PGCP DSML 08

Module: Machine Learning

Assignment-1 (synthetic datasets)

1) Generate the 2D data using the following piece of code:

```
import numpy as np
import matplotlib.pyplot as plt
rng = np.random.RandomState(1)
x = 10 * rng.rand(50)
y = 2 * x - 5 + rng.randn(50)
plt.scatter(x, y);
```

   i) Build a linear regression model to fit the above generated data and plot the function. Also calculate MSE and R2 score of the fitting.
   ii) Build a linear-SVM regression on the above generated data and plot the function. Also calculate MSE and R2 score of the fitting.

2) Generate a synthetic datasets having four classes and 1000 samples using the following piece of code:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import make_blobs
X, y = make_blobs(n_samples=1000, centers=4,
             random_state=0, cluster_std=1.0)
```

   i) Show the scatter plot of the dataset (each class patterns from a different color)
   ii) Split the dataset into 75% training and 25% testing patterns.
   iii) Apply standard scaler function to normalize the data as per normal distribution.
   iv) Apply softmax regression to classify the data into respective classes. Make use of an appropriate function to show the decision boundaries. Also show the classification report and Confusion matrix.
   v) Apply nonlinear-SVM to classify the test data into respective classes. Make use of RBF kernel. Plot the decision boundaries. Also show the classification report and Confusion matrix.
   vi) Apply k-NN to classify the test data into respective classes. Comments on the best suitable value of the parameter “k”.

3) Generate a binary classification dataset with 1000 samples in 2D which is having 95% patterns from negative class and the rest 5% patterns from the positive class. Split the training and testing sets. Apply Logistic regression model and k-NN algorithm to classify the testing dataset and print their respective classification reports. Apply an oversampling method to balance this imbalanced dataset. Again, Split the training and testing sets from the oversampled data. Apply Logistic regression model and k-NN algorithm (with same parameters) to classify the testing dataset and print the classification report. Make your conclusion on these two classification results.

-End of the Assignment-