Lecture 3: ML fundamentals

Firas Moosvi (Slides adapted from Varada Kolhatkar)

Announcements

- My weekly office hours will be announced soon!
- Homework 2 (hw2) is released already!
  - You are welcome to broadly discuss it with your classmates but final answers and submissions must be your own.
  - Group submissions are not allowed for this assignment.
- Advice on keeping up with the material
  - Practice!
  - Reminder to run the lecture notes on your laptop and experiment with the code.
  - Start early on homework assignments.
- Last day to drop without a W standing is this Friday: **May 15, 2026**

Dropping lowest homework (Update)

- CPSC 330 has 9 homework assignments that are all an integral part of your learning
- To account for illnesses, other commitments, and to preserve your mental health, there has long been a policy of dropping your lowest HW score.
- After some analysis from the data from previous terms (Learning Analytics!), there is a slight modification to this policy:

> **With the exception of HW5,** we will drop your lowest homework grade - all students must complete HW5.

- This is to encourage all students to complete HW5! It's important!

Recap

- Importance of generalization in supervised machine learning
- Data splitting as a way to approximate generalization error
- Train, test, validation, deployment data
- Overfitting, underfitting, the fundamental tradeoff, and the golden rule.
- Cross-validation

Finish up demo from last class

For this demo, each student should click this link to create a new repo in their accounts, then clone that repo locally to follow along with the demo from today.

Recap

A typical sequence of steps to train supervised machine learning models

- training the model on the train split
- tuning hyperparamters using the validation split
- checking the generalization performance on the test split

Clicker 3.1

Participate using Agora (code: canvas)

Select all of the following statements which are TRUE.

- a. A decision tree model with no depth (the default `max_depth` in `sklearn`) is likely to perform very well on the deployment data.
- b. Data splitting helps us assess how well our model would generalize.
- c. Deployment data is scored only once.
- d. Validation data could be used for hyperparameter optimization.
- e. It's recommended that data be shuffled before splitting it into train and test sets.

Clicker 3.2

Participate using Agora (code: canvas)

Select all of the following statements which are TRUE.

- a. k-fold cross-validation calls fit k times
- b. We use cross-validation to get a more robust estimate of model performance.
- c. If the mean train accuracy is much higher than the mean cross-validation accuracy it's likely to be a case of overfitting.
- d. The fundamental tradeoff of ML states that as training error goes down, validation error goes up.
- e. A decision stump on a complicated classification problem is likely to underfit.

