EE 518: Machine Learning Theory & Algorithms **Spring 2024**

Mini Project 1

Due: May 17, 2024, 11:59PM PT

Student Name: *Instructor Name: John Lipor*

Introduction

Most homework problems encountered in machine learning courses are well-formulated problems on nicely curated datasets. However, real world problems are never this nice. First, the actual task at hand is typically poorly defined (e.g., "here's my data, tell me something from it"). Second, even when a task is clearly defined, datasets contain features of different types (binary, real-valued, categorical, location), missing data, and outliers. One aim of these mini projects is to acquaint you with the task of dealing with real problems.

As stated in class, a problem of interest (at least to the professor) is that of using historical geophysical data to predict geothermal energy favorability, i.e., which locations in the U.S. are amenable to producing electricity by using the earth's heat [1]. For this mini project, you will work with data from the United States Geological Survey (USGS) and the INGENIOUS project [2] to predict a surrogate for geothermal energy favorability. While the geothermal favorability problem is naturally modeled as a classification problem, there are very few known high-favorability locations in the Western U.S., and the resulting imbalance makes it very difficult to train reliable classifiers (see Dr. Lipor's work [3, 4]). For this project, we will instead consider predicting a surrogate for geothermal favorability referred to as the *heat flow residual.* One key indicator of geothermal favorability is when the measured heat flow at a location is much higher than the expected value. Since heat flow is relatively easy to measure, high-quality heat flow estimates have been produced in recent years. As a result, we can compare the actual well measurements to the predicted heat flow to obtain the "label" at these sites corresponding to the difference between the measured and estimated heat flow. We refer to this quantity as the heat flow residual.

You are provided with the heat flow residuals at each point in the training set. While the problem may be viewed as one of pure regression, in practice, we wish to categorize heat flow residuals into four categories. For a residual r_i, we define the categories

1. low: $r_i \leq 25$
2. transition: $r_i \in (25, 50]$
3. high: $r_i \in (50, 200]$
4. very high: $r_i > 200$

The resulting labels 1-4 are *ordinal labels.* In contrast to pure classification, for ordinal data, there is a notion of distance between classes; e.g., predicting "very high" for a true class "low" is worse than predicting "very high" for a true class "high." Due to the ordinal nature of our problem, your final predictor will be evaluated using the following loss

$$\ell(y, \hat{y}) = |y - \hat{y}| \,, \tag{1}$$

where $y, \hat{y} \in \{1, \cdots, 4\}$.

Your task is to train an accurate predictor of heat flow residual that can be used to predict other locations likely to have a high residual, and therefore likely to have high geothermal favorability. You are provided with a training dataset consisting of 3112 examples, each with 28 features, as well as the corresponding heat flow residual. You will also be provided with the features only of a test dataset. As part of your project, you will submit your predictions on the test set, and I will return your score evaluted using (1).[1]

Notes

1. DSS rules apply for all mini projects.
2. Be sure to cite any sources you used, including/especially Kaggle notebooks.
3. For this project, you are free to use any type of predictor or available library (e.g., `xgboost` or algorithms in `sklearn`). However, you may **not** use a neural network.[2]

The Dataset

The training dataset is included in `heatflow_resid_train.csv` and the test dataset is in `heatflow_resid_test.csv`. The target (label) that we are trying to predict is `hfqc_resid`.

The file `featureNames.xlsx` contains a brief description of each feature used. Of the 53 features included, it is very likely that many are not useful for prediction. In our work on this problem, we have utilized the following 16 features:

```
['hf22_tgwt','cond_sur','cond_lcr','cond_man','cond_mcr','cond_uma','geop_mag',
 'geop_grv','geop_dtb','eqi200n5','eqd200n5','geod_shr', 'geod_dil','geod_2nd',
 'faultAl','vent_ing']
```

You may choose to utilize only these features, but your EDA (see below) should explore all available features. Selecting which features to drop based on EDA and preliminary analysis will result in a higher score for the "EDA" and "Approach" categories (see Grading below).

