1 Overview

The aim of this project is to build familiarity with the machine learning practice. This is designing to test the ability of designing experiments and using the latest ML tools to carry out these experiments.

2 Background

There are a lot of packages - SCIKIT-LEARN, PYTORCH, TENSORFLOW etc which can help the ML engineer automate several workflows. This might give a false perception that these are all one would need to learn to do ML in practice. However, the main job of the ML engineer is not to use these tools, but rather to build "value" from these.

The skill to create "value" comes from ability to (i) design good hypothesis to test, (ii) testing the hypothesis robustly either in controlled setting or uncontrolled setting, and (iii) finally drawing meaningful conclusions from the outcomes. This lab is aimed at testing these core skills.

3 Project Statement

You can select one of the following problems as a starting point for designing your projects. Note that these projects are open-ended on purpose and you are expected to come up with novel and meaningful experiments within the scope of each problem statement.

A. Examining Bias-Variance Tradeoff in Hypothesis Classes We have discussed a lot in class about hypothesis classes. In fact, the stand we took was - ML is about designing a good hypothesis class!. And a key information which would allow us to choose an hypothesis class is the *Bias-Variance Trade-off*. So, in this problem select various kinds of hypothesis classes - Decision Trees, Neural Networks, Kernel Methods etc and do the following:

1. Generate a synthetic dataset using a function from the hypothesis class $\mathcal{H}_0$.
2. Then manipulate the complexity by designing different hypothesis classes with $\mathcal{H}_{-n} \subseteq \mathcal{H}_{-n+1} \subseteq \cdots \subseteq \mathcal{H}_0 \subseteq \mathcal{H}_1 \subseteq \cdots \mathcal{H}_n$.
3. Examine the bias-variance trade-off curve look for various models and draw your conclusions from it.

B. Decision Trees with Linear Splits: Decision trees can be found efficiently with axis-aligned splits and hence it is used widely in practice. However, it is possible to do linear splits within the decision trees as well.

1. Implement decision trees with linear splits efficiently using the existing tools. Comment on the complexity of your implementation?
2. Compare the boundaries you get here vs the boundaries you obtain with neural networks. Is there any relation between these two classes?
3. Compare this with vanilla decision trees, random forests and boosting. Can you come up with an example in favour of each of these classes over the decision trees with linear splits? and vice-versa?

C. Which hypothesis classes are better tuned for time series prediction? A time series dataset is nothing but a data in the form $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_T$, where each of the $\boldsymbol{x}_i \in \mathbb{R}^d$. This can be converted to

a supervised regression problem by asking - Can you predict the value next time given the previous k values?. That is we want to identify the relationship

$$\boldsymbol{x}_{t-k}, \boldsymbol{x}_{t-k+1}, \cdots, \boldsymbol{x}_{t-1} \rightarrow \boldsymbol{x}_t. \tag{1}$$

So, the aim is to identify the right hypothesis class to do this problem?

1. First, setup the ML problem correctly - (i) What is the right metric to evaluate the model? (ii) How should you do the train/valid/test splits? etc.
2. For each model of hypothesis class, do a hyper-parameter optimization and accordingly select the best model.
3. Comment on the performance of linear models, decision trees and neural networks for time series. Can you identify where the errors in these models are coming from? and propose a solution?

You are required to form a group of 3-4 members. You should submit a report (not exceeding 5 pages [1]) by April 15th 2023 with the following sections

1. Problem Statement
2. Methodology
3. Experimental Results and Validation
4. Conclusion and Future Work

Also, you or one of your group members may be asked to give a presentation of around 10 minutes at the end of the semester.

3.1 Problem Statement and Hypothesis

What constitutes a good problem statement? - A good problem statement should explicitly state what the goal of the project is. And moreover it should also include a measure of the extent to which the problem is solved. Usually a good problem statement is a hypothesis which can be shown to be either true or false.

3.2 Methodology

For problems which include a well framed hypothesis, methodology involves framing an experiment. A good experiment is the one which either proves or disproves your hypothesis. In other words, it should not leave you uncertain with regards to the hypothesis. For theoretically oriented problem statements, a good methodology involves seeing how much of the current known theory is consistent with the problem statement and "twisting" the question to give positive answers.

As you might have gathered this is an open-ended task. As to how much is expected in this project is discussed in evaluation section below.

3.3 Experimental Results and Validation

It is rarely the case that we trust a piece of information coming from a single source. All information should be validated from different perspectives to ensure correctness.

If your problem statement involves an hypothesis, and you have shown it to be TRUE/FALSE you should discuss how well it correlates with current knowledge, and/or identify what aspects made this hypothesis TRUE/FALSE. For theoretically oriented problem statements, experiments are usually on simple datasets whose behaviour is well understood. This allows inference on the theoretical aspects.

[1] Reports exceeding 5 pages will be penalized

3.4 Conclusion and Future Work

You should then summarize your work in simple words and identify precisely what you have proved. More importantly, you should also state what you have not proved and should be considered for future work

4 Evaluation of the projects

As stated above, these projects are expected to be done in groups of 3-4. The evaluation would reflect the standard peer-review procedure to evaluate scientific literature. With an exception that the peers shall be the instructors/TAs. In the sense that,

- If we can identify holes in your arguments easily, they would be penalized. For instance, if you have used the same set for both train and validation, it would be heavily penalized.
- Originality and Creativity would be rewarded. If the questions and results you generate are novel and surprising then it would be highly rewarded.
- You are encouraged to go through the video at https://youtu.be/SPVWSG7-i_E?t=1742 and slides at https://drive.google.com/file/d/15hPTA64h31ShaoybLWeU3moZan7zVbr_/view?usp=sharing (made for the top conference on ML) to get an idea of the peer-review procedure.

5 Final words

This project constitutes of 10% of your grade. While the remaining 90% is designed to evaluate your technical ability, the aim of this project is to test some important soft-skills such as (a) Critical Thinking and (b) Communication of ideas. While the actual results you get from this project may/may-not be useful in your career progression, we can assure you that the soft-skills you acquire would be a great deal beneficial. Hope you enjoy this excercise!