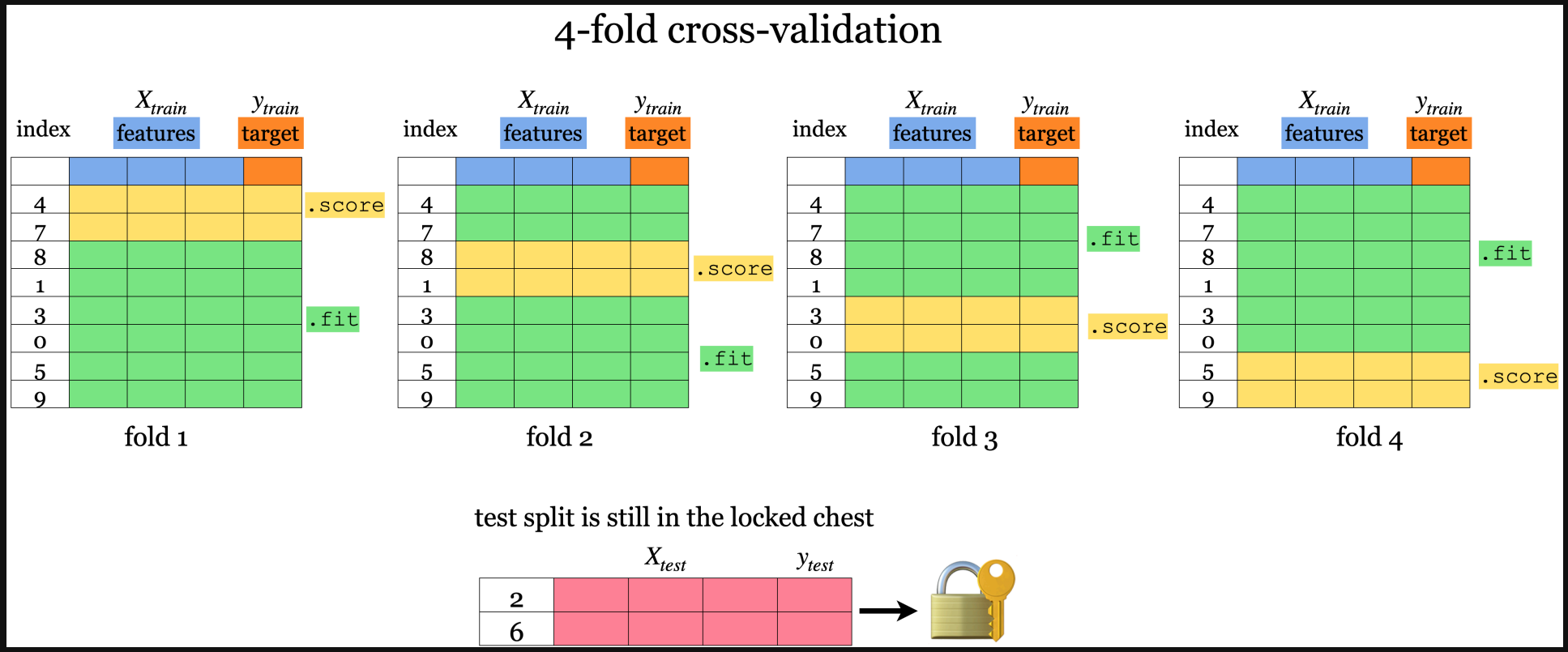
Recap from videos

- Why do we split the data? What are train/valid/test splits?
- What are the benefits of cross-validation?
- What is underfitting and overfitting?
- What's the fundamental trade-off in supervised machine learning?
- What is the golden rule of machine learning?

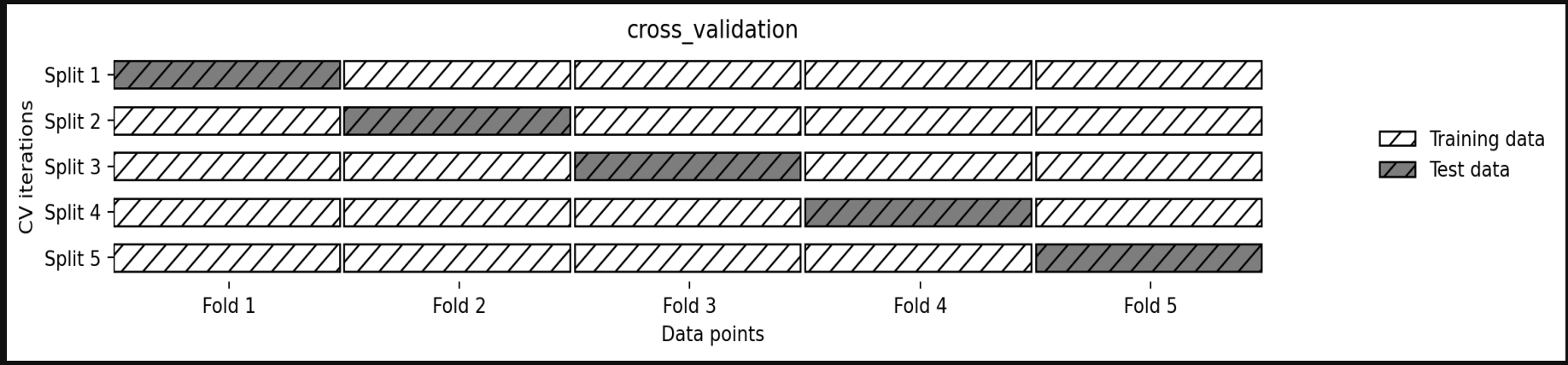
Summary of train, validation, test, and deployment data

	`fit`	`score`	`predict`
Train	✓	✓	✓
Validation		✓	✓
Test		once	once
Deployment			✓

