Group 11 (19CS10071, 19CS30041)
Anurat Bhattacharya
Sayantan Saha

Machine Learning Assign 3

13th November 2021

OVERVIEW

The task is to implement and use a Support vector machine(SVM) Classifier for Occupancy Detection from the database https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+ and to plot the reduced dimensional data from Principal Component Analysis(PCA) and Linear Discriminant Analysis(LDA)..

GOALS

1. First, we need to read the data.
2. Then, the data read is prepared to be fed for the implementation like reducing the data into two-dimensional feature space in case of PCA and reducing the data into one-dimensional feature space in case of LDA after train, validation and test split.
3. Thereafter, an SVM classifier is trained on the dimensionally reduced data(by PCA or LDA) after trying different kernel types by varying the appropriate hyperparameters of the classifier.
4. The accuracy of the classifier with the best kernel is to be determined.

Description of the dataset

The dataset given to us contains following attributes:

1. date: It is a time series data. As of date, it contains the time year-month-day hour-minute-second. We break up this field into *day* and *time* as the data is only of one month only the day and the time matters.
2. Temperature: It contains the temperature of that time in celcius.
3. Humidity: It contains the relative humidity of that time in percentage.
4. Light: This contains measurements of lights at a time in the ‘Lux’ unit.
5. Humidity Ratio: It contains derived quantity from temperature and relative humidity, in ‘kg water-vapor/kg-air’ unit.

6. CO2: Measurements of CO@ in the environment at a particular time in the ‘ppm’ unit.
7. Occupancy: It is the data which is used as dependent on the other attributes. The SVM classifier is used to predict this and test the accuracy. Occupancy is 1 if it is in occupied status and 0, otherwise.

Our aim is to use the date(only the day),time,temperature, humidity, light, CO2, humidity ratio feature as input attributes and predict the occupancy using the SVM classifier model .

Steps to Run:

