

## Homework 2 Question 1

01/11/2024

50 Points Possible

Attempt 1



In Progress

NEXT UP: Submit Assignment



Add Comment

Unlimited Attempts Allowed

05/10/2024

## 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.

Please do not create groups yourself, but join existing groups (otherwise Canvas will not register your group).

Even if you are staying in the same group, make sure to join the group for Homework 2 - these are independent of the groups for Homework 1).

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 11 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/2249409/files/179561963?wrap=1)

(<https://iu.instructure.com/courses/2249409/files/179561963?wrap=1>).

([https://iu.instructure.com/courses/2249409/files/179561963/download?download\\_frd=1](https://iu.instructure.com/courses/2249409/files/179561963/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) [features\\_presence540.txt](https://iu.instructure.com/courses/2249409/files/179561947?wrap=1)

(<https://iu.instructure.com/courses/2249409/files/179561947?wrap=1>).

([https://iu.instructure.com/courses/2249409/files/179561947/download?download\\_frd=1](https://iu.instructure.com/courses/2249409/files/179561947/download?download_frd=1)), you will only use columns 4 to 14 as the attributes (features) and column 3 (token number) to separate train, validation and test data (see below). See this website for a detailed description of the dataset: <https://osf.io/cvwu9/wiki/Data%20File%20Descriptions/>.

(<https://osf.io/cvwu9/wiki/Data%20File%20Descriptions/>). We will use only the first 480 rows (so ignore rows 481 to 720).

**Answer the questions below directly in your Jupyter Notebook, using Markdown cells.** Be sure to clearly indicate that your comment is an answer to a particular question.

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 labels. This involves computing the Pearson Correlation Coefficient (PCC) and generating scatter plots. [3 points]
3. For training data, use token numbers 1-10, for validation 11 to 13, and for testing 14 to 16 (each of the 30 rock subtypes has 16 token numbers). [2 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 Machine (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. Once you have found a good one, try it on the test set. Describe and discuss your findings. [8 points]

6. Is your method better than a human? Test that by taking human data from [trialData.csv](#)

(<https://iu.instructure.com/courses/2249409/files/179561985?wrap=1>). ↓

([https://iu.instructure.com/courses/2249409/files/179561985/download?download\\_frd=1](https://iu.instructure.com/courses/2249409/files/179561985/download?download_frd=1)) (see [here](#) ↗

(<https://osf.io/cvwu9/wiki/Data%20File%20Descriptions/>) for a description of the file). Compute human accuracy on train and test data (use only rocks with numbers 1 to 480 and note that Block number 1-3 is training, number 4 is test). How does the human accuracy compare to the accuracy of your best model? [2 points] Compute the average human accuracy and standard deviation for each of the 480 rocks (regardless of whether they are train or test rocks). Make a plot with the x-axis showing average human accuracy (values between 0 and 1) and y-axis showing model probability (also values between 0 and 1) for 480 rocks (regardless of whether they were used for train or test). Each rock should be represented with a dot in this plot. Color rocks from three different categories in different colors. [2 points] Compute the correlation coefficient between average human accuracies and model probabilities for each rock category (120 rocks per category) and for all rocks (all 480 rocks). Report the p-value. Is the correlation significant? [1 point]

✓ View Rubric

### Assignment 2 (1)

Criteria	Ratings	Pts
Question -1	<p>2 to &gt;0 pts Full Marks</p> <p>Statistical descriptions and Visualizations :1.5 If any special treatment required :0.5</p>	<p>0 pts No Marks</p> <p>/ 2 pts</p>
Question-2	<p>3 to &gt;0 pts Full Marks</p> <p>Computing the PCC:1.5 Scatter Plots :1.5</p>	<p>0 pts No Marks</p> <p>/ 3 pts</p>
Question-3	<p>2 to &gt;0 pts Full Marks</p> <p>For training data, use token numbers 1-10 :0.66 pts, for validation 11 to 13 :0.66 pts, and for testing 14 to 16 (each of the 30 rock subtypes has 16 token numbers) :0.66 pts</p>	<p>0 pts No Marks</p> <p>/ 2 pts</p>
Question 4 a Multinomial Logistic Regression	<p>10 to &gt;0 pts Full Marks</p> <p>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</p>	<p>0 pts No Marks</p> <p>/ 10 pts</p>
Question 4 b <a href="#">view longer description</a>	<p>10 to &gt;0 pts Full Marks</p> <p>Model is implemented correctly :2 Different hyperparameters (C, Kernel, Gamma, degree) 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>	<p>0 pts No Marks</p> <p>/ 10 pts</p>
Question 4 c <a href="#">view longer description</a>	<p>10 to &gt;0 pts Full Marks</p> <p>Model is implemented correctly :2 Different hyperparameters(no. of trees, max depth ,the minimum number of samples required to split an internal node,</p>	<p>0 pts No Marks</p> <p>/ 10 pts</p>

Assignment 2 (1)

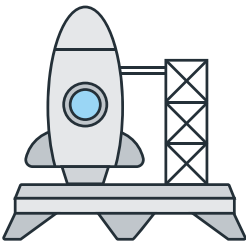
Criteria	Ratings		Pts
	the minimum number of samples required to be at a leaf node) have been tried:3 Training, Validation and Testing Performance have been reported :3 Discussion on the impact of different hyper parameters has been done :2		
Question 5 <a href="#">view longer description</a>	8 to >0 pts Full Marks 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 :1	0 pts No Marks	/ 8 pts
Question 6 part 1 <a href="#">view longer description</a>	2 to >0 pts Full Marks Compute human accuracy on train and test data. :1 Compare to the accuracy of your best model. :1	0 pts No Marks	/ 2 pts
Question 6 part 2 <a href="#">view longer description</a>	2 to >0 pts Full Marks compute the average human accuracy and the standard deviation. :1 Make a plot with the x-axis showing rock numbers and the y-axis showing average human accuracy and standard deviation for each of 480 rocks. :1	0 pts No Marks	/ 2 pts
Question 6 part 3 <a href="#">view longer description</a>	1 to >0 pts Full Marks add the accuracy of your model for each of those rocks. :0.5 Discuss if your model making similar errors as humans? :0.5	0 pts No Marks	/ 1 pts
			Total Points: 0

Keep in mind, this submission will count for everyone in your Project Groups group.

Choose a submission type


Upload


More



Choose a file to upload

or

 Webcam Photo

 Canvas Files

<

(<https://iu.instructure.com/courses/2249409/modules/items/33380041>)

>

(<https://iu.instructure.com/courses/2249409/modules/items/333>)