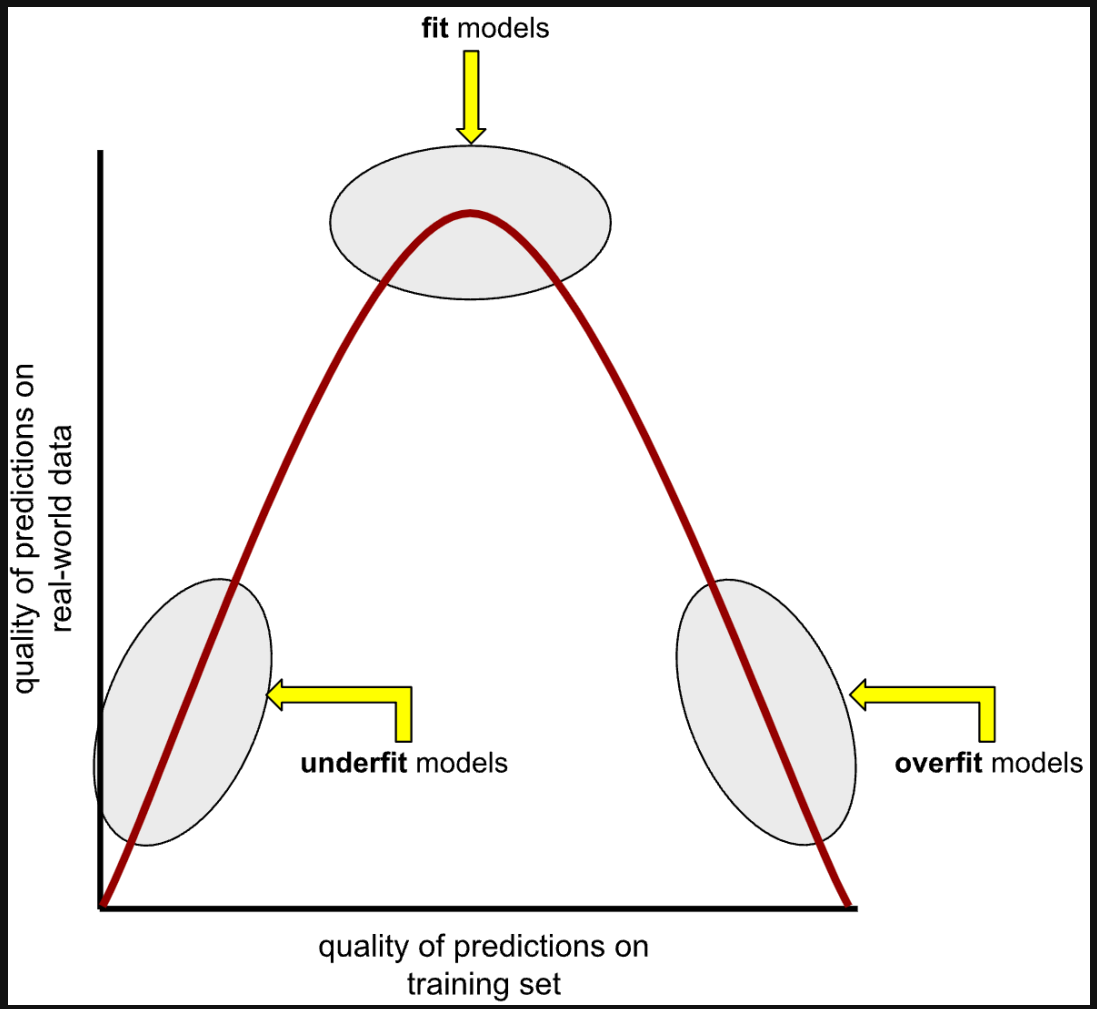
Cross validation



Cross validation



Overfitting and underfitting

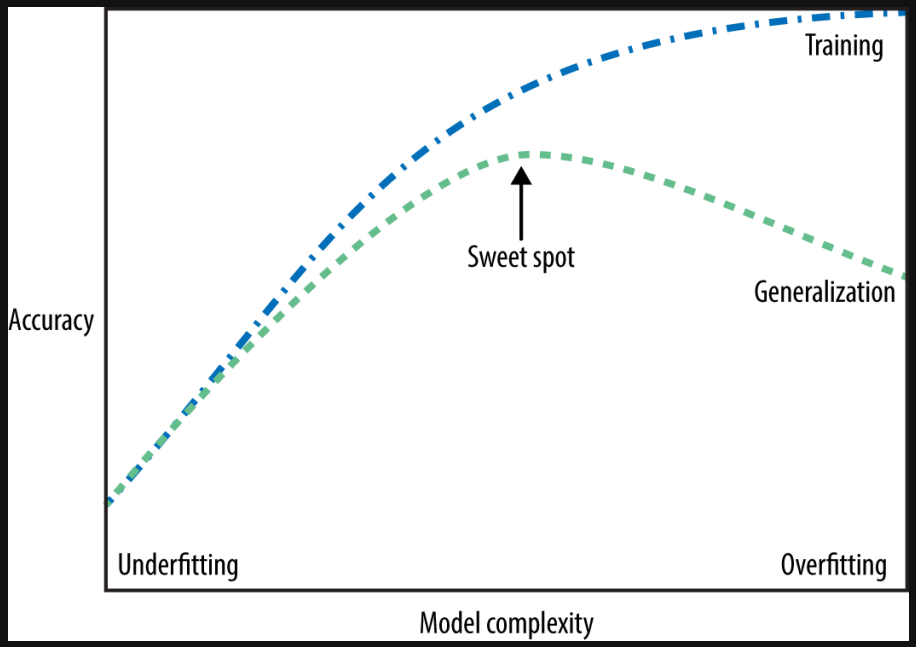


- An **overfit model** matches the training set so closely that it fails to make correct predictions on new unseen data.
