Machine Learning Club, Summer Induction Assignment, 2023 - Q2

1 Predicting Bacterial Property

Considering this is a Tabular Dataset it is best to use a tree based models such as XG Boost, LightGBM, CatBoost etc. The selection of the model will depend on the dataset. It is better to use these models instead of the model due to the following reasons:

1. Tree-based models remain state-of-the-art on medium-sized data (10K samples) even without accounting for their superior speed.
2. Tree-based models are more robust to uninformative features than Neural Networks (NNs).
3. Tree-based models are better at learning irregular patterns in the data (non-smooth target functions) than deep learning models.

References

[1] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," *arXiv preprint arXiv:2207.08815*, 2022. [Online]. Available: `https://arxiv.org/abs/2207.08815`

Even if you do have access to an GPU there is a very high chance tree based models will perform better. However, where the datasets can be represented in some other format such as a graph (the bacterias are the nodes, and if they make contact there is an edge between them) possibly we can try out graph neural networks, but this totally depends on the dataset.

2 Predicting number of people on the beach

Considering this is also a tabular dataset, there is a very high chance that tree-based models will be hard to beat.

However, if we consider the problem as a fit for temporal GNNs, they might outperform tree-based models. In this approach, the nodes would represent the different places at the beach, and the other features such as temperature, precipitation, humidity, wind speed, etc., can be considered as node features. We can create a fully connected graph with the edge weights representing the distances between the places. We would create such a graph for each timestamp and apply temporal GNNs for prediction.

I suggested this model because a similar problem of traffic forecasting has a lot of literature on GNNs, and they have shown to outperform tree-based models.

References

[1] W. Jiang and J. Luo, "Graph Neural Network for Traffic Forecasting: A Survey," *arXiv preprint arXiv:2101.11174*, 2021. [Online]. Available: `https://arxiv.org/abs/2101.11174`

The above paper summarises the development of Temporal GNNs for traffic forecasting. Use only if you have a **GPU**

3 Matrix Multiplication

Just do a simple matrix multiplication. The most efficient and with correct solutions with 0 RMSE. No need to use any models.