1.Run pip install -r requirements.txt

2.Run python mlassgn3_Grp11.py

PROCEDURE

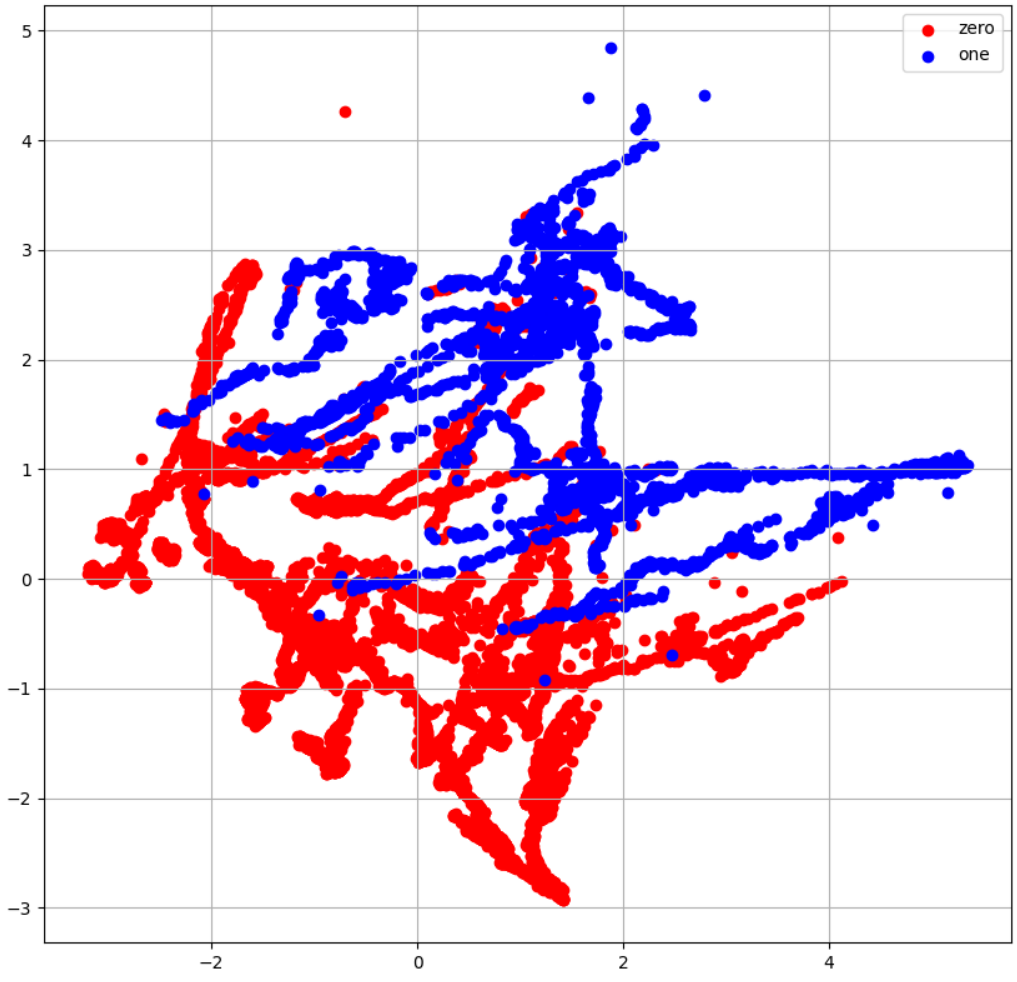
1. There are three text files ‘datatest’, ‘datatest2’, ‘datatraining’ with no proper ratio for training, validation and test set. So, we combine the total 20560 number of instances from the three text files as our preprocessed data set. The date field is then split into time and date(with only the day).
2. This combined data set is normalized first with respect to each attribute barring the attribute ‘occupancy’ which is to be predicted. This normalized data is split into training, validation and test data as 70:10:20 ratio.Now the data is processed and ready for being used in the SVM classifier model.
3. Now the processed training data is reduced into two dimensional feature space using Principal Component Analysis(PCA)(components determined from train set). The reduced dimensional data of the train split is plotted in a two-dimensional plane.
4. The reduced dimensional data generated by the PCA is used to train an SVM classifier using sklearn. The appropriate hyperparameters are varied by trying different kernel types. Then the classification accuracy is computed in the validation set. The validation accuracies are tabulated for each of these combinations.The kernel with the highest validation accuracy is chosen and test accuracy is computed using that kernel.
5. A copy of the processed data from step-2 is kept for this part. Here we reduce the feature dimension of the training data into a one-dimensional feature space using LDA. Then the reduced dimensional data of the train split is plotted.
6. Similar to step 4, the reduced dimensional data generated by the LDA (components generated from train state) is used to train an SVM classifier using sklearn. The appropriate hyperparameters are varied by trying different kernel types. Then the classification accuracy is computed in the validation set. The validation accuracies are

tabulated for each of these combinations.The kernel with the highest validation accuracy is chosen and test accuracy is computed using that kernel.

