Exercise 5

Practical Data Science (PDS)

Agenda

1. Assignment 2 – Overview

2. Training Machine Learning Models

3. Assignment 3 – Predicting Video Sale Games with Deep Learning

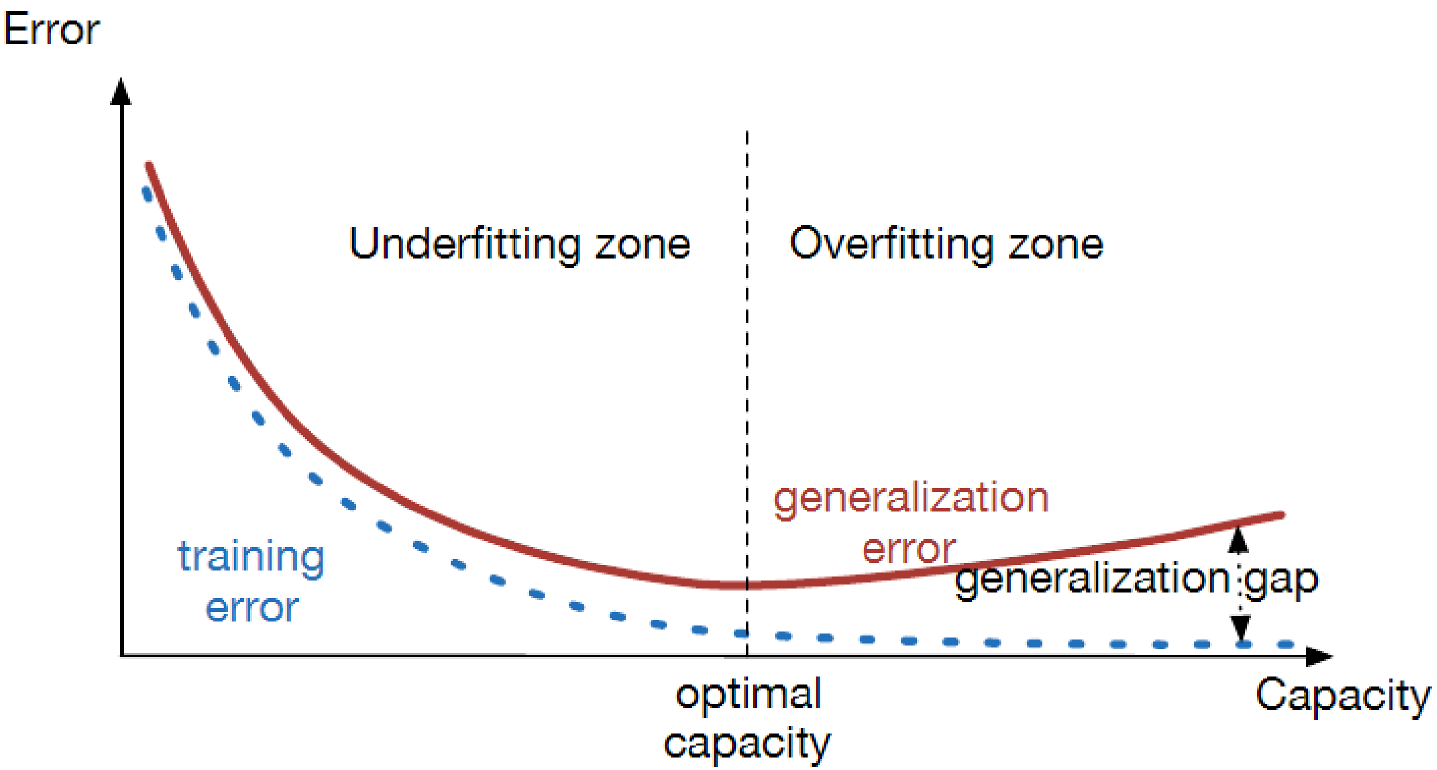
Assignment 2

- Passing Criteria: **all tasks until the model training task** are complete & runnable

- Not penalizing buggy implementations (will do for the next assignment)
  - Use correct encoding algorithms but incorrect data structure
  - Loss is way over 1.0

- Interesting ways to improve the results: encoding strategy (K-NN), new models (Catboost, XGBoost), etc.
  - Some justifications are incorrect (e.g., did not consider loss visualization to prevent overfitting)
  - Good pre-processing techniques (advanced imputation, encoding) can yield better performance

Overfitting vs. Underfitting



https://srdas.github.io/DLBook/ImprovingModelGeneralization.html

Overfitting vs. Underfitting – In a nutshell

- *Underfitting*: train (in-sample) and validation (out-sample) losses can still **decrease** after training

