School of Computer Science Engineering and Technology

Course- BTech
Course Code- CSET301
Year- 2022
Date- 19-09-2022

Type- Core
Course Name-AIML
Semester- Even
Batch- V Sem

Lab Assignment No.5 . 1 . 1

Exp. No.	Name	CO-1	CO-2	CO-3
5.1.1	SVM classifier	✓	✓	--

Objective: To implement Support vector machines (SVMs) (using Scikit-learn) to predict Pulsar Star **(as 0 or 1) and handling of numerical and categorical features.**

Download the dataset from /kaggle/input/predicting-a-pulsar-star/pulsar_stars.csv (10)

About Dataset: Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter. Machine learning tools are now being used to automatically label pulsar candidates to facilitate rapid analysis. Classification systems in particular are being widely adopted, which treat the candidate data sets as binary classification problems. there are 9 variables in the dataset. 8 are continuous variables and 1 is discrete variable. The discrete variable is target class variable. It is also the target variable.

Attribute Information:

Each candidate is described by 8 continuous variables, and a single class variable. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the DM-SNR curve . These are summarised below:

1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.
9. Class

Load the data and print first 10 and last 10 rows, *view dimensions of dataset, and datatype o f each attributes* using a suitable function. (5)
view the column names of the dataframe and check, is there any leading spaces in column n ames, if yes, remove leading space from column names using str.strip() funcrion.

Rename the column names

1. Check the presence of missing values. Handle it if present.
2. Check the presence of Categorical columns. Handle it if present. i.e., Transform categorical features into numerical features. (Hint: Use eitherone hot encoding, label encoding or any other suitable pre-processing technique).
3. Scale the numerical columns value using StandardScaler() function.
4. Check the outliers in continuous variables by plotting box plot
5. Assign all the features of dataframe excluding "target class" feature into feature vector **X** and "target class" feature into **y** vector i.e., target feature. (5)
6. Split the dataset into 80% for training and rest 20% for testing (sklearn.model_selection.train_test_split function) (5)
7. Run SVC classifier of SVM using built-in function on the training set(SVC from sklearn.svm) with following parameters and show the accuracy in each case. (10)

    a) SVM with Default parameter means C=1.0, kernel=rbf and gamma=auto among other parameters)
    b) SVM with rbf kernel and C=100.0
    c) SVM with rbf kernel and C=1000.0
    d) SVM with linear kernel and C=1.0
    e) SVM with linear kernel and C=100.0
    f) SVM with linear kernel and C=1000.0
    g) Choose option b and e with gamma =100
    h) Choose option c and f with gamma =1000

    Note: 1) Low C implies we are allowing more outliers and high C implies less outliers.
    2) The linear, polynomial and RBF or Gaussian kernel are simply different in case of making the hyperplane decision boundary between the classes. The kernel functions are used to map the original dataset (linear/nonlinear) into a higher dimensional space with view to making it linear dataset.
    3) gamma defines how much influence a single training example has. The larger gamma is, the closer other examples must be to be affected. Gamma high means more curvature, Gamma low means less curvature.

8. Use the trained model to **predict** on the **test set** and then (15)
    a. Print 'Accuracy' obtained on the testing dataset i.e. (sklearn.metrics.accuracy_score function)
    b. Confusion matrix (sklearn.metrics.confusion_matrix),

Suggested Platform: Python: Jupyter Notebook/Azure Notebook/Google Colab.