Utrecht University
Information and Computing Sciences
AI & Data Science

INFOMLDDE - Programming Assignment

March, 2025

In this course we have introduced the topic of *Stream Learning* (Machine Learning in Data Streams). We have covered several topics including designing Stream Learning algorithms (How to learn in a data-stream), evaluating them (How well do our models perform on a given data-stream), and we have covered several algorithms suitable for the stream setting (Ensembles, Decision Trees, etc.)

In this assignment you are tasked with utilizing this knowledge to evaluate several ML techniques on several stream data-sets. All of this work should be done using Python and the River (`https://riverml.xyz/0.22.0/` framework. The models you will be evaluating are:

1. Linear Model
2. Hoeffding/VF Decision Tree Model
3. An Ensemble Technique (Bagging, Boosting or Stacking)
4. K-Nearest Neighbours

Note that the exact models that you will use have not been specified. This is because it will depend on the data-stream you're evaluating (some are Classification tasks, others are regression tasks). You will have to look at what River provides and make a decision.

The data-streams you will be using for this evaluation are:

1. RandomRBFDrift - A synthetic dataset. Make sure to use a suitable number of features (5 or more) and that you vary the drift speed to simulate different intensities of concept drift.
2. LEDDrift - A synthetic dataset where the goal is to predict the digit on a LED display based on several binary signals. Don't forget to experiment with different noise values and number of drifted features.
3. AirlinePassengers - A regression task where you are asked to predict that number of passengers an airport will receive.
4. Elec2 - A binary classification task where a model must predict whether the price of electricity will go up or down.

1 Task 1: Parameter Tuning

The first task for you to complete is a parameter tuning of your selected models. We often take for granted that the default values provided by frameworks may not be the most optimal.

There are many ways to go about parameter tuning but OFT (One factor at a time) analysis might be the easiest. The idea behind OFT is that you choose one hyper-parameter to vary, and you freeze all of the others. This then allows you understand how that one parameter affects your selected model. You can of course use more complex techniques (such as multi-factor analysis or even metaheuristic techniques for parameter tuning but that is beyond the scope of this course).

When parameter tuning your model, take care of the following:

1. You need to choose your evaluation metrics beforehand. Think about which ones would be appropriate given your chosen model and dataset. Sure accuracy might be fine, but is it appropriate if the dataset is imbalanced? It is also perhaps a good idea to look at multiple metrics to get a broader understanding of how your models behave.

2. Choose reasonable parameters to explore. Some of the models you're working with have a lot of tunable parameters. It would be infeasible (and unnecessary) to explore them all. For example, you might explore different optimizers for a Logistic Regression model, each of those optimizers also might have tunable learning rates. You may also want to look at L1 and L2 regularization, but hopefully you're seeing how this can blow out of proportion quickly.

3. Choose reasonable ranges to explore for each parameter. Just like choosing which parameters to tune, be senisble about the values of these parameters. For the ensemble technique, it makes sense to look at the number of learners in the ensemble. You may think any value $\in [2, 100]$ is reasonable. You don't necessarily need to explore every possible integer value in that range.

4. You may find a range of parameters that achieve similar performance, or even that certain parameter ranges are better on certain metrics, that's ok. The important take-away from this task is that you are able to motivate why you chose these evaluation metrics, parameters to tune and parameter value ranges. You should also take note of what trends emerge as a result of these choices.

2 Task 2: Recovery Analysis

The second task is for you to perform Recovery Analysis on these models that you've chosen. This means identifying two or more Concepts in the data-streams, isolating them, and plotting the α, β, lower bound and actual model's performance curves (See Exercise Class 4 for examples of what this looks like).

Note that this might be quite difficult to do on the real-world data-streams so it would be best to complete this task on the synthetic datasets where you have control over the drift intensity.

It would also be best to conduct these analyses over several data-streams with varying drift intensity. This will give you a greater understanding of how each of your models react to concept drift. You will also then be able to fairly compare each of models across these different data-streams and be able to suggest which models work best under different conditions. If you can do that, you have sufficiently completed this part of the Assignment.

3 Other Notes

- Recall that this assignment **is not for marks!**. However, you will need to have engaged with it to answer certain questions in the final exam.
- The objective of this assignment is for you to gain deeper practical insights into the ML techniques we've covered in class. You have a lot of freedom about how you choose to do that. What is most important is that you can motivate any decisions you make, as well as discuss the findings and limitations of your work.
- If you are unsure about where to start, or have any questions, feel free to email Brandon `b.gower-winter@uu.nl`.