Practical Data Science – 2022/2023

Final Exam – 19/06/2023

Full name: ____

Student Number: ____

This exam is divided in two parts. Part 1 is a set of short, independent questions. Part 2 contains two small cases which will require longer answers. Part 1 is worth 40% of the grade. Part 2 is worth 60% of the grade.

Please, economize your time. I advice you to not get stuck in one question if you are not sure about the answer. Locate the ones that are easy and feel obvious to you and focus on them first.

The space you are given to provide answers to each question has been designed intentionally. If a question only has a small area to answer, that is because a brief, to the point answer is expected. Don’t try to find space somewhere else to extend the length of your answer.

Your exam includes a few pages of blank, draft paper at the end. You can use it for anything you need. I won’t review the contents of it.

Please, use a pen. I will ignore anything written with a pencil.

I wish you good luck.

Part 1

1. What is simulation-based optimization?

2. Why is simulation modeling useful when studying complex systems?

3. Imagine you have to model an uncertain, real world phenomenon like demand for a product, but you don't have historical data that you can rely on. How would you decide how to simulate it?

4. What tools can be used to create a simulation?

5. We discussed how random search is tipically a poor algorithm choice to run simulation based optimization. But still, it is very useful for one thing. What is it?

6. What are the differences between heuristics and metaheuristics?

7. Provide two circumstances where linear or integer programming would be a better tool than simulation.

8. Provide two circumstances where simulation-based optimization would be a better tool than exact methods such as linear or integer programming.

9. Provide two advantages of designing an inventory policy by using simulation-based optimization over a classical Operations Management formula.

10. Simulations with built-in randomness will yield different results between executions even with identical input parameters. For example, a run might show a service level of 93% and then the next one of 98%. How can you deal with this when trying to draw conclusions from your simulation results?

11. Explain the difference between Supervised and Unsupervised Machine Learning.

12. Why is labeled (as in, contains both X and y) data necessary for supervised Machine Learning?

13. What is the difference between classification and regression problems?

14. Why do we split data between train and test sets?

15. What defines a good split when building a decision tree?

16. Is letting a decision tree grow fully a good idea or not? Why? How would you decide what is the optimal size?

17. Why is accuracy typically not enough to measure performance in classification problems?

18. How can you know if a Machine Learning model is overfitting?

19. What would you do if your Machine Learning model is underfitting?

20. Why is the choice of evaluation metric important? What are the advantages of developing a customized evaluation metric over picking a standard one like accuracy, f1-score, RMSE, MAPE, etc.?

Part 2

First case

You are working for an airport. The airport is designing a new terminal. The management needs to decide how many lines to put in the security control area (a line is composed of a metal detection arch + two conveyor belts, one at each side). Management is concerned about the purchase and operation costs of the rather expensive machines that need to be bought and maintained for each line that gets placed. But at the same time, they must make sure that the waiting times for passengers are bearable, avoiding long queues that make passengers unhappy and cause problems in the operations of the airport.

Describe how simulation can help the airport management in this situation.

What is the decision that needs to be made? What are the goals? How can the trade-off be managed?

What data would you ask for to design your simulation? How would you use it to model different parts of the reality in the simulation?

Describe (high level, no code) a heuristic to optimize the decision. You can assume you have access to a simulation code that allows you to run simulations of the security area in action.

Second case

You work for a residential security alarms company. Your company offers a subscription to an alarm service that also includes the alarm hardware and physical security staff nearby customer locations. If an alarm gets triggered, your company will send one of their employees to check if everything is fine at the customer location.

Your company structures their security staff in teams. Each team covers a specific geographical area. The company reviews the sizing of each team every six months to adjust it to the density of customers in the area. There is a trade-off between having a too large team (which means there is always some employee available to check on emergencies, but implies a high cost for the company) and a too small team (which leads to a reduced cost, but a higher chance that there will be no employees available when an emergency happens).

In order to improve their sizing, the company wants to build models to estimate the changes in the customer base for each area during the six months period so their sizings are more accurate. As part of this, one specific subproblem is predicting the departure of existing customers. **The company would like to understand how Machine Learning can be used to assess each individual customer's probability of abandoning the service during the next sixth months**.

- What data would you propose using? Propose at least 5 features that you would build out of the data to prepare the training and testing dataset. You can assume the company has good data on their interactions with their customers (financial transactions, customer service, operational interactions such as triggered alarms and security staff visits, sign-ups and cancellations, etc).

Decide if you would model this as a classification or a regression problem. Decide on one performance metric for it and motivate your choice.

Propose a baseline algorithm for the problem.

After some work, you are pondering whether to build a simple decision tree classifier or go for a random forest classifier. Explain two advantages of each model type.

You finally decide to move forward with the random forest classifier. How would you decide what hyperparameters (tree size, forest size, etc) to use?