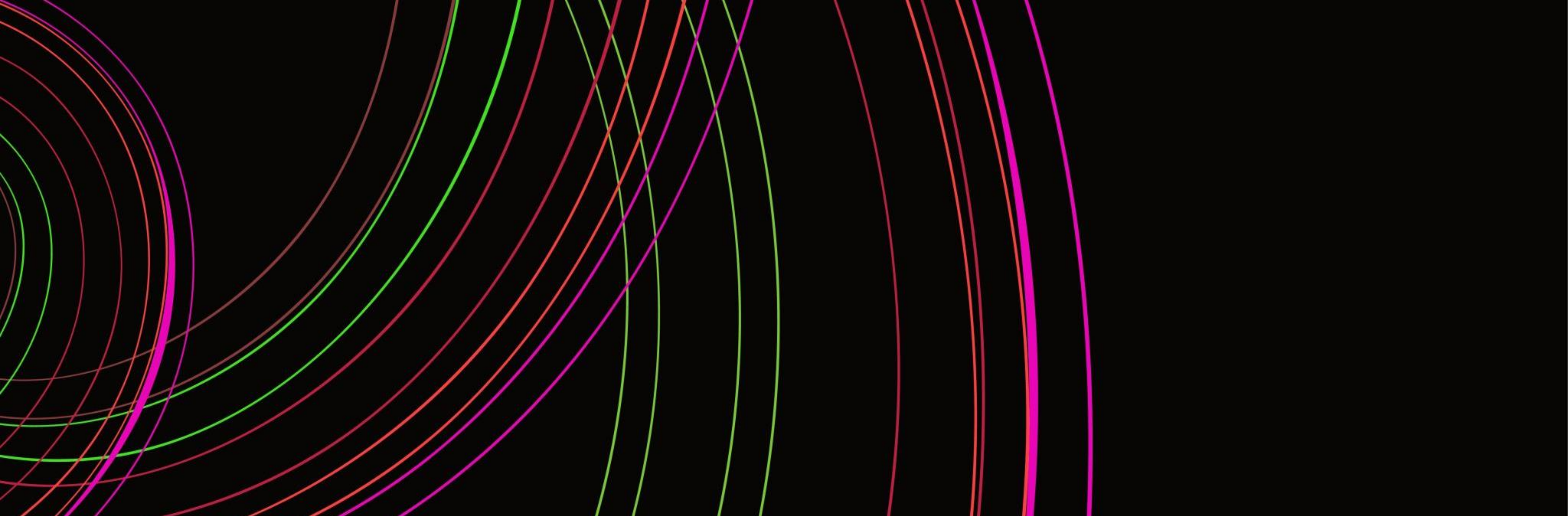


ML Advice and concluding thoughts

Nakul Gopalan

nakulgopalan

Slides inspired from Byron Boots, Rodrigo Borela

Announcements

- Last class
- CIOS survey and bonus
- Feel free to contact me meanwhile
- Office hours with me on an appointment basis. Shoot me an email if you want to chat.
- Final Project due 4th of May 2021 AOE with a 7 min video and a complete report with an ethics statement about your project.

What is Machine Learning?

- “Learning is any process by which a system improves performance from experience.” - Herbert Simon

- Definition by Tom Mitchell (1998):
  - Machine Learning is the study of algorithms that improve their performance P
  - at some task T
  - with experience E.

  A well-defined learning task is given by .

Slide from Byron Boots

Supervised Learning

- Input data X with Labels Y
- Has Training and Testing accuracy
- Built on strong assumptions about independence of data
- Built on strong assumptions about noise present in the data

Train vs Test data

- Do not let your ML algorithm cheat by looking at the test data.
- Learning is generalization to novel data
- Test data is sacred!!!!
- Use validation data to improve model

Different ways to improve your model

- More training data
- Features
    1. Use more features
    2. Use fewer features
    3. Use other features
- Better Training
    1. Run for more iterations
    2. Use a different algorithm
    3. Use a different classifier
    4. Play with regularization

Different ways to improve your model

- More training data
- Features
  1. Use more features
  2. Use fewer features
  3. Use other features
- Better Training
  1. Run for more iterations
  2. Use a different algorithm
  3. Use a different classifier
  4. Play with regularization

Needs an organized approach!

First step: diagnose your model

Some possible problems:

- Overfitting (high variance)
- Underfitting (high bias)
- Your learning does not converge
- Are you measuring the right thing?