7. Then the final test accuracy of the SVM classifier after reducing feature space by PCA and LDA are compared.

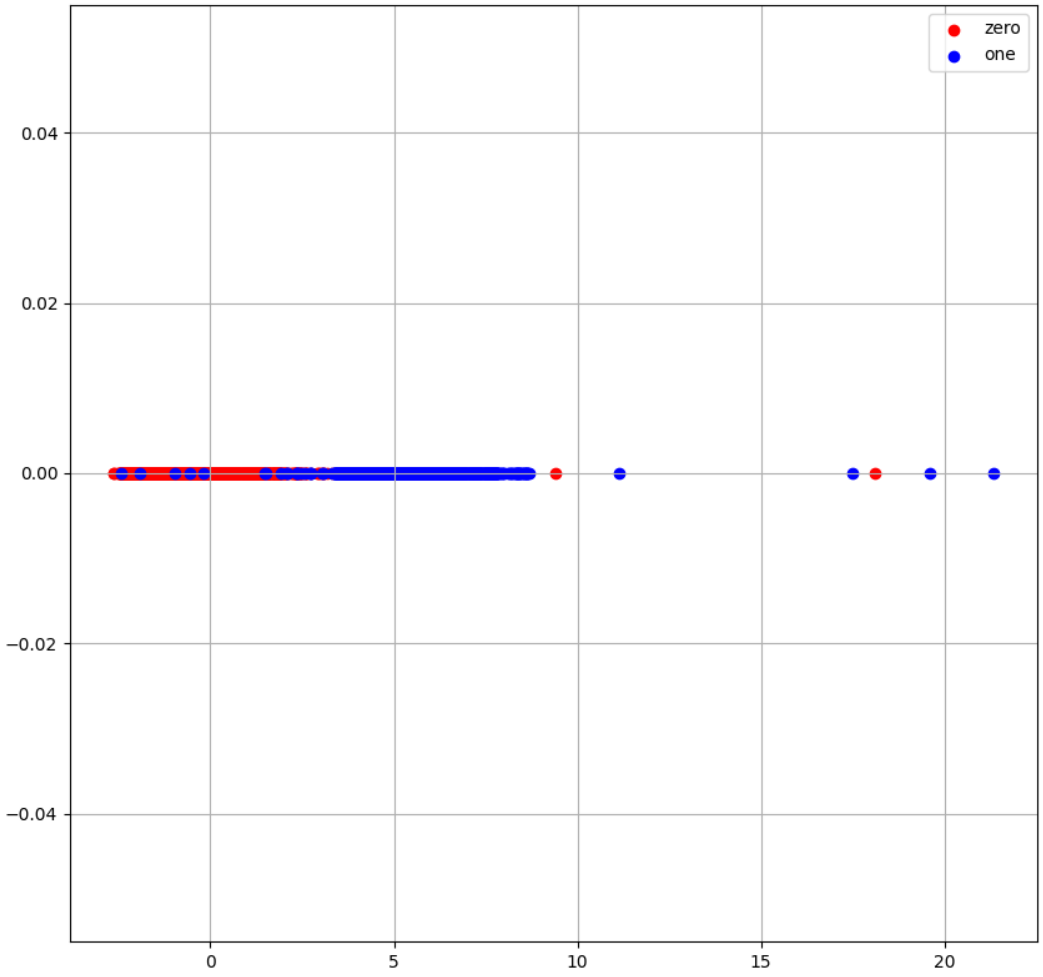
RESULTS :

PCA



Kernal Type	Validation Accuracy(%)	Test Accuracy
Linear	92.31	
Polynomial Degree 2	86.77	
Poly Degree 3	91.43	
Poly Degree 4	88.52	
Radial Basis	94.94	95.622
Sigmoid	83.75	

LDA



Kernal Type	Validation Accuracy(%)	Test Accuracy
Linear	98.69	99.0029
Polynomial Degree 2	98.59	
Poly Degree 3	98.59	
Poly Degree 4	98.54	

Radial Basis	98.69	
Sigmoid	93.677	

```
anurat@anurat-Inspiron-3576:~/ML/mlassgn3/ML_assgn3_19CS10071_19CS30041
$ python3 mlassgn3_Grp11.py
Accuracy with different kernals on validation (PCA) : {'linear': 0.9231
517509727627, ('poly', 2): 0.867704280155642, ('poly', 3): 0.9143968871
59533, ('poly', 4): 0.8852140077821011, 'rbf': 0.9494163424124513, 'sig
moid': 0.8375486381322957}
Kernal taken for test data:  rbf
Test Accuracy with PCA:  0.9562256809338522
Accuracy with different kernals on validation (LDA) : {'linear': 0.9868
677042801557, ('poly', 2): 0.9858949416342413, ('poly', 3): 0.985894941
6342413, ('poly', 4): 0.9854085603112841, 'rbf': 0.9868677042801557, 's
igmoid': 0.9367704280155642}
Kernal taken for test data:  linear
Test Accuracy with LDA:  0.9900291828793775
```

Inference:

Different Kernels :

While using PCA we see that radial basis function has highest validation accuracy of 94.94% and test accuracy of 95.622%.

While using LDA we see that Polynomial f degree 2 gives highest validation accuracy of 98.69% and a test accuracy of 99.0029%.Here the validation accuracies of polynomials,linear and radial basis functions are quite close indicating we have almost reached the max accuracy possible(as is also evident from the high accuracy value of ~ 99%)

LDA vs PCA :

Accuracy obtained using LDA is much higher that that using PCA.This is because PCA focuses on maximizing the accuracy,while the cost metric for LDA is

$J = \frac{D^2}{s_1^2 + s_2^2}$; It maximizes the distance between the two classes and minimizes the individual scatters , thus ensuring better separability.Hence we see a stark increase of accuracy while using LDA.