Machine Learning for Large-Scale Data Analysis and Decision Making (MATH80629A) Fall 2021

Week #3 - **Summary**

Announcement

- **Hybrid classroom**: Mondays 8:30 am - 11:30 am
  Class room: Manuvie. This classroom is located on the 1st floor of Côte-Sainte-Catherine building.
  Zoom: Zoom link.

- **Hybrid office hour**: Mondays 11:30 am - 1 pm
  Office:4.834
  Zoom: Zoom link.

- **Lab session** on week #5 (September 27)
  Lab room: Laboratoire Lachute

Today

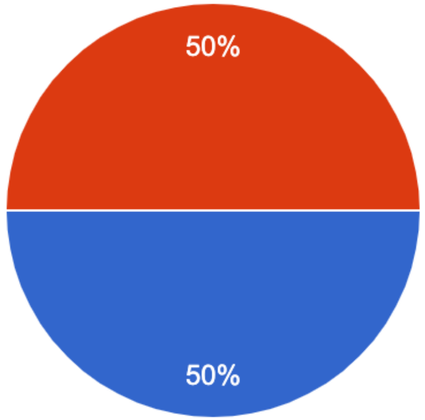
- **First Quiz** on Gradescope!
- **BE PREPARED** for next week! We will have a quiz almost every week at the beginning of the class. You can check the schedule on the website.
- Summary of Machine learning fundamental
- Q&A
- Hands-on session

Quiz 0

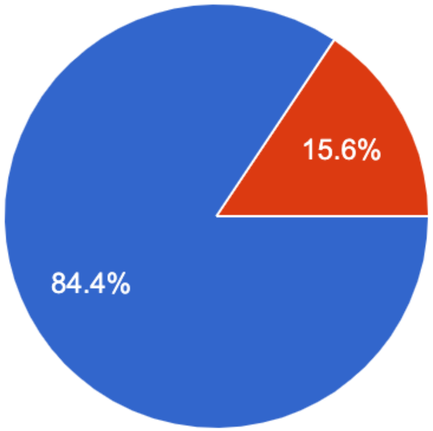
Login to your Gradescope account

Class statistics

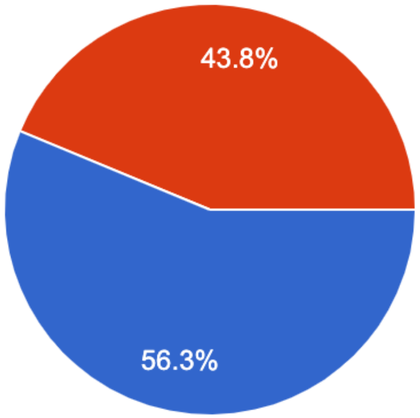
Gender



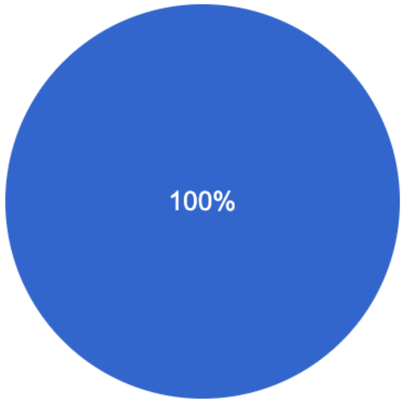
Python



ML



Laptop



Student Introduction Survey form **due tonight**

Machine Learning Problem

The three components of an ML problem:

1. **Task.** What is the problem at hand?

   - Model. How are you parametrizing your solution.

2. **Performance.** How well you are doing?

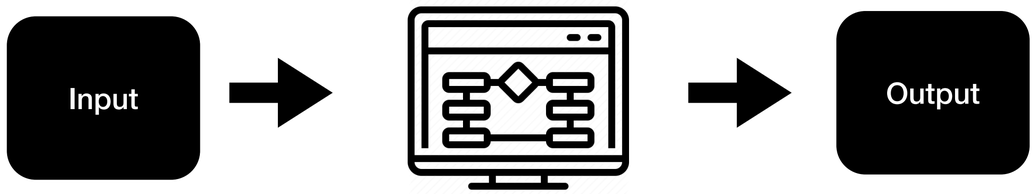
3. **Experience.** What kind of data do you have access to?