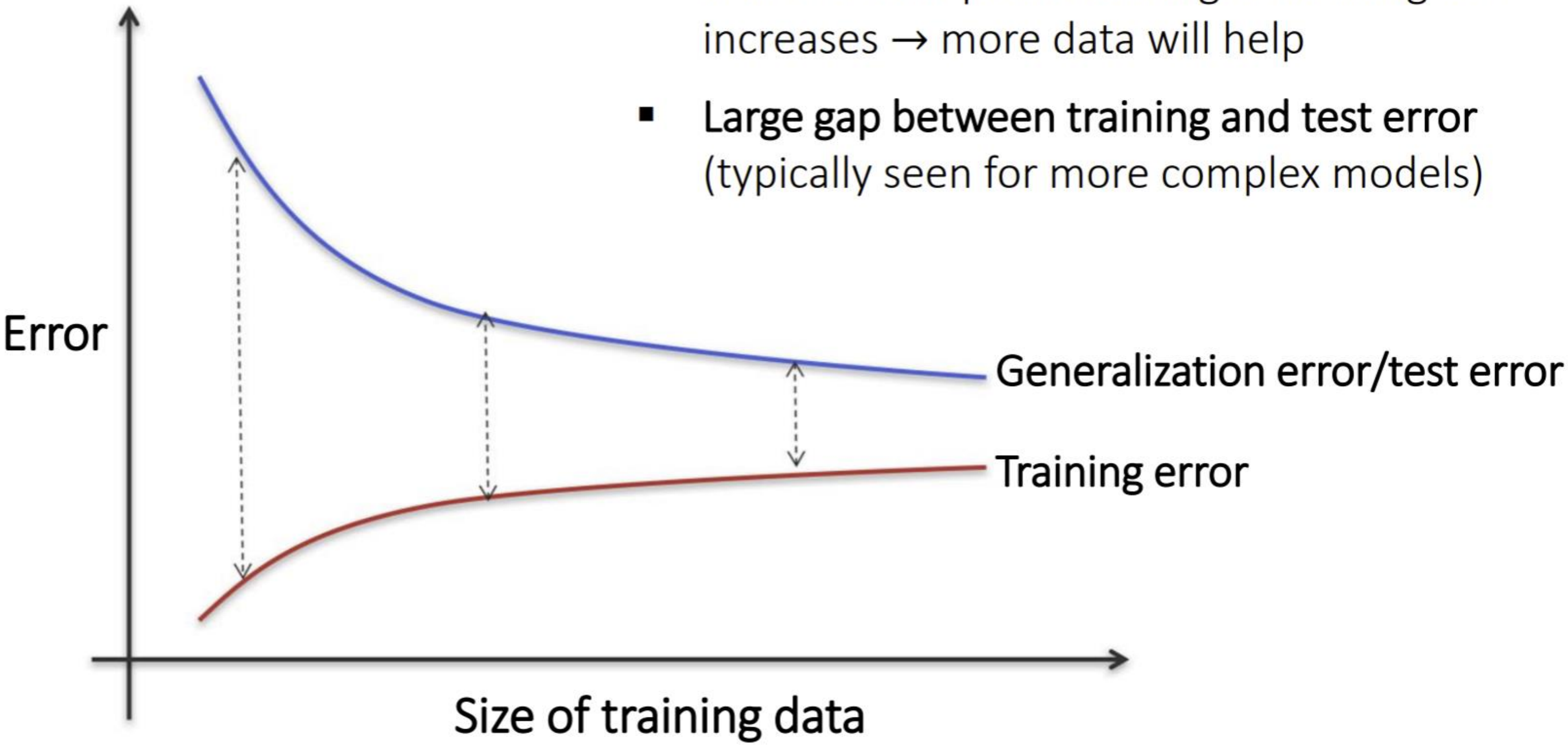
Overfitting vs. underfitting

- Overfitting: the training accuracy is much higher than the test accuracy
    - The model explains the training set very well, but poor generalization

- Underfitting: both accuracies are unacceptably low
    - The model can not represent the concept well enough

