Data Mining Assinment 2

Ahmet Oral

November 2020

.

DECLARATION OF HONOR CODE:
Student ID: 180709008
Name: Ahmet
Surname: Oral
In the course of Data Mining (CENG 3521), I take academic integrity very seriously and ask you to do as well. That's why, this page is dedicated to some clear statements that defines the policies of this assignment, and hence, will be in force. Before reading this assignment booklet, please first read the following rules to avoid any possible violation on academic integrity.

- This assignment must be done individually unless stated otherwise.
- You are encouraged to discuss with your classmates about the given assignments, but these discussions should be carried out in an abstract way. That is, you cannot copy code (in whole or in part) of someone else, cannot share your code (in whole or in part) with someone else either.
- The previous rule also holds for the material found on the web as everything on the web has been written by someone else.
- You must not look at solution sets or program code from other years.
- You cannot share or leave your code (in whole or in part) in publicly accessible areas.
- You have to be prepared to explain the idea behind the solution of this assignment you submit.
- Finally, you must make a copy of your solution of this assignment and keep it until the end of this semester.

I have carefully read every of the statements regarding this assignment and also the related part of the official disciplinary regulations of Mugla Sıtkı Kocman University and the Councilof Higher Education. By signing this document, I hereby declare that I shall abide by the rulesof this assignment to prevent any violation on academic integrity.

Signature

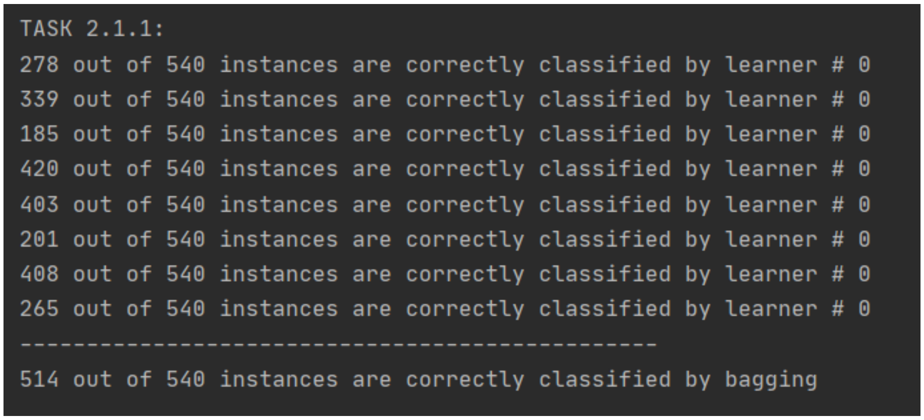
2 Ensemble Learning

2.1 Different training sets

2.1.1 Bagging Pasting

.

In this task, my goal is to create an instance of multi-layer perceptron network with four hidden layers that have 16, 8, 4,2 neurons with more than 1,000 iterations to 1/8 subsets of data set given to each base classifier.So I loaded digits dataset, splitted it 30/70 ratio then I created a MLPClassifier.After that applied BaggingClassifier to it.Then I created a for loop that runs on bcf.estimators_ as a list to get the correctly classified instances of each learner. I took the prediction then I get the accuracy score of it.I used the accuracy score to calculate number of correctly classified instances and printed the results:



```
TASK 2.1.1:
278 out of 540 instances are correctly classified by learner # 0
339 out of 540 instances are correctly classified by learner # 0
185 out of 540 instances are correctly classified by learner # 0
420 out of 540 instances are correctly classified by learner # 0
403 out of 540 instances are correctly classified by learner # 0
201 out of 540 instances are correctly classified by learner # 0
408 out of 540 instances are correctly classified by learner # 0
265 out of 540 instances are correctly classified by learner # 0
------------------------------------------------
514 out of 540 instances are correctly classified by bagging
```

Figure 1: Results of Task 2.1.1

2.1.2 Boosting

In this task I loaded moons data set and add gaussian noise with derivation vale of 0.2 . I splitted the data wit 30/70 ratio and I created an instance of logistic regression algorithm with SGD solver.Then I applied AdaBoost classifier with four base classifiers to it.Then I visualized testing samples and decision boundary that are learned by each base algorithm. Here is the figure:

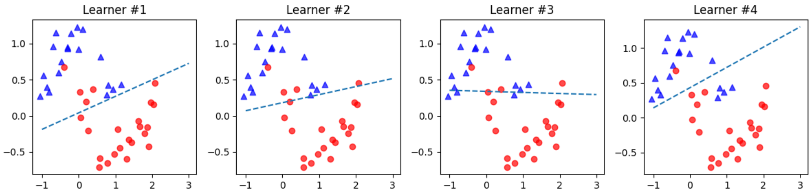
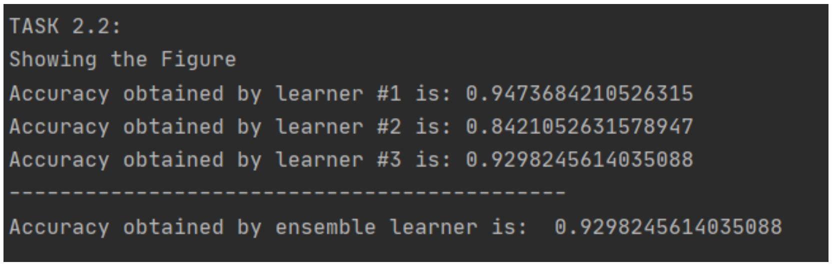