- An **underfit model** is too simple and does not even make good predictions on the training data

Source

The fundamental tradeoff

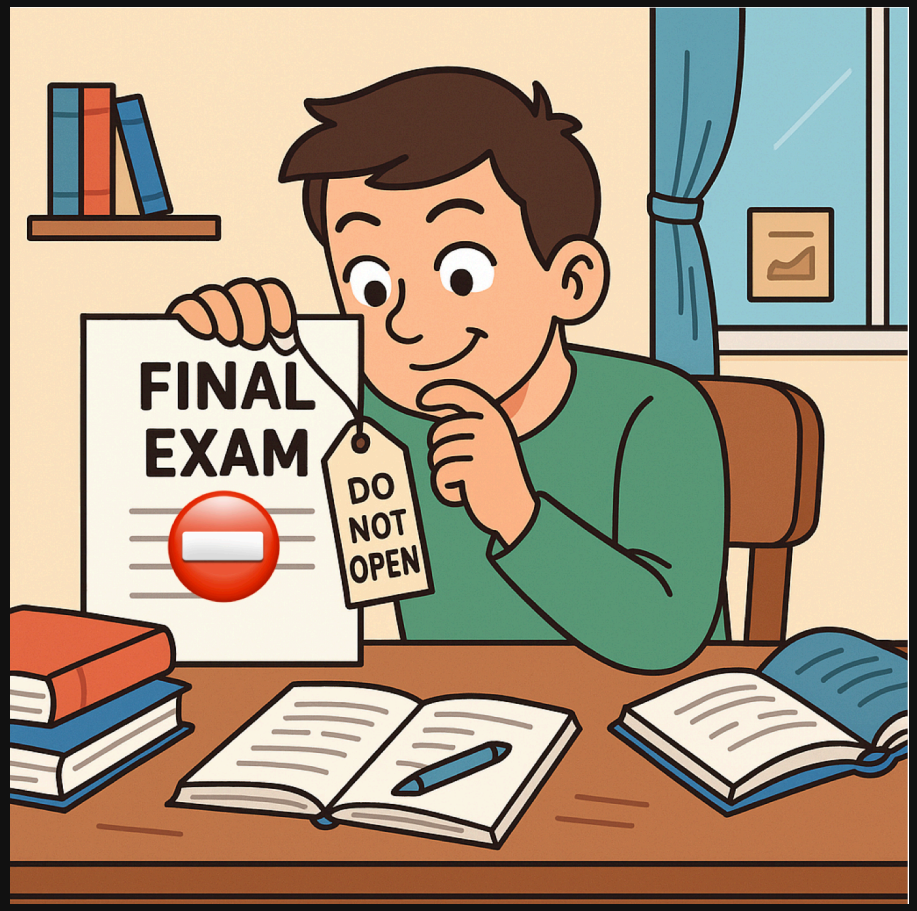
As you increase the model complexity, training score tends to go up and the gap between train and validation scores tends to go up.



