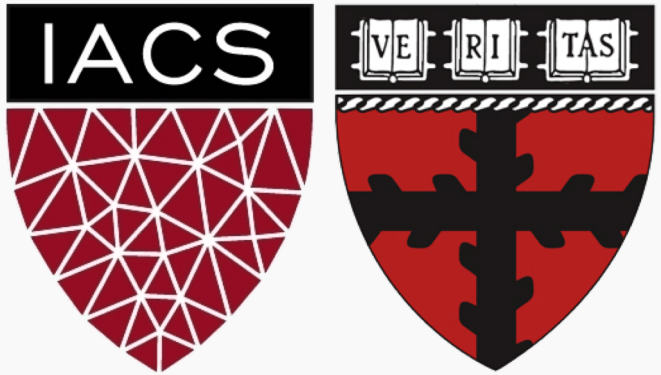
LAB TIME

Lab #4: Demonstration of Dataset Splits

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Tanner



- We are given this data and can do whatever we want with it.

Data

60 observations

- We are given this data and can do whatever we want with it.
- We can use it to train a model!

~~Data~~ Training Data

60 observations

- We are given this data and can do whatever we want with it.
- We can use it to train a model!
- The assumption is that there exists some other, hidden data elsewhere for us to apply our model on. During the training of our model, we never have access to it.

~~Data~~ Training Data

60 observations

Testing Data

10 obs.

- The assumption (and hope) is that our training data is representative of the ever-elusive testing data that our trained model will use

- The assumption (and hope) is that our training data is representative of the ever-elusive testing data that our trained model will use
- Let's say that our model performed poorly on the testing data. What are possible causes?

~~Data~~ Training Data

60 observations

Testing Data

10 obs.

- The assumption (and hope) is that our training data is representative of the ever-elusive testing data that our trained model will use
- Let's say that our model performed poorly on the testing data. What are possible causes?
- How do we know our trained model was trained well?

~~Data~~ Training Data

60 observations

Testing Data

10 obs.

- The assumption (and hope) is that our training data is representative of the ever-elusive testing data that our trained model will use
- Let's say that our model performed poorly on the testing data. What are possible causes?
- How do we know our trained model was trained well?
  - Let's make a synthetic "test" set from our training, for evaluation purposes

- Now we at least have some feedback as to our model's performance before we deem the model to be final.

- Now we at least have some feedback as to our model's performance before we deem the model to be final.
- "Validation Set" is also called "Development Set"
- But some of the same issues exist

- Validation set may be small. Training set may be small.
- In order to (1) train on more data, and; (2) have a more accurate, thorough assessment of our model's performance, we can use ALL of our training data as validation data (in a round-robin fashion)
- This is cross-validation

For a specific parameterization of a model **m**:

Testing Data

10 obs.

Run #	Training Data	Validation Data
1	$x_1 - x_{55}$	$x_{56} - x_{60}$
2	$x_1 - x_{50}; x_{56} - x_{60}$	$x_{51} - x_{55}$
•		
•		
•		
11	$x_6 - x_{60}$	$x_1 - x_5$

- Perform all k runs (k-fold cross validation) for each model **m** that you care to investigate. Average the k performances
- Pick the model **m** that gives the highest average performance
- Retrain that model on all of the original training data that you received (e.g., all 60 observations)