HW1: Model evaluation

[PART 1]

This part focuses on the attached basketball shot dataset. Our goal is to predict ShotType using all the other variables. You may assume that the dataset is a representative sample of the data generating process. Optimal performance is not important, so no need for data transformations, feature selection, and you can rely on default parameters, unless otherwise noted.

We want to compare 3 models:

- Baseline classifier that only learns and predicts the relative frequencies of classes,
- Logistic regression,
- A model of your choice (it has to be a model whose performance is known to be very sensitive to the choice of at least one parameter).

For the model of your choice you will have to select the tunable parameters for each training fold in cross-validation. Do this in two different ways (we can count these as two separate models):

- Optimizing training fold performance.
- Nested cross-validation.

Our metrics of choice are log-score and classification accuracy.

Implement a model evaluation and comparison of these four models and two metrics using cross-validation. Report the results and your interpretation. Include and motivate any further methodology choices you had to make.

[PART 2, grades 8, 9]
(will only be graded if you successfully submit and defend Part 1)

After performing the analysis in Part 1 you get two additional requests.

First, we suspect that error depends on Distance. Provide further results that confirm (or disprove) our suspicions. Describe and motivate your methodology.

Second, it turns out that the dataset is not entirely representative of the data generating process – the difference is that the true relative frequencies of Competition types are 0.6 for NBA and 0.1 for the other four types (instead of the approximately equal representation of types in the given dataset). Estimate how the models would perform on data with the true relative frequencies of Competition type. Report the results and your interpretation. Describe and motivate your methodology.

Hint: You don't have to re-do Part 1. Part 2 can be done well enough by analyzing the errors data from Part 1.

[PART 3, grade 10]
(will only be graded if you successfully submit and defend Parts 1 and 2)

This part focuses on the paper: Wainer, J., & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. Expert Systems with Applications, 182, 115222.

- Study the paper for a good understanding of the terminology used in the paper, their methodology, results, and implications for machine learning practice. You don't have to report on this part – you will demonstrate your understanding at the homework defense.
- The authors support their claims using experiments that compare different approaches to selecting tunable parameters. This is again a case of model evaluation and comparison. Write a critical assessment of their methodology (evaluation process, metrics used, statistical tests).

[GENERAL NOTES]

- **Part 1 can be done in Python or R. Part 2 must be done in R** (unless you are not enrolled in the Data Science Track – then you can also use Python for Part 2).
- Submit a pdf report (all parts combined into a single pdf; **no more than 1 page per part!**) and easy-to-reproduce code (each part separately).
- **The evaluation process** (CV, block bootstrap, nested CV, etc.) **and evaluation metrics** (log-score, MSE, etc**.) should be your own code (= don't rely on autoML-like libraries).** For everything else you are encouraged to use existing libraries.
- **Feel free to use any tools, including LLMs and collaboration with others, but keep in mind that our goal is to understand what we are doing and not merely to do. Your work will be graded based on your understanding of your code, report, and the subject matter in general.**