Types of Experiences

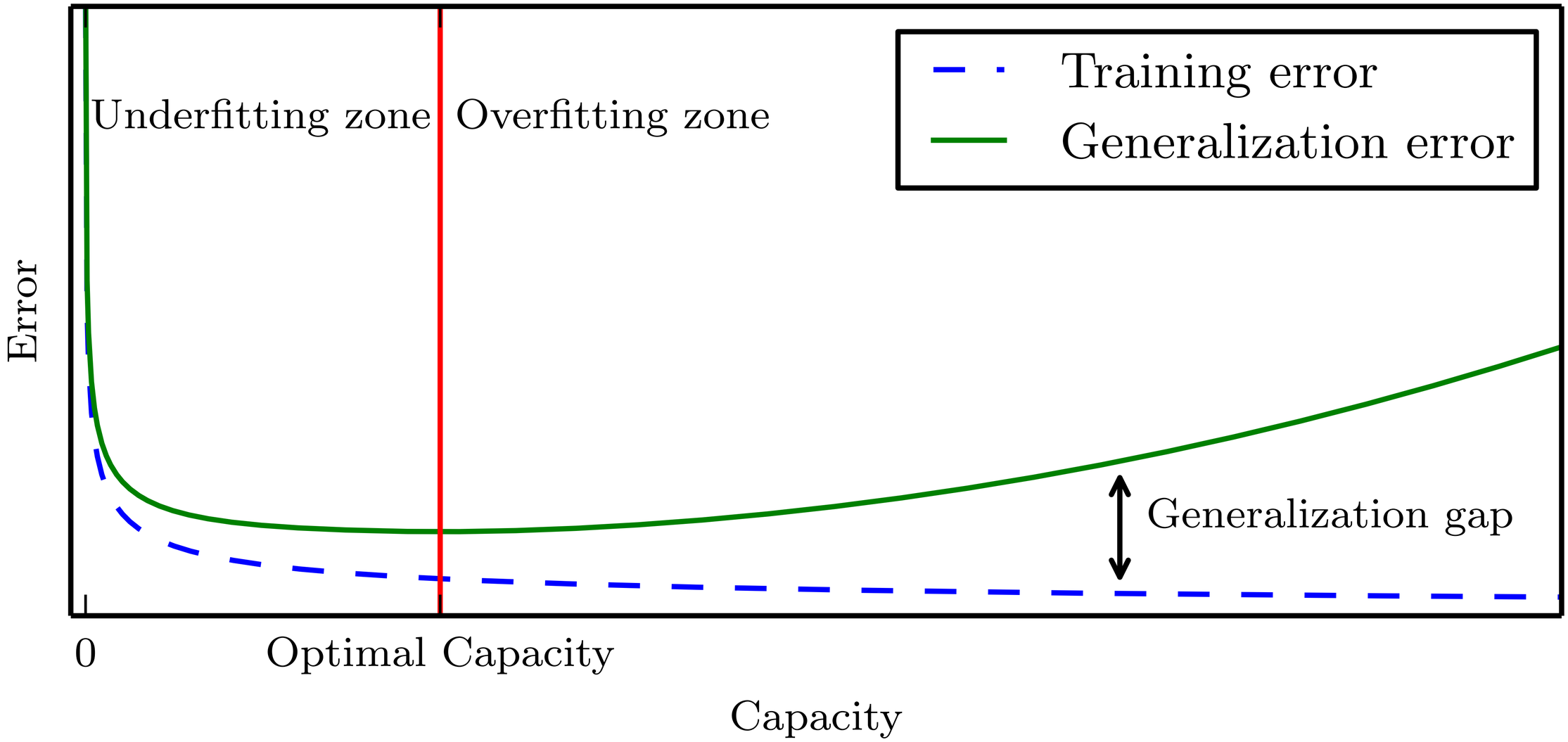
- **Supervised {(x,y)}.** e.g., regression, classification. f: X -> Y
- **Unsupervised {(x)}**. e.g., clustering, dim. reduction, density estimation
- **Reinforcement learning**. Agent takes actions in an environment.

Model Evaluation

- **Given:**
  - A performance measure
  - A train dataset
  - A model



- **Can calculate:**
  - Train error: used to learn (to train).
  - Train error cannot be used to evaluate your model
  - Must use a separate dataset for evaluation

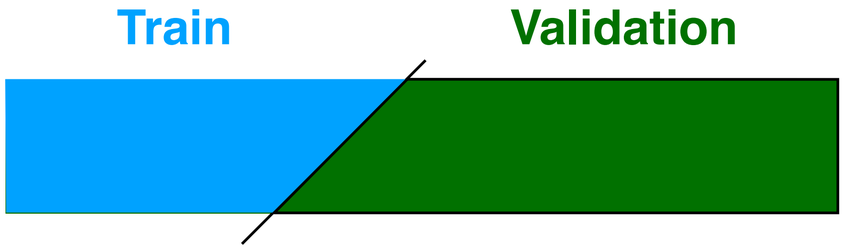

Underfitting zone
Overfitting zone
Training error
Generalization error
Error
Generalization gap
0
Optimal Capacity
Capacity

Regularization

- Can be thought of as way to limit a model's capacity
- $\text{Loss} := \text{MSE}^{\text{train}} + \underbrace{\lambda \mathbf{w}^\top \mathbf{w}}_{||W||_2}$

Validation set

- How do we choose the right model and set its hyper parameters (e.g.)?
  - **Use a validation set**
    - **Split the original data into two:**
      1. Train set
      2. Validation set
        - Proxy to the test set



  - **Train different models/hyper-parameter settings on the train set**
  - **Pick the best according to their performance on the validation set**

Bias / Variance

- The goal is to hit the bull's eye
- Each blue dot represents the "performance" of a fixed model on different data from the same distribution