- Underfitting: Both accuracies rise
- Sweet spot: Validation accuracy peaks
- Overfitting: Training ↑, Validation ↓
- Tradeoff: Balance complexity to avoid both

The golden rule

- Although our primary concern is the model's performance on the test data, this data should not influence the training process in any way.



Source: Image generated by ChatGPT 5

- **Test data = final exam**
- You can practice all you want with training/validation data
- But **never peek** at the test set before evaluation
- Otherwise, it's like sneaking answers before the exam ⟶ **not a real assessment of your learning**.

Additional Resource on Cross Validation

MLU-EXPLAIN

CROSS VALIDATION

Reduce, Reuse, Resample

Jared Wilber & **Jasper Croome** May 2022

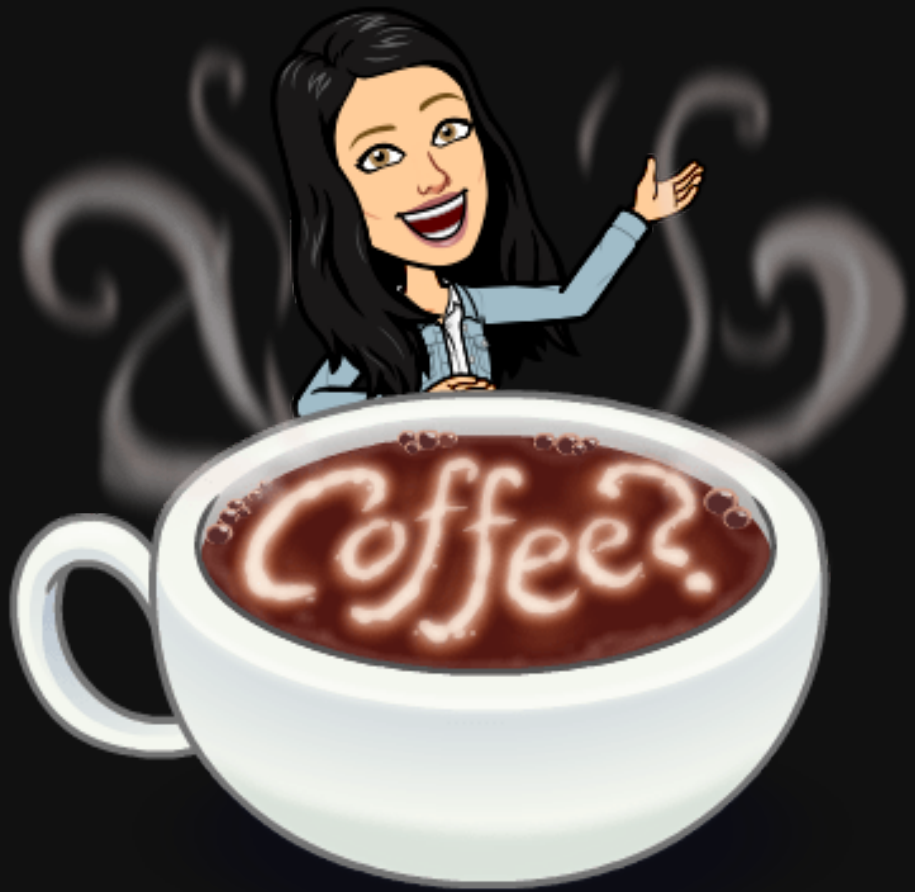
In machine learning we need to estimate the performance of a model before we put it into production. While we could just evaluate our model's performance on the same data that we used to fit its parameters, doing so will give us unreliable assessments of our model's ability to generalize to unseen data. Because obtaining new data may be difficult, we'd like to find a way to assess the generalization capabilities of a model without having to wait for new data. This article discusses one of the most common approaches for this task: **K-Fold Cross-Validation**. We'll

Reference: MLU-Explain - Cross Validation

UBC Computer Science

Break

Let’s take a break!



Group Work: Class Demo & Live Coding

For this demo, each student should click this link to create a new repo in their accounts, then clone that repo locally to follow along with the demo from today.