Homework 1: Random Forests

Checkpoint Due Date: 08:00 on September 18, 2023 [50 pts.]

Final Due Date: 18:00 September 29, 2023 [100 pts.]

Problem Statement

Use a random forest to predict a quantity $\mathtt{y}$, given inputs $\mathbf{X}$. You may set the problem up as a regression or classification task.

Data

The data you use is completely up to you, as are the predictors and predictand(s). I would recommend to try and use data from your research, but this is not required. Some things to keep in mind:

- Random forests are a supervised learning approach, so your data needs to be capable of being split such that the predictors $\mathbf{X}$ are used to predict $\mathtt{y}$
- Machine learning is data hungry. Too little data and you've already set yourself up for failure.[1]
- Too much data and your code will take a long time to run. You will likely be running your code dozens of times, so this extra time can add up - but more (good) data is always better.
- When learning a new machine learning approach it is good to start with a data set that you generally understand so that you can identify when the results are out-to-lunch[2]. At the same time, choose a data set where the relationships may be complicated enough that you can fully explore the power of the machine learning approach.

Checkpoint

By the Checkpoint Due Date listed above, submit (via canvas) a document listing 2-3 issues you are facing with this homework as well as a link to your homework github repository (more on this below). This document does not need to be formal (bullets are fine!), but include any helpful figures if desired. You will be sharing these issues with small groups, and possibly the class, so come ready to discuss. Some example issues could include:

- My model is only predicting one class.
- My model never likes predicting the extremes for my regression task.
- My model is overfitting and I don't know how to fix it.
- I am having issues with memory during training.
- My model is no better than random chance.

[1] Alas, I cannot tell you what is "enough" data - that is something you will need to determine for your data sets

[2] Not in touch with the real world, crazy.

Final submitted Write-up

Submit a full write-up on your scientific process and results by the Final Due Date listed above. *You should not submit your code.* The write-up should include text and relevant figures that cover the following:

- scientific motivation and specific problem statement
- description of the data including explicit identification of the predictors and predictands
- description of any data pre-processing performed and why you did it
- training/validation/testing split
- machine learning setup and reasons for hyperparameter choices when relevant
- results (e.g. testing accuracy)
- a detailed discussion of why you don't think you have overfit
- a detailed discussion of why you think the results are better (or worse if that is the case) than a baseline approach of your choice (e.g. random chance, linear regression, climatology, etc)
- concluding thoughts including any insights gained from your efforts
- link to your github repository

Github Repository

As part of this homework you are also expected to use git/github and include a link to your public or private github repository. Your repo must show *at least two commits* by the Checkpoint Due Date and *at least four commits* by the Final Due Date. Your code does not need to be complete, and we will not assess your code, but we do want to see that it is there. A few additional notes:

- If you choose to make your repo private, please be sure to add me (eabarnes1010) and Charlie 2connollyc152 as Collaborators so that we can see your repo.
- Do not upload your data! Github does not deal well with large files / data sets. Just commit and push your code (.ipynb, .py, etc).

Also, some useful links:

- GitHub: `https://github.com`
- Git and GitHub Start-up Guide
- Git cheat sheet: `https://education.github.com/git-cheat-sheet-education.pdf`
- GitHub Desktop: `https://desktop.github.com/`
- ROSSyndicate Best Practices
- ATS GitHub Tutorial run by Justin Hudson Spring 2023