Figure 2: Results of Task 2.1.2

2.2 Different learning algorithm

Here, I started by loading breast cancer data set but this time I split data with 20/80 ratio because in introduction you said; we use an approach that uses 20/80 ratio and you didn't specifically ask for us to use 30/70 ratio.Apply Ensemble learning framework that produce final estimation with respect to the majority of votes returned by base classifiers created with 5-fold cross validation strategy.For the three base classifiers I used LogisticRegression, SGDClassifier and SVC. Here are the results:

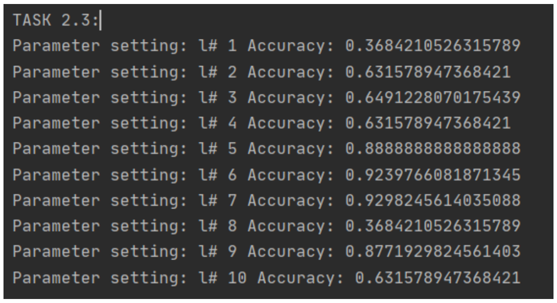


```
TASK 2.2:
Showing the Figure
Accuracy obtained by learner #1 is: 0.9473684210526315
Accuracy obtained by learner #2 is: 0.8421052631578947
Accuracy obtained by learner #3 is: 0.9298245614035088
--------------------------------------------
Accuracy obtained by ensemble learner is:  0.9298245614035088
```

Figure 3: Results of Task 2.2

2.3 Different parameter setting

Here, my task was to apply Ensemble learning framework that produce final estimation with respect to the majority of votes returned by created base classifiers. So I loaded the data, splitted it with 30/70 ratio. Then inside a for loops that runs with range(10) I created MLPClassifier with 10 hidden layer sizes and in order to neurons you asked in assignment.Then I printed the accuracy score.After that I get the score of ensamble learning algorithm I used and printed it. Results are:



```
TASK 2.3:
Parameter setting: l# 1 Accuracy: 0.3684210526315789
Parameter setting: l# 2 Accuracy: 0.631578947368421
Parameter setting: l# 3 Accuracy: 0.6491228070175439
Parameter setting: l# 4 Accuracy: 0.631578947368421
Parameter setting: l# 5 Accuracy: 0.8888888888888888
Parameter setting: l# 6 Accuracy: 0.9239766081871345
Parameter setting: l# 7 Accuracy: 0.9298245614035088
Parameter setting: l# 8 Accuracy: 0.3684210526315789
Parameter setting: l# 9 Accuracy: 0.8771929824561403
Parameter setting: l# 10 Accuracy: 0.631578947368421
```

Figure 4: Results of Task 2.2

3 k-Nearest Neighbors Classifier

In this part you asked us to apply an instance-based classifiers to solve classification problems.I loaded moons dataset with more then 100 tuple size and I used 'noise=0.3' parameter to give Gaussian noise to data with a deviation value of 0.3.I split the data as only four tuples are used for testing while remaining tuples are used for training.I applied KNeighborsClassifier for each testing tuples with k = 5 setting.After that I wrote a function to get nearest k neighbors of that sample.I will use this function to show nearest k neighbors of that sample with green border color.I visualized training samples, corresponding testing sample and nearest k neighbors:

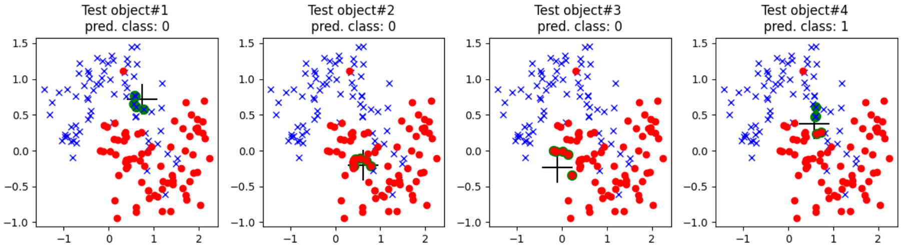


Figure 5: Results of Task 2.2

.

Conclusion

In this assignment I practiced on how data mining tasks such as classification, regression can be handled by ensemble and lazy learning algorithms.My task is to analyse problems such as Different learning algorithms on same problem, same algorithms with different parameters on same problem and different training sets (bagging, pasting, and boosting), different representations of the same input. While doing the assignment I learned and used BaggingClassifier, VotingClassifier and AdaBoostClassifier as well as KNeighborsClassifier. I also practiced to visualize models which was a lot harder for me in previous assignments.In my code I tried to explain my code as clear as possible and I tried to write my code very clean.I explained each step with comments so it is understandable.