ELG 5255 Applied Machine Learning Fall 2021

Assignment 1 (Dimensionality Reduction, Feature Selection and Unsupervised Learning)

Submission

You must submit your assignment on-line with Bright Space. This is the only method by which we accept assignment submissions. We do not accept assignments sent via email, and we are not able to enter a mark if the assignment is not submitted on Bright Space! The deadline date is firm since you cannot submit a assignment passed the deadline. It is student's responsibility to ensure that the assignment has been submitted properly. A mark of 0 will be assigned to any missing assignment.

Goal

The goal of this assignment is to apply Dimensionality Reduction and Feature Selection methods along with the unsupervised methods.

Dataset

During this assignment, pokemon dataset is used. The provided dataset is preprocessed so please use the provided file.

In python:

```
import pandas as pd

pokemon_train = pd.read_csv("file path/Pokemon_train.csv")
pokemon_test= pd.read_csv("file path/Pokemon_test.csv")
```

Questions

Please submit your code and report (including screen shot of code and the relevant figures).

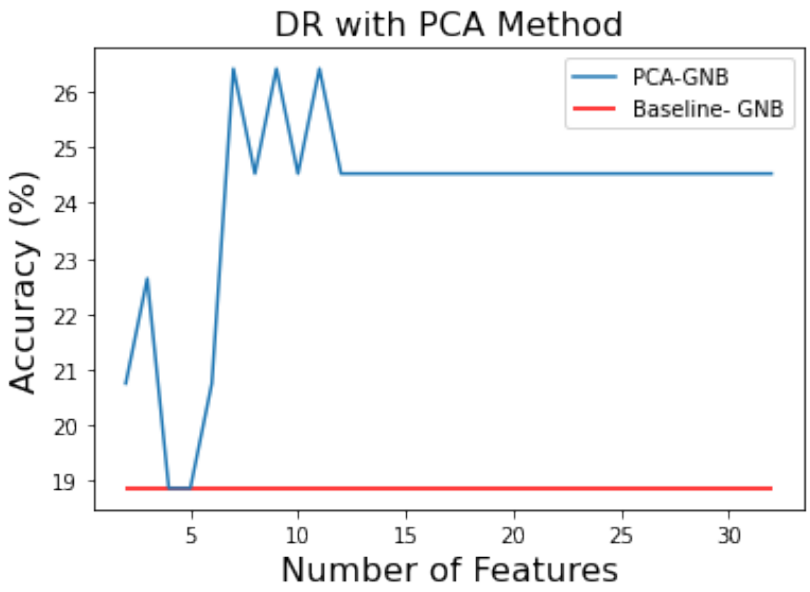
1. Load the Pokemon dataset and split features and labels for training and test sets.

2. Apply Gaussian Naïve Bayes classifier (GNB) and Support Vector Machine (SVM) to Pokemon dataset. (**10 marks**)

- Provide the accuracy of GNB and SVM classifier as baseline performances
- Apply TSNE(n_components=2, random_state=0) to training and test set and visualize

3. Choose the best number of cluster for k-means clustering algorithm(**15 marks**)
    - Using the elbow rule, plot the distortion score (a.k.a inertia) vs the number of clusters
    - Determine the optimal number of clusters for k-means
    - Plot clustered data with optimum number of cluster

4. Apply one of the following Dimensionality Reduction (DR) methods to data. Find the best value for n_components based on the GNB and SVM classifiers test accuracies. Plot the number of features-Accuracy graph with baseline performances for each classifier as shown below. DR method should be applied training set, and test set should be transformed accordingly. Graph should be plotted based on the test accuracy.(**15 marks**)
    - Principal Component Analysis, PCA(n_components=n, random_state=0)
    - Linear Discriminat Analysis, LDA(n_components=n, random_state=0)



Note: If there are multiple same maximum value, choose the lowest number of features.

5. Use the following Feature Selection methods (one for each method). Find the best number of features based on the GNB and SVM classifiers' test accuracies on the data that is obtained after Q4. Plot the number of features versus accuracy graph for each case with the improved baseline performance as shown in Q4.(**20 marks**)
    - **Filter Methods** (Information Gain, Variance Threshold etc.)
    - **Wrapper Methods** (Forward or Backward Feature Elimination, Recursive Feature Elimination etc.)

6. Choose the best number of cluster for k-means clustering algorithm on the processed data (using the best features or dimensionality from Q4 and Q5)(**10 marks**)
    - Using the elbow rule, plot the distortion score (a.k.a inertia) vs the number of clusters
    - Determine the optimal number of clusters for k-means

7. Choose the best number of neurons for SOM algorithm using the best features or dimensionality from Q4 and Q5. (**20 marks**)
    - Using the elbow rule, plot the distortion score (a.k.a inertia) vs the number of neurons (max 30 neurons)
    - Determine the optimal number of neurons for SOM
    - Plot the initial and final Neuron positions

8. Tune the epsilon (0.2-3) and minpoints (2-15) values in the given intervals to obtain same number of clusters in Q6 by using DBSCAN. Plot the epsilon and minpoints values using a 3D figure to show the best 10 combination of epsilon and minpoints that brings you closer to the desired cluster number. The z-axis should illustrate the number of clusters you obtain when using the corresponding epsilon and minpoints. (**10 marks**)

    Note: If you cannot find the parameters to obtain desired number of clusters, you can use the parameters of the closest number to desired number of cluster to plot the figure.