- *Overfitting*: train (in-sample) loss **decreases**, and validation (out-sample) loss **increases**

- Desirable case:
  - Train (in-sample) loss decreases
  - Validation (out-sample) loss decreases
  - They decrease stably without too much fluctuations
  - Small (generalization) gap between train and validation losses

Recipe to start a deep learning project

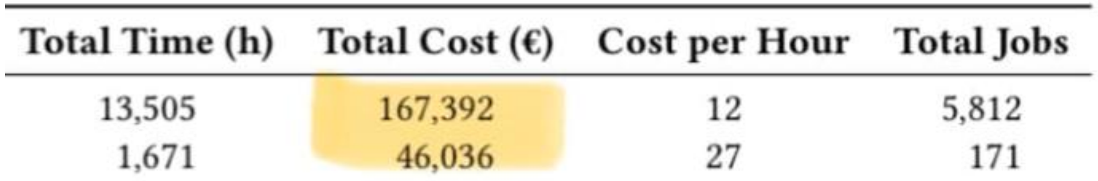
- Sanity check 1: use only **a single** training sample
  - Model should memorize (overfit) the sample with 100% accuracy
  - This prevents any unwanted bugs in the implementation

- Sanity check 2: increase to **a small subset** of training samples
  - Similar reason, making sure model works correctly

- Start increasing samples for training, e.g., 20%, 50%, 80% and 100% of the train set
  - Generalization should occur at some point

- Find a good learning rate

- Why all of these? **GPUs are expensive**!

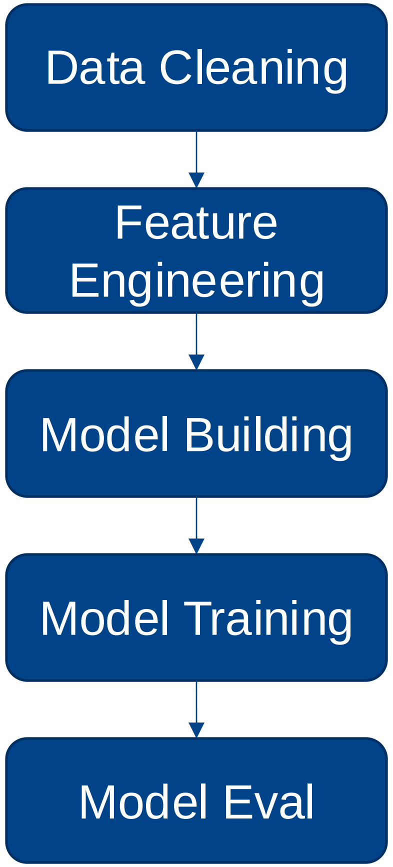


Total Time (h)	Total Cost (€)	Cost per Hour	Total Jobs
13,505	167,392	12	5,812
1,671	46,036	27	171

Assignment 3 - Predicting Video Game Sales with Deep Learning

- **Deadline**: 07.12.2024 at 23:59 pm

Assignment 2



Data Cleaning

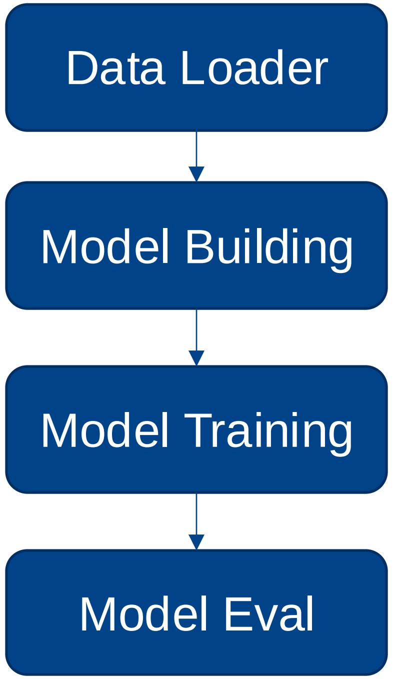
Feature Engineering

Model Building

Model Training

Model Eval

Assignment 3



Data Loader

Model Building

Model Training

Model Eval

Additional Materials

- Practical Deep Learning lectures: https://cvg.cit.tum.de/teaching/ws2024/i2dl
  - Implement DL models with PyTorch
  - More in-depth engineering skills for training Neural Networks

- Improving model generalization: https://srdas.github.io/DLBook/ImprovingModelGeneralization.html

- Analyze different loss curves: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

- Practical Deep Learning book: https://udlbook.github.io/udlbook/
  - Beginner-friendly with many examples
  - Each chapter is accompanied with Jupyter Notebooks