

Homework 2 Question 1

50/50 Points

21/10/2023

Attempt 1

Review Feedback
21/10/2023Attempt 1 Score:
50/50

Add Comment

Anonymous Grading: no

Unlimited Attempts Allowed

Details

The goal of this assignment is for you to explore different classification algorithms. This assignment is designed to give you practical programming experience with the data preprocessing and evaluation concepts that were discussed in class. Please carefully read all the instructions below. Do not hesitate to use Slack and Q&A community to ask questions.

You can complete this assignment individually or in a group (up to 4 members). To select your group (even if you work individually) go to *People* section in Canvas and select *Groups*. There, you can join the desired group. Please communicate with your colleagues before forming the groups. Your assignment should be submitted by uploading your code (in the form of a **Jupyter Notebook (.ipynb) AND pdf copy of the files** – so we can make comments directly on the file) to Canvas.

Be sure to run the file before committing so that we can directly see your results. Please mention all the resources that were used to solve the problem (e.g., websites, books, research papers, other people, etc.). To complete the assignment, you can use any Python (or R) package that you want, but we recommend using Scikit-Learn.

Question

In this problem, you will apply different classification methods. You will use a Rock dataset where you will use 19 different rock features to predict the rock category. The data you need are included in these two files: 1) **aggregateRockData.xlsx**

(<https://iu.instructure.com/courses/2165858/files/162923781?wrap=1>)_ ↓

(https://iu.instructure.com/courses/2165858/files/162923781/download?download_frd=1) you will only use 2nd column that contains the rock category number (1 = Igneous, 2 = Metamorphic, 3 = Sedimentary) - that will be the label. 2) **norm540.txt**

(<https://iu.instructure.com/courses/2165858/files/162923788?wrap=1>)_ ↓

(https://iu.instructure.com/courses/2165858/files/162923788/download?download_frd=1) you will only use columns 4 to 22 - those will be the attributes (features). See this website for a detailed description of the dataset: <https://osf.io/cvwu9/wiki/Data%20File%20Descriptions/>. ↗



(<https://iu.instructure.com/courses/2165858/modules/items/30838974>)

Attempt

(https://iu.instructure.com/

1. Display the statistical values for each of the attributes, along with visualizations (e.g., histogram) of the distributions for each attribute. Are there any attributes that might require special treatment? If so, what special treatment might they require? **[2 points]**
2. Analyze and discuss the relationships between the data attributes, and between the data attributes and label. This involves computing the Pearson Correlation Coefficient (PCC) and generating scatter plots. **[3 points]**
3. Select 20% of the data for testing and 20% for validation and use the remaining 60% of the data for training. Describe how you did that and verify that your test and validation portions of the data are representative of the entire dataset. **[5 points]**
4. Train different classifiers and tweak the hyperparameters to improve performance (you can use the grid search if you want or manually try different values). Report training, validation and testing performance (classification accuracy, precision, recall and F1 score) and discuss the impact of the hyperparameters (use markdown cells in Jupyter Notebook to clearly indicate each solution):
 - A. Multinomial Logistic Regression (softmax regression); hyperparameters to explore: C, solver, max number of iterations. **[10 points]**
 - B. Support vector machines (make sure to try using kernels); hyperparameters to explore: C, kernel, degree of polynomial kernel, gamma. **[10 points]**
 - C. Random Forest classifier (also analyze feature importance); hyperparameters to explore: the number of trees, max depth, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node. **[10 points]**
5. Combine your classifiers into an ensemble and try to outperform each individual classifier on the validation set (try to get above 80% accuracy). Once you have found a good one, try it on the test set. Describe and discuss your findings. **[10 points]**

✓ **View Rubric**

Select Grader





Mohit Mathrani (TA)



<https://iu.instructure.com/courses/2165858/modules/items/30838974>

<https://iu.instructure.com/c>

Assignment 2

Criteria	Ratings		Pts
Question -1	2 to >0 pts Full Marks Statistical descriptions and Visualizations :1.5 If any special treatment required :0.5 	0 pts No Marks	2 / 2 pts
Question-2	3 to >0 pts Full Marks Computing the PCC:1.5 Scatter Plots :1.5 	0 pts No Marks	3 / 3 pts
Question-3	5 to >0 pts Full Marks Splitting the data in testing , validation and training sets correctly 2.5 Verification of splitting 2.5 	0 pts No Marks	5 / 5 pts
Question 4 a Multinomial Logistic Regression	10 to >0 pts Full Marks Model is implemented correctly :2 Different hyperparameters (C, solver,max number of iterations) have been tried:3 Training, Validation and Testing Performance have been reported :3 Discussion on the impact of different hyper parameters has been done :2 	0 pts No Marks	10 / 10 pts
Question 4 b view longer description	10 to >0 pts Full Marks Model is implemented	0 pts No Marks	10 / 10 pts


<https://iu.instructure.com/courses/2165858/modules/items/30838974>
<https://iu.instructure.com/courses/2165858/modules/items/30838974>


Assignment 2

Criteria	Ratings	Pts
	<p>tried:3 Training, Validation and Testing Performance have been reported :3 Discussion on the impact of different hyper parameters has been done :2</p> 	
<p>Question 4 c</p> <p>view longer description</p>	<div> <div> 10 to >0 pts Full Marks </div> <div> 0 pts No Marks </div> </div> <p>Model is implemented correctly :2 Different hyperparameters(no. of trees, max depth ,the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node) have been tried:3 Training, Validation and Testing Performance have been reported :3 Discussion on the impact of different hyper parameters has been done :2</p> 	10 / 10 pts
<p>Question 5</p> <p>view longer description</p>	<div> <div> 10 to >0 pts Full Marks </div> <div> 0 pts No Marks </div> </div> <p>Ensemble classifier has been implemented via all the models with the best hyperparameters :4 Accuracy of the ensemble is greater than all the individual classifiers :2 Test set Accuracy :1 Discussion on Findings</p> 	10 / 10 pts
Total Points: 50		

File Name

Size


<https://iu.instructure.com/courses/2165858/modules/items/30838974>
<https://iu.instructure.com/courses/2165858/modules/items/30838974>

File Name		Size	
	AML A2 Q1.ipynb	3.65 MB	





<https://iu.instructure.com/courses/2165858/modules/items/30838974>

<https://iu.instructure.com/courses/2165858/modules/items/30838974>