MACHINE LEARNING ASSIGNMENT

UE17CS303

Course instructor : Dr. Saha

Team Members:

- PES1201701419 Venugopal N B
- PES1201701375 Chetan B C
- PES1201701558 Anirudh Joshi

Classification Of Stars And Quasars Using Supervised Learning

Anirudh Joshi,Chetan B C,Venugopal N B

26th November,2019

1. Abstract

This is a standard application of Machine Learning problem where, given the dataset we are to classify it into two main classes i.e STARS and QUASARS. We estimate the accuracy using KNN Classifier and classify the data into STARS and QASARS for all given catalogs(cat1.csv,cat2.csv,cat3.csv,cat4.csv). the dataset. To evaluate the correctness of the classifiers, we report the accuracy and other performance metrics and find reasonably satisfactory results.

Keywords: Machine Learning, KNN Classifier, Catalog

2. Introduction

The main application of Machine Learning is to classify the data into different categories.For us ,given the dataset we applied K Nearest Neighbors briefed as KNN classifier to classify the data into two main categories STARS and QUASARS.The KNN classifier belongs to Supervised Learning Techniques.In the given catalog the last column is for represtation of classes.We used numeric indication for representing the classes i.e '0' for 'STARS','1' for 'QUASARS'.

3. Procedure

3.1

Step 1:
We split the dataset line-by-line and then again split it column wise and store the resultant in the nested list.So each row is stored as list in a list.

Step 2:
Now we split the data into two categories:
(1)Training Sample
(2)Test Sample
We divide the dataset based on the 'ratio' value given,which takes many values like 0.50 ,0.60, 0.80, 0.90, etc.
Multiplying the 'ratio' number with the number of data samples we get portion of data that is to be grouped as Training Sample and the rest as Test Sample. **For example:** if number of data sample=100 ,ratio=0.80 then 80 of the samples of the catalog are grouped as Training Sample and rest 20 of the data samples as Test Sample. These grouping is done randomly i.e we choose random data samples every time to group them as Test Sample or Training Sample.

Step 3:
We take the first sample from Test Sample and all samples from Training Samples and find Euclidean Distance between those points. We sort them based on the distances in ascending order and then we choose first 'k' values and find the the maximum repeated class.Once we find that we assign that Test Sample to that class which is the predicted value. Now,we choose next sample from Test Sample and follow the same Procedure above. We repeat the whole process for all of the Test Samples.Now we compare the Predicted values with the True values and calculate Accuracy.This is done all of the given Catalogs
We pass the first 'k' samples which are sorted based on Euclidean Distance to a function which returns us whether the sample belongs to class '0' or '1' .This takes first 'k' samples and then counts number of class 0's and class 1's, and returns max of them indication that sample belongs to that class.

The $akkarucy(tv.predo)$ function gives us the accuracy of our Predicted values. It takes 'tv' true value , 'predo' Predicted values as arguments and compares and returns accuracy of it. The formula is :

$$Accuracy = no.ofcorrectpredo * 100/no.oftv \quad (1)$$

Where,
no.ofcorrectpredo :- number of correct Predicted values,
no.oftv :- number of True values

3.2 *K-fold Cross Validation:*

Till now we calculated Accuracy by dividing the given catalog into 'Test sample ' and 'Train Sample' based ratio.Now, we calculate the Accuracy by dividing the whole sample into 'K' groups and iteratively we assign 1-group as Train Sample and rest of the groups as Test Sample.We calculate Accuracy for each of the cases.We find average of all the Accuracy's.This way we take all of groups as Train Sample and calculate the Accuracy, which may turn out to be more precise.
We do it by dividing the catalog into k groups groups.Let's say k=3, so we divide it into 3 groups and choose sample randomly and append the first group as list into a list. We do same to rest of the groups.We are left with a list of list.We first assign 1st group as Train Sample and rest 2 as Test Sample We proceed with same procedure and calculate Accuracy. Next we assign 2nd group as Train Sample and rest 2 as Test sample. We do this until all of the groups are assigned as Train Sample and calculate Average of all the Accuracy's

