

Homework 1 Question 1

50/50 Points

29/09/2023

Attempt 2

Review Feedback
29/09/2023Attempt 2 Score:
50/50

Add Comment

Anonymous Grading: no

Unlimited Attempts Allowed

01/09/2023

Details

Introduction

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

The goal of this problem is for you to explore how to properly analyze, visualize, split, clean and format data and perform linear regression, polynomial regression and regularization. You will use the happiness data (e.g. happiness data.csv) located [here](https://iu.instructure.com/courses/2165858/files/160010582?wrap=1)

(<https://iu.instructure.com/courses/2165858/files/160010582?wrap=1>) .

(https://iu.instructure.com/courses/2165858/files/160010582/download?download_frd=1) . The data consists of the following attributes:



(<https://iu.instructure.com/courses/2165858/modules/items/30397200>)

Attempt

(<https://iu.instructure.com/courses/2165858/modules/items/30397200>)

- Life ladder: information about how happy people are
- Log GDP per capita: market values of goods and services in a country
- Social support: how people feel they are supported by those around them
- Healthy life expectancy: rank of the country based on the happiness score
- Freedom to make life choices: how much freedom contributes to one's feeling of happiness
- Generosity: have you donated money
- Perceptions of corruption: how do people perceive that there is corruption
- Positive affect: do you feel happiness, laughter and enjoyment?
- Negative affect: do you feel worry, anger or sadness?

Your task is to build a model that given attributes/features: Country name, Log GDP per capita, Social support, Freedom to make life choices, Generosity, Perceptions of corruption, Positive affect and Negative affect, Healthy life expectancy predicts Life ladder (note that this means that you can ignore Year).

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.


- Summarize the data. How much data is present? What attributes/features are continuous valued? Which attributes are categorical? **[5 points]**
- Display the statistical values for each of the attributes, along with visualizations (e.g., histogram) of the distributions for each attribute. Explain noticeable traits for key attributes. Are there any attributes that might require special treatment? If so, what special treatment might they require? **[5 points]**
- 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. **[5 points]**
- Select 20% of the data for testing. Describe how you did that and verify that your test portion of the data is representative of the entire dataset. **[5 points]**
- Train a Linear Regression model using the training data with four-fold cross-validation using appropriate evaluation metric. Do this with a closed-form solution (using the Normal Equation or SVD) and with SGD. Perform Ridge, Lasso and Elastic Net regularization – try a few values of penalty term and describe its impact. Explore the impact of other hyperparameters, like batch size and learning rate (no need for grid search). Describe your findings. For SGD, display the training and validation loss as a function of training iteration. **[10 points]**
- Repeat the previous step with polynomial regression. Using validation loss, explore if your model overfits/underfits the data. **[10 points]**
- Make predictions of the labels on the test data, using the trained model with chosen hyperparameters. Summarize performance using the appropriate evaluation metric. Discuss the results. Include thoughts about what further can be explored to increase performance. **[10 points]**

✓
S



(<https://iu.instructure.com/courses/2165858/modules/items/30397200>)

(<https://iu.instructure.com/courses/2165858/modules/items/30397200>)

Md Rysul Kabir (TA) 





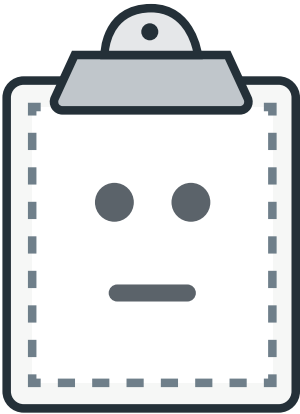
HW1 Rubric			
Criteria	Ratings		Pts
Question-A	5 to >0 pts Summarization How much data is present? - 1 pt, What attributes/features are continuous valued? - 2 pts, Which attributes are categorical? - 2 pts 	0 pts No Marks	5 / 5 pts
Question-B	5 to >0 pts Visualization and summary Statistics Visualization and summary Statistics -3 pts , Special Treatment Needed - 2 pts 	0 pts No Marks	5 / 5 pts
Question-C	5 to >0 pts Correlation PCC table has been computed -1 pt, Scatter Plots -2 pts, Discussion - 2 pts 	0 pts No Marks	5 / 5 pts
Question-D	5 to >0 pts Splitting - Test/Train Correctly splitting into test/train -2 pts, Verification - 3 pts 	0 pts No Marks	5 / 5 pts
Question-E Part 1 view longer description	2 to >0 pts Regression Linear Model using K-Fold with Normal form and SGD (train and val loss) 	0 pts No Marks	2 / 2 pts
Question-F Part 2			3 / 3 pts


<https://iu.instructure.com/courses/2165858/modules/items/30397200>
<https://iu.instructure.com/courses/2165858/modules/items/30397200>

HW1 Rubric			
Criteria	Ratings		Pts
	Regularization with different penalty terms - Ridge - 0.5 pt, Lasso - 0.5 pt. Elastic Net - 0.5 pt, Impact - 0.5 pt 		
Question-E Part 3	5 to >0 pts Hyper parameter tuning and Description Hyper parameter tuning - (Batch size - Learning rate) - 2 pts, Description of models - 3 pts 	0 pts No Marks	5 / 5 pts
Question-F Part 1 view longer description	2 to >0 pts Regression Polynomial Model using K-Fold with Normal form and SGD (train and val loss) 	0 pts No Marks	2 / 2 pts
Question-F Part 2	3 to >0 pts Regularization Regularization with different penalty terms - Ridge - 0.5 pt, Lasso - 0.5 pt. Elastic Net - 0.5 pt, Impact - 0.5 pt 	0 pts No Marks	3 / 3 pts
Question-F Part 3	5 to >0 pts Hyper parameter tuning and Description Hyper parameter tuning - (Batch size - Learning rate) - 2 pts, Description of models - 3 pts 	0 pts No Marks	5 / 5 pts
Question-G Part -1 view longer	7 to >0 pts Prediction on Test Labels	0 pts No Marks	7 / 7 pts
<div> <div>  </div> <div> https://iu.instructure.com/courses/2165858/modules/items/30397200 </div> </div> <div> https://iu.instructure.com/courses/2165858/assignments/15204031?module_item_id=30397201 </div>			

HW1 Rubric		
Criteria	Ratings	Pts
	Evaluation Metric-2 pts ▲	
	Comments The predictions should be under part G.	
Question-G Part -2 view longer description	3 to >0 pts Conclusions Summarize the results - 1 pts, Future work - 2 pts ▲	3 / 3 pts
	0 pts No Marks Comments Well written!	
		Total Points: 50

File Name		Size	
	Homework_1_1-1.ipynb	3.07 MB	✓
	Homework_1_1.pdf	1.52 MB	✓



Preview Unavailable
Homework_1_1-1.ipynb

 [Download](#)

[https://iu.instructure.com/files/162963439/download?
download_frd=1&verifier=phf5pzMaGLDSguY4r7SbMRCyWM0ECmbwdw3L175X](https://iu.instructure.com/files/162963439/download?download_frd=1&verifier=phf5pzMaGLDSguY4r7SbMRCyWM0ECmbwdw3L175X)



<https://iu.instructure.com/courses/2165858/modules/items/30397200>

<https://iu.instructure.com/courses/2165858/modules/items/30397200>