Homework 2

COM SCI X 450.4 - Machine Learning

1 Introduction

This homework will be separated into three modules: Regression, Classification, and Model Implementation. In the first one, you will follow the instructions on the Jupyter notebook that was shared and will compare the performance of Linear Regression (we will use SGDRegressor because it is equivalent but it lets you change the regularization) and KNN for regression. You will then do the same for classification, now comparing Logistic Regression and KNN for classification. Finally, the last part of your homework will be to implement the models yourself and compare the results to the ones from scikit-learn. A large portion of the code has already been structured, you will need to complete the missing pieces.

2 Part I - Regression (40 points)

For this section, you will follow the instructions on the notebook that was shared. The dataset that will be used is the student grade prediction dataset from Kaggle. Our goal is to predict academic performance and learn which factors contribute to good grades. The notebook will guide you through some pre-processing steps. Feel free to change them if you think there are alternative ways to process the data. that could yield improvements.

2.1 Pre-Processing

Follow the instructions on the notebook and answer the questions below.

- Explain, in your own words, what were the pre-processing steps taken.
- Why were the different types of variable pre-processed differently and why is it important?
- What conclusions can you make after looking at the correlation matrix? Which features will likely be the most useful for your model?
- Explain, in your own words, why we need to have a training, a validation and a test set.
- Why do we go through a separate process for test and validation? Why do we have to pre-process the training set first?

2.2 Making predictions and evaluating the model

Train your Linear Regression and your KNN models. Report the metrics for each model and answer the questions below.

Remember: you will do your analysis, select your parameters, and make conclusions by fitting to the training data and making predictions to the validation data. Once you finish your conclusions, you will report the metrics on your test data and compare which model is better.

- Create a baseline. What baseline would you propose for this task?
- Which model performs better? Why do you think so? Is it better than the baseline?
- Which model is faster during training? And during inference time? Hint: You can use the function %timeit here
- What happens when you change the regularization? Remember that alpha=0 means no regularization.
- Were there any major differences between the results of your correlation matrix and the coefficients?
- Pretend you were consulting for this school. The Principal wants to know what are the main factors that contribute to good and poor academic performance. Based on your results, what would you tell the Principal?

3 Part II - Classification (45 points)

We will follow a similar process, but now for Classification. We will use the churn prediction dataset from Kaggle. This is a binary dataset where the objective is to predict if a customer has closed his bank account or if they are still a customer.

3.1 Pre-Processing

It is now your turn to pre-process the dataset. Following the previous logic from the Regression part, you will handle the different kinds of attributes and make the necessary transformations to your training, validation, and test sets. You don't have to implement everything yourself, feel free to explore the documentation from scikit-learn and use any of the existing methods.

- Explain, in your own words, what were the pre-processing steps taken.
- Is there anything interesting that you learned by looking at the data?
- What is the average credit score of a customer that has closed their account?

3.2 Making predictions and evaluating the model

Train your Logistic Regression and your KNN models. Report the metrics for each model and answer the questions below.

Remember: you will do your analysis, select your parameters, and make conclusions by fitting to the training data and making predictions to the validation data. Once you finish your conclusions, you will report the metrics on your test data and compare which model is better.

- Create a baseline. What baseline would you propose for this task?
- Which model performs better? Why do you think so? Is it better than the baseline?
- Which model is faster during training? And during inference time? Hint: You can use the function %timeit here
- What happens when you change the regularization? Remember that alpha=0 means no regularization.
- Based on the coefficients of your model, what can you tell about your features? Which ones are the most informative? Is anything surprising to you?
- How many False Negatives is your model predicting? and False Positives?
- Pretend you were hired by the bank to predict how likely a customer is to churn in the near future. Which metric would be more important for your problem, Precision or Recall? Why?

4 Part III - Implementing the Models (10 Points + 10 bonus points)

You will now modify the three files that were shared with you, *my_logistic_regression.py*, *my_linear_regression.py*, and *my_knn.py*.

4.1 Implementations

- What is the *_prepare_X* function doing and why is it important?
- The weights in the linear/logistic regression files is being initialized by sampling from a normal distribution. What happens if you initialize it differently? (ie. uniform distribution).
- The weights in the linear/logistic regression files is being initialized by sampling from a normal distribution centered on zero. What happens if you initialize it differently? (ie. centered in 1).

- Which models get better results for regression? The ones you implemented or the ones from Scikit-Learn?
- Which models get better results for classification? The ones you implemented or the ones from Scikit-Learn?
- (Bonus) The code implements Logistic Regression for the binary case. Modify the code to work with multiple classes. Hint: Softmax might be a better function than the Sigmoid here. Load the Iris dataset using Scikit-Learn and report your results.

5 How long did it take? (5 Points)

Please add to your written document how many hours you took to finish this assignment. Provide any feedback on the assignment that you might want to share.

6 Deliverable

You will submit your assignment before Sunday, November 1st, before 11:59pm. You should submit a zipped file with:

- The Python source files where you implemented your algorithms
- The Python Notebook with the code you wrote
- The PDF verison of the Python Notebook with the code you wrote
- A separate PDF document with the answers to the questions.

Start early, especially if you are not familiar with Python! If you require help, please contact instructor. Office hours can be scheduled if necessary as long as the instructor is contacted with 48 hour notice.