3.3 *Up Sampling / Down Sampling:*

Given a catalog with errors in it, we need to filtrate it or fix it. We use Up Sampling/Down method to do it. We divide the catalog into Train Sample and Test Sample. We take Train sample and calculate number of each classes. The class with minimum count is replicated till both the classes are are of same number.This method is called Up Sampling,whereas reducing the number of classes of maximum one's is called Down Sampling. The need is to balance both classes if they are imbalanced else our precise will be less.Once we replicate the classes we follow same procedure to calculate Accuracy. i.e find Euclidean Distance for all samples in Test and Train samples divides and assigning the majority occurred class to that sample.Calculate the Accuracy by comparing with the True value.

3.4 *Graphs:*

In graph pic.1,
The graph distinguishes False Positive Rate Vs False Negative Rate using two different Classifying algorithms.We see that using KNN classifier Algorithm, the area under the curve(AOC) is 0.90 squnits. whereas using Naive-Bayes Classifier algorithm AOC is 0.77. squnits

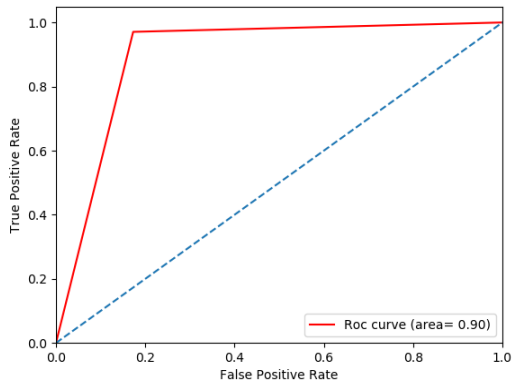
In Graph pic.2:
The Graph plotted between different values of K and respective Accuracy. It shows Accuracy vs k.It is clear that Accuracy is maximum for k=3. So it classifies for K=3. Thus model predicts well for for K value equals to 3.

In pic.3
It is basically represents how samples are changed after Up Sampling/Down Sampling. Before applying Up Sampling we see that there approximately 480 of samples belonging to class '0' and nearly 3200 to class '1'. After Up sampling we see that there are equal number of samples belonging to both the classes.

4. Result:

After successful completion of whole process we must be able to:
(1) Classify the samples to their respective classes Accurately. (2) Measure of Accuracy of classifying the samples for given dataset. (3) How precise is our classification. (4) Important graphs to understand clearly for different constraints

Fig 1.AUC curve of KNN and Naïve-Bayes



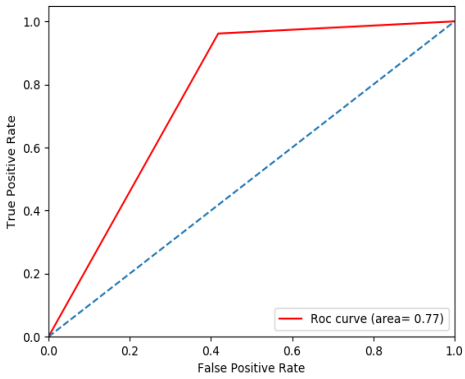


Fig 2. Accuracy vs. K-values for KNN

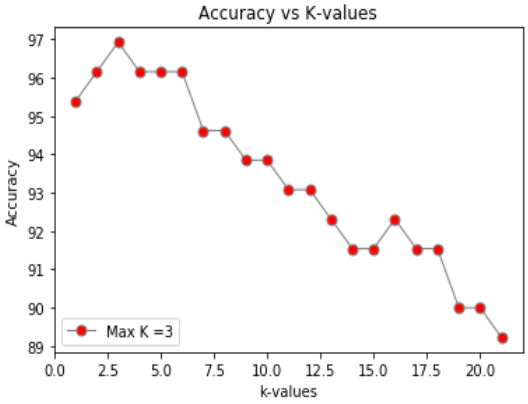


Fig 3. Before Upsampling and after Upsampling Class frequency