Requirements

You must create a report in either LaTeXor a Jupyter notebook that contains the following sections:

1. Problem description
   - Explain what data you have to work with, what algorithms you will use, what your goal is, and why anyone should care.
2. Exploratory data analysis (EDA)
   - This dataset likely contains many features that will hurt prediction. Further, the data types are heterogeneous, and so some sort of preprocessing will likely be required. The first evaluation of a dataset is known as EDA and is an essential part of a data science pipeline. You must perform an EDA on this dataset and explain your decisions.
   - There are many excellent tutorials on EDA online. Cite whichever ones you found helpful.
   - Be sure to make any significant findings stand out and try to keep this concise.

[1] The top scoring project will receive 500 Lipor points.

[2] There is nothing wrong with neural networks, but there are several other classes on campus devoted to their study. Further, tuning their parameters can be a time sink that will get in the way of your learning for this project.

3. Challenges

   - What were the challenges you encountered when applying machine learning to this dataset?
   - Did these challenges mainly result from the data? From results? Installing libraries to perform preprocessing?

4. Approach

   - Provide a thorough but concise description of your approach, including (but not limited to) your approach to wrangling, preprocessing, and improving the performance of your predictor.
   - This should be much clearer than what is typically found in Kaggle notebooks, but make sure it stays concise. A summary paragraph at the beginning of this section could be helpful.
   - **(NEW)** Since we have now formally studied model selection, you must use either a validation set or cross validation to tune your model parameters. Be sure to describe how you went about this and why.

5. Evaluation and summary

   - **(NEW)** Describe why/where your approach falls short in terms of estimation and approximation error. You may wish to generate a learning curve, as described by Ch. 11 of the text.
   - Consider diving into the data to see if you can determine any trends regarding which points were misclassified.
   - Likewise consider if there were any features that were particularly important or unimportant.
   - Summarize your solution, describing what worked, what didn't, what the main limitations are, and any general conclusions about the dataset.

6. What I learned

   - Describe the main skills/tools that you learned and used for this project and how you learned them.

Grading

The grading breakdown is below. Each section will be graded according to technical correctness, effort, and creativity. Note that clarity of writing is a major component. You should put yourself in the place of writing for a boss or senior in a workplace. If your writing is terrible, they will assume your work is terrible.

Item	**Percentage**
Description	5%
EDA	20%
Challenges	10%
Approach	30%
Evaluation	20%
What I learned	5%
Clarity/conciseness of written communication	10%

Advice

- To start, you may approach this problem as either a pure regression or pure classification problem. This will simplify the process, allowing you to think more carefully about minimizing the ordinal loss later on.
- In our work on this problem, we found the simple approach of performing regression using XGBoost worked well, but only when we trained the predictor using residuals in the "low" and "transition" categories.
- You may make use of linear dimensionality reduction methods such as PCA or dropping features when classifying.
- If helpful, you may make use of nonlinear embedding/dimensionality reduction methods such as t-SNE or UMAP for visualization.
- You may find articles on "data wrangling" and "exploratory data analysis" useful. Two examples are [5, 6]. Please feel free to share these in the `#interesting-reading` channel.
- A great example of really thinking through the results of an algorithm is given in [7].

References

[1] Office of Energy Efficiency & Renewable Energy. (2022) Geothermal basics. [Online]. Available: https://www.energy.gov/eere/geothermal/geothermal-basics#:~:text=Geothermal%20energy%20is%20heat%20energy,depths%20below%20the%20Earth's%20surface.

[2] Great Basin Center for Geothermal Energy. (2022) Ingenious project. [Online]. Available: https://gbcge.org/current-projects/ingenious/

[3] S. Mordensky, J. Lipor, J. DeAngelo, E. Burns, and C. Lindsey, "When less is more: How increasing the complexity of machine learning strategies for geothermal energy assessments may not lead toward better estimates," *Geothermics*, 2023.

[4] S. Mordensky, J. Lipor, E. Burns, and C. Lindsey, "What did they just say? building a rosetta stone for geoscience and machine learning," in *Geothermal Rising Conference (GRC)*, 2022.

[5] (2022) A simple tutorial on exploratory data analysis. [Online]. Available: https://www.kaggle.com/code/spscientist/a-simple-tutorial-on-exploratory-data-analysis

[6] S. Ray. (2016) A comprehensive guide to data exploration. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

[7] (2019) Ml explainability: Deep dive into ml model! [Online]. Available: https://www.kaggle.com/niyamatalmass/ml-explainability-deep-dive-into-the-ml-model