Homework 2: Model Selection

Update: 21 February 2021

In this homework, you are asked to apply the model selection techniques in lectures 5 and 6 to select the best model for the labeled dataset from HW#1. You **MUST** create and write down your code and answers into ONE Jupyter notebook.

Your Jupyter notebook should contain the followings:

- Load and prepare the dataset (e.g., feature scaling, categorical features, etc.)
- Train/Valid/Test split or Cross-validation
- Feature selection
- Model selection
- Evaluation (e.g., accuracy, precision, etc.)

Also, you **MUST** justify your answers in the Jupyter notebook:

- Why do you choose these features? Any preprocessing techniques you apply? If so, why?
- Which approach do you choose to evaluate your model: (1) Train/Valid/Test split or (2) cross-validation? Why?
- Why do you choose this model (e.g., decision tree, logistic regression, etc.)? How do you come up with the hyperparameters for the model (e.g., tree depth, etc.)?
- What are the performance metrics that you use to evaluate your model? Why?

This is *individual* homework. We encourage you to discuss with your friends regarding techniques that you have tried. However, you must write code and answer be yourself. You can ask the instructors during class time or by email. **Copy your friends' answers will result in severe punishments.**

Note: DO NOT worry if the performance of your selected model is not as well as your friends. The goal of this homework is to see whether you understand and can apply the model selection properly to the real-world dataset. You can still get a full score even though your model performs not as well as your friends.

Optional: you will get an extra score if you can apply the Optuna (an open-source hyperparameter optimization framework to automate hyperparameter search) in the model selection. You can see an example here.

Submission: Upload your Jupyter notebook to MyCourse
Deadline: 8 March 2021 @ 23:55