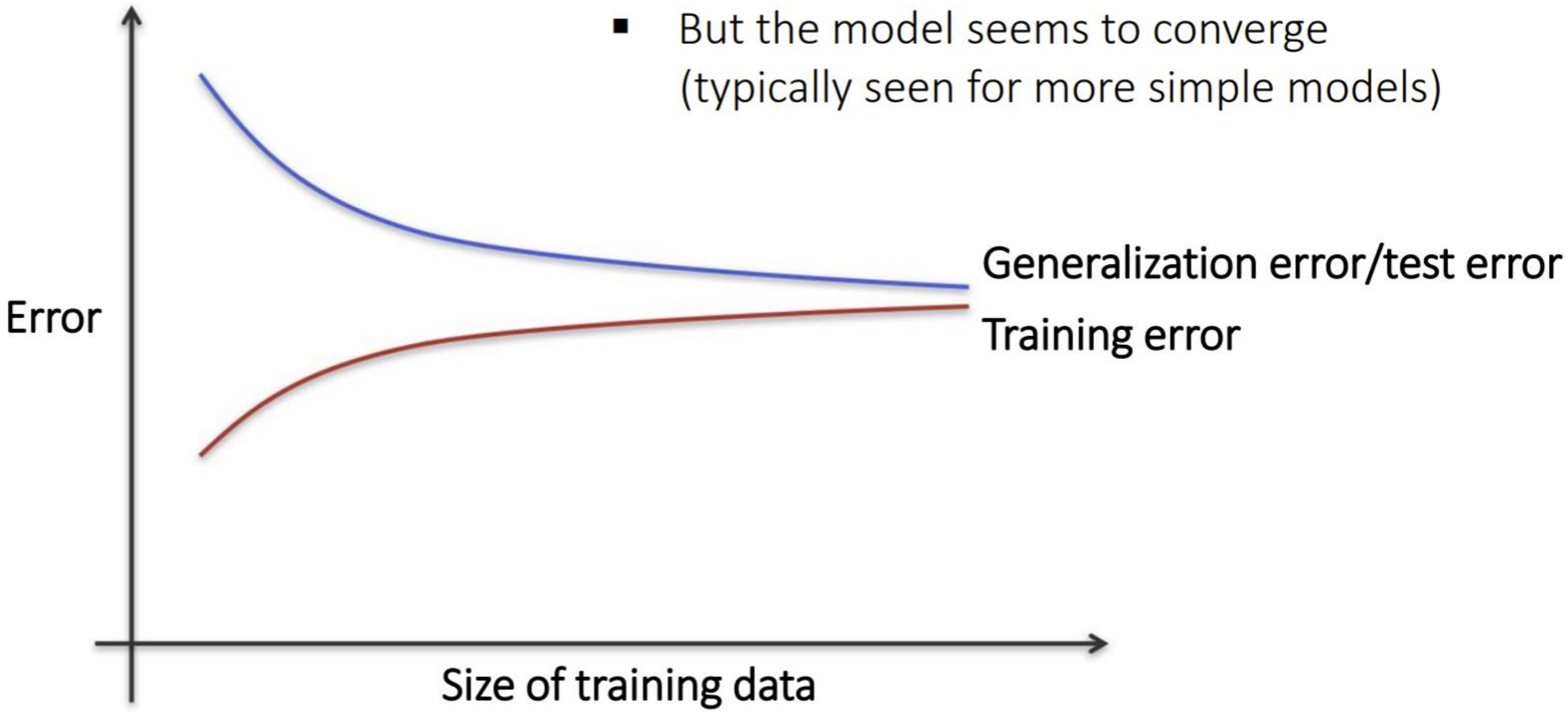
Overfitting

- Test error keeps decreasing as training set increases → more data will help
- **Large gap between training and test error** (typically seen for more complex models)



Underfitting (high bias)

- **Both the train and test error are unacceptable**
- But the model seems to converge (typically seen for more simple models)



Different ways to improve your model

- More training data -> Tackles overfitting
- Features (can add one at a time, measure importance)
    1. Use more features -> Tackles underfitting (kernels, complex models)
    2. Use fewer features -> Tackles overfitting (not all features are important)
    3. Use other features -> Tackles both over and underfitting
- Better Training
    1. Run for more iterations
    2. Use a different algorithm
    3. Use a different classifier
    4. Play with regularization -> Tackles both over and underfitting

First step: diagnose your model

Some possible problems:

- Overfitting (high variance) ☑
- Underfitting (high bias) ☑
- Your learning does not converge
- Are you measuring the right thing?

Learning curves modulo SGD noise

Gradient Descent

- **Local Minima**
- Needs parameter tuning
- Powerful
- Very simple to implement
- Batch gradient descent

Different ways to improve your model

- More training data -> Tackles overfitting
- Features
  1. Use more features -> Tackles underfitting
  2. Use fewer features -> Tackles overfitting
  3. Use other features -> Tackles both over and underfitting
- Better Training
  1. Run for more iterations -> Track objective until convergence
  2. Use a different algorithm
  3. Use a different classifier
  4. Play with regularization -> Tackles both over and underfitting

First step: diagnose your model

Some possible problems:

- Overfitting (high variance) ☑
- Underfitting (high bias) ☑
- Your learning does not converge ☑
- Are you measuring the right thing?

What to measure

- Accuracy / F1 / Performance
- Label imbalance

First step: diagnose your model

Some possible problems:

- Overfitting (high variance) ☑
- Underfitting (high bias) ☑
- Your learning does not converge ☑
- Are you measuring the right thing? ☑

Different ways to improve your model

- More training data -> Tackles overfitting
- Features
    1. Use more features -> Tackles underfitting
    2. Use fewer features -> Tackles overfitting
    3. Use other features -> Tackles both over and underfitting
- Better Training
    1. Run for more iterations -> Track objective until convergence
    2. Use a different algorithm -> Compare your measurement
    3. Use a different classifier -> Compare your measurement
    4. Play with regularization -> Tackles both over and underfitting

Understand your data

- Visualizations are critical
    - *PCA*
    - *Scatter Plots*
    - *Histograms*
- Features might be zeros, or too high or too small
    - *Mean center, scale variance of each feature*
    - *Normalize data (Min-Max scaling* [-1, +1]
    - *Whiten the Data (center the mean, identity covariance)*

Software ethics

- Write clean code
- Understand the operations you are performing on your matrices
- Know matrix shapes before and after every operation
- Write unit tests
- You can test individual parts of your model

Big is not necessarily better

- Simple models
- Ensemble methods
- Do not buy into the hype
- Do what is best for your application

Ethical considerations

- Make life better/ enjoyable for everyone
- Powerful methods
- Prone to biases
- Biased data is everywhere
- Biased models only propagate bias
- Clean data
- Understand where your model could be biases
- Work in rich, diverse teams and create equitable products