Lecture 3: ML fundamentals

Firas Moosvi (Slides adapted from Varada Kolhatkar)

Announcements

- Homework 2 (hw2) has been released (Due: Sept 16, 11:59pm)
  - 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!
  - Make sure you run the lecture notes on your laptop and experiment with the code.
  - Start early on homework assignments.
- If you are still on the waitlist, it’s your responsibility to keep up with the material and submit assignments.
- Last day to drop without a W standing: **Sept 16, 2023**

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

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

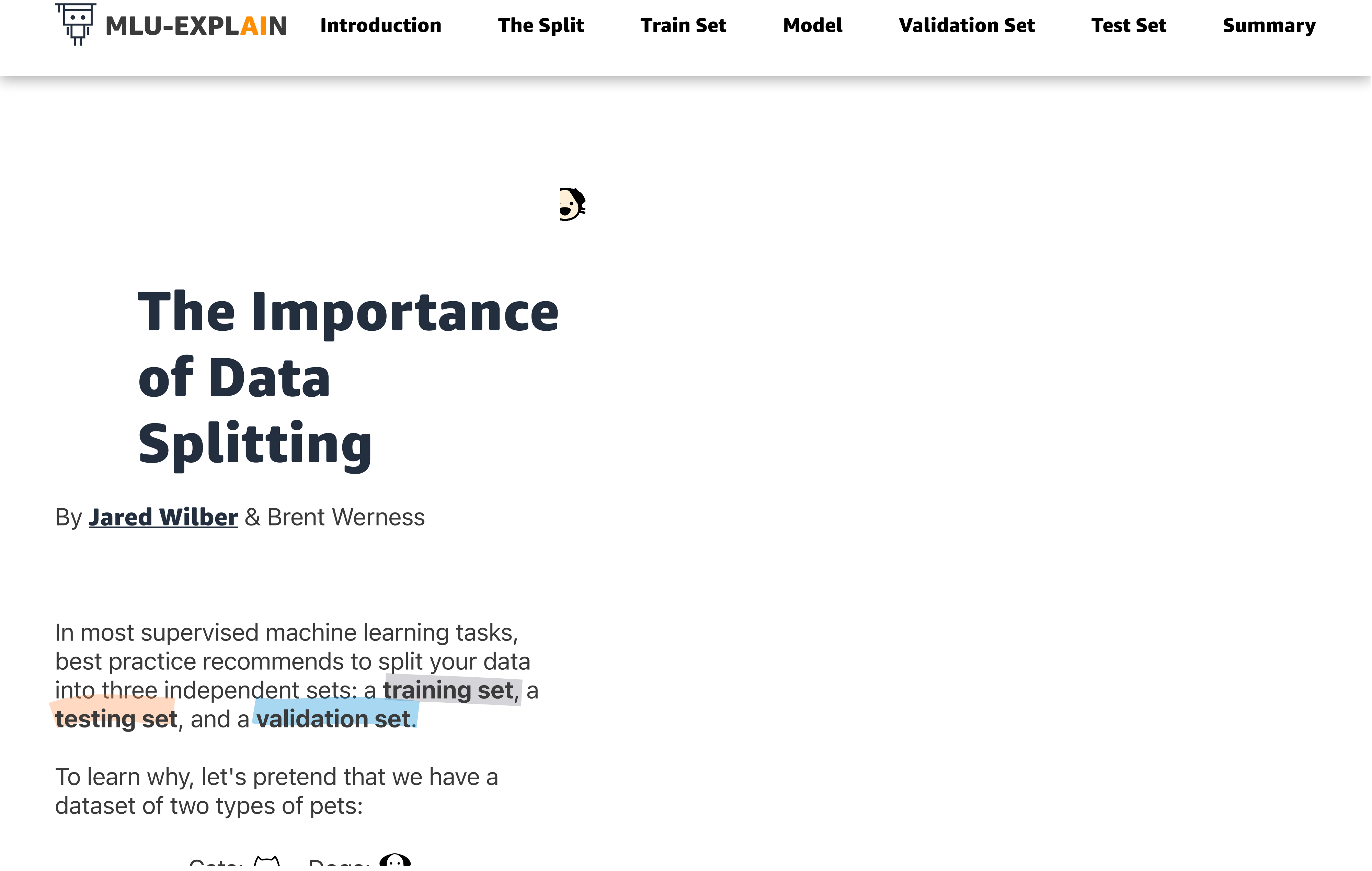
iClicker 3.1

Clicker cloud join link: https://join.iclicker.com/VYFJ

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 only scored 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.

Additional Resource



The Importance of Data Splitting

By **Jared Wilber** & Brent Werness

In most supervised machine learning tasks, best practice recommends to split your data into three independent sets: a **training set**, a **testing set**, and a **validation set**.

