous classifier? Try again with an SGDClassifier. How much does PCA help now?

Q2: Use t-SNE to reduce the first 5,000 images of the MNIST dataset down to 2 dimensions and plot the result using Matplotlib. You can use a scatterplot using 10 different colors to represent each image's target class. Alternatively, you can replace each dot in the scatterplot with the corresponding instance's class (a digit from 0 to 9), or even plot scaled-down versions of the digit images themselves (if you plot all digits the visualization will be too cluttered, so you should either draw a random sample or plot an instance only if no other instance has already been plotted at a close distance). You should get a nice visualization with well-separated clusters of digits. Try using other dimensionality reduction algorithms, such as PCA, LLE, or MDS, and compare the resulting visualizations.

Q3: A hybrid method in feature selection combines multiple feature selection techniques to improve the overall performance and robustness of the selection process. Here's an example of a hybrid method. You are given a dataset (train,csv) with a large number of features and we want to select the most relevant ones for a classification task. We can create a hybrid feature selection method by combining a variance threshold approach with a filter-based approach and a wrapper-based approach.

Step1: Dropping Constant Features using VarianceThreshold

Step2: Apply a filter-based approach to rank the features based on their individual relevance to the target variable. This can be done using statistical measures such as chi-square, ANOVA, correlation, or mutual information.The filter-based approach helps us identify a subset of features that have a high correlation with the target.

Step3: using wrapper-based method and evaluate the selected subset of features.