Modules
Grades
Inbox
Calc

CoC & Instructions
Pre_Processing
Model_Building
Model_Building Assessment
Assignment

Model_Building Assessment

The due date for submitting this assignment has passed.
Due on 2024-04-13, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2024-04-07, 18:54 IST

Click here to view the sklearn library reference

Click here to view the Colab File

Click here to view the Questions.MD file

Load the dataset.

The last column is the target column.
Last 30% rows of the dataset constitute test set and remaining rows form the training set.
Do not shuffle the dataset while splitting
You must have to use only training set to train all the estimator in questions below.
First row of the file has column names/ids, and it has no index column.

1) Which dataset are you using for this exam? Write the last two letters of the dataset file name. ***0 points***

Click here to view the dataset

- V1
- V2
- V3

Yes, the answer is correct.
Score: 0

Accepted Answers:

V3

2) Instructions (Q2-Q3)
Instantiate a perceptron classifier that with following parameters:

- random_state = 1729
- learning rate = 1
- Train for appropriate number of iterations
- Do not shuffle the dataset for each iteration.
- Include the intercept (bias) term.
- Use 10% of the data for validation fraction.
- Do not apply regularization.
- Set warm start to true.

Hint: one iteration of training indicates going over each sample exactly once.

Train the classifier on the training data.

Train the perceptron classifier for 5 iterations. What is value of bias (intercept) after 5th iteration?

-4

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) -6.005,-5.995

4 points

3) In continuation of the previous question, compute precision accurate upto 2 decimal places on training data for positive class (i.e. class value 1), after 5 iterations.

[Hint: Use estimator trained from the previous question]

[Hint: Use estimator trained from the previous question]

0

Yes, the answer is correct.

Score: 5

Accepted Answers:

(Type: Range) 0.0,0.005

5 points

4) Train (on training data only) logistic regression using SGDClassifier. Use the following parameters:

Choose appropriate loss value to obtain logistic regression
penalty='l2',
eta0=0.001,
alpha=0,
learning_rate='constant'
random_state=1729.
warm_start = True

Train the classifier for 5 iterations and note the value of the loss in each iteration. What will be the loss value after second iteration? Answer upto three decimal places.

Note: Set the remaining parameters, if any, accordingly to be able to get the loss value after second iteration. Also note that the classifier has to be trained for 5 iterations.

0.474

No, the answer is incorrect.

Score: 0

Accepted Answers:

(Type: Range) 0.206,0.216

5 points

5) Use GridSearchCV with SGDClassifier. Following are the classifier's parameters: ***5 points***

loss = 'log_loss'
learning_rate = 'constant'
random_state = 1729

Following are parameters to examine:

alpha = [0.0001, 0.0005, 0.001, 0.005]
eta0 = [0.01, 0.05, 0.1, 0.5]

What are the best values of alpha and eta0 respectively?

- (●) 0.0001, 0.01
- ( ) 0.001, 0.01
- ( ) 0.0001, 0.5
- ( ) 0.0005, 0.01
- ( ) 0.005, 0.5
- ( ) 0.05, 0.1

No, the answer is incorrect.

Score: 0

Accepted Answers:

0.05, 0.1

6) Create a new estimator using SGDClassifier that uses the best parameters obtained in Gridsearch earlier ((learning rate to be constant, random_state to be '1729' and use appropriate loss for logistic regression)) and set the weight of class 0 to be 0.1 and the weight of class 1 to be 2. How many samples of class 1 from the test set are correctly predicted by this estimator? ***5 points***

- ( ) 47
- ( ) 46

- 55
- 40
- 41
- None of these

No, the answer is incorrect.
Score: 0

Accepted Answers:

41

7) Fit an SVM classifier with following parameters: ***5 points***

kernel='rbf'
decision_function_shape='ovr'
random_state=1729
C=1

Train the model on training data, and print the confusion matrix on test data.

- [[1142 0] [58 0]]
- [[1141 0] [59 0]]
- [[1139 0] [61 0]]
- [[1150 0] [50 0]]
- [[1000 0] [200 0]]

Yes, the answer is correct.
Score: 5

Accepted Answers:

[[1139 0] [61 0]]

8) **Instructions for Q8-10** ***5 points***

Train a Decision Tree Classifier with the following properties:

criterion = 'entropy',
splitter = 'random',
min_samples_split = 4,
min_impurity_decrease = 0.0001,
random_state = 1729

What is the resultant depth of the tree?

- 18
- 20
- 21
- 24
- 26

No, the answer is incorrect.
Score: 0

Accepted Answers:

18

9) How many nodes are there in the tree? ***5 points***

- 495
- 515

313

491

589

571

No, the answer is incorrect.
Score: 0

Accepted Answers:

571

10) What is the value of entropy at the left child of root?

0.018

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 0.0245,0.0255

5 points

11) Out of DecisionTreeClassifier, KNeighborsClassifier and LogisticRegression, which one performs the best when used as base estimator in BaggingClassifier on the test data in terms of accuracy score when 20 base estimators are used ?
(Use random state 1729 for BaggingClassifier, DecisionTreeClassifier and LogisticRegression) ***8 points***

The metric for best performance will be the lowest 'absolute' difference in the train and test score.

DecisionTreeClassifie

KNeighborsClassifier

LogisticRegression

Yes, the answer is correct.
Score: 8

Accepted Answers:

LogisticRegression

12) When the above three individual classifiers (with same settings) are used in VotingClassifier, how much absolute difference do we obtain in train and test scores? Enter your answer correct upto 4 decimal places.

0.0135

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 0.0077,0.0087

8 points