To learn why, let's pretend that we have a dataset of two types of pets:

Cats: Dogs:

Reference: MLU-Explain - Data Splitting

iClicker 3.2

Clicker cloud join link: https://join.iclicker.com/VYFJ

Select all of the following statements which are TRUE.

a. k-fold cross-validation calls fit k times

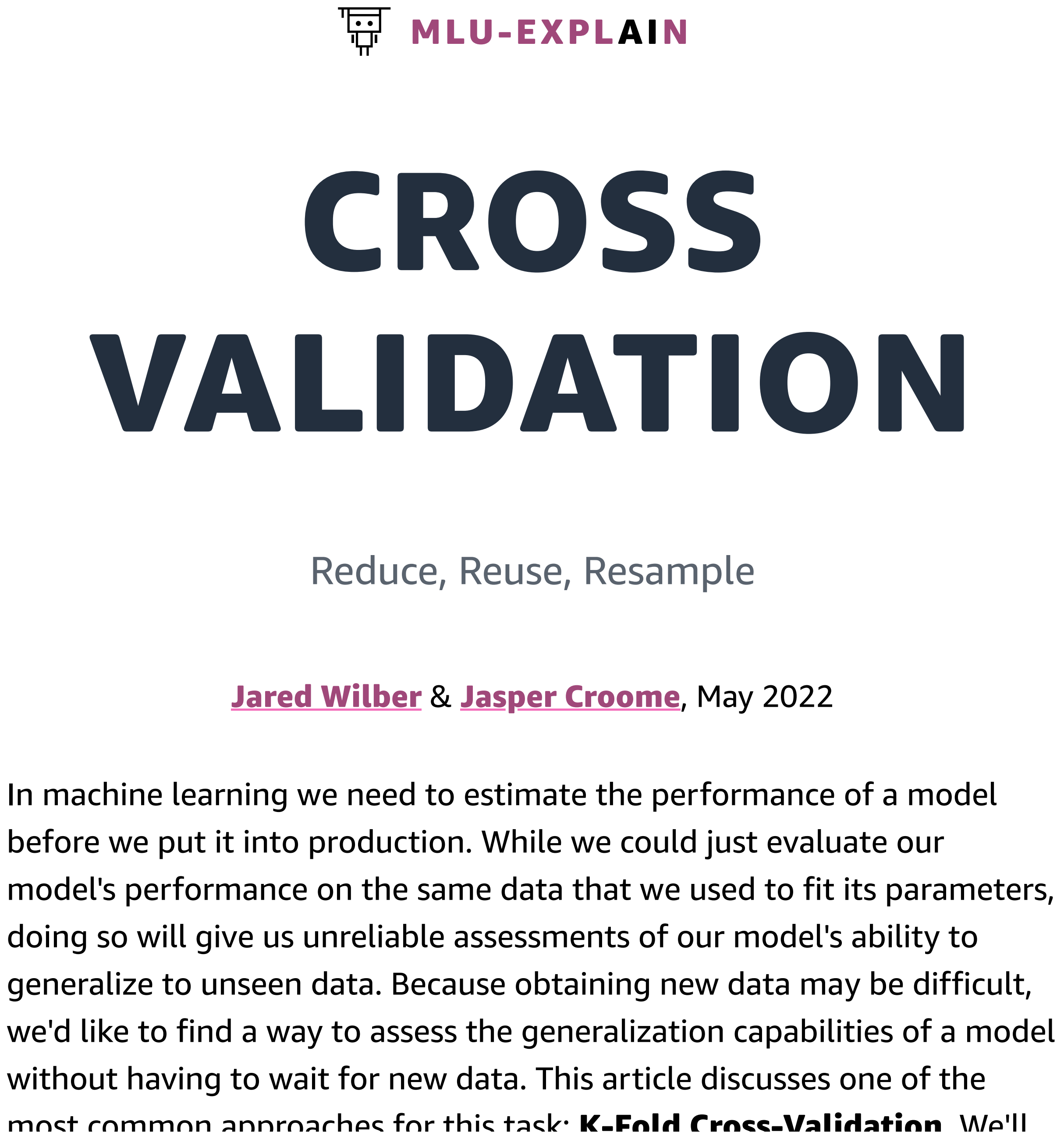
b. We use cross-validation to get a more robust estimate of model performance.

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.

Additional Resource



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

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.

If you really don’t want to create a repo,

- Navigate to the `cpsc330-2024W1` repo
- run `git pull` to pull the latest files in the course repo
- Look for the demo file here: `lectures/102-Firas-lectures/class_demos/`.