Week 2:

Testing/Training, Cross-Validation, and Bias-Variance

By: Sean Betancourt, Erik Fisher, and Jenny Wang

Outline

A model is useless if we cannot evaluate how well it performs. The main goal of these slides is to teach you how to evaluate the models you build.

a) Training and Testing Workflows
   i) **Getting to Know Your Dataset**
   ii) Cross Validation
   iii) Bias and Variance

Crabs!

Can we predict frontal lobe of crab given other body metrics?

Dataset:

```
import pandas as pd
crab_df = pd.read_csv('data.csv')
crab_df.head()
```

	sp	sex	index	FL	RW	CL	CW	BD
0	B	M	1	8.1	6.7	16.1	19.0	7.0
1	B	M	2	8.8	7.7	18.1	20.8	7.4
2	B	M	3	9.2	7.8	19.0	22.4	7.7
3	B	M	4	9.6	7.9	20.1	23.1	8.2
4	B	M	5	9.8	8.0	20.3	23.0	8.2

...

```
X = crab_df[['RW', 'CL', 'CW', 'BD']].to_numpy()
y = crab_df[['FL']].to_numpy()
```

X

```
array([[ 6.7, 16.1, 19. ,  7. ],
       [ 7.7, 18.1, 20.8,  7.4],
       [ 7.8, 19. , 22.4,  7.7],
       [ 7.9, 20.1, 23.1,  8.2],
       [ 8. , 20.3, 23. ,  8.2],
```

...

y

```
array([[ 8.1],
       [ 8.8],
       [ 9.2],
       [ 9.6],
       [ 9.8],
```

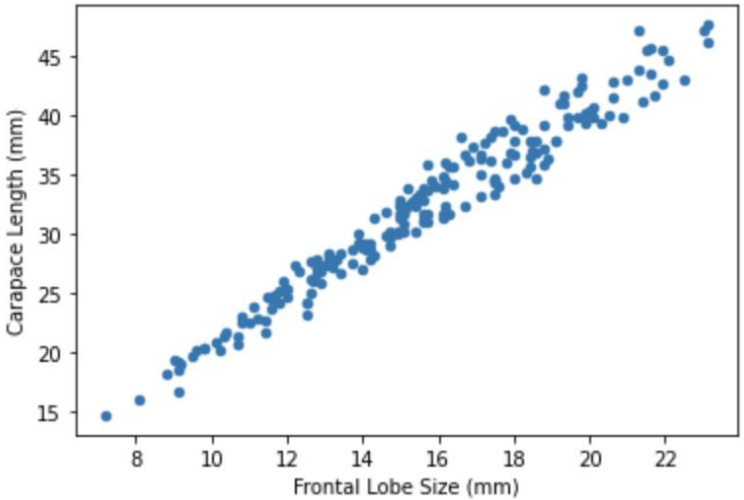
...

Data is from “Crab body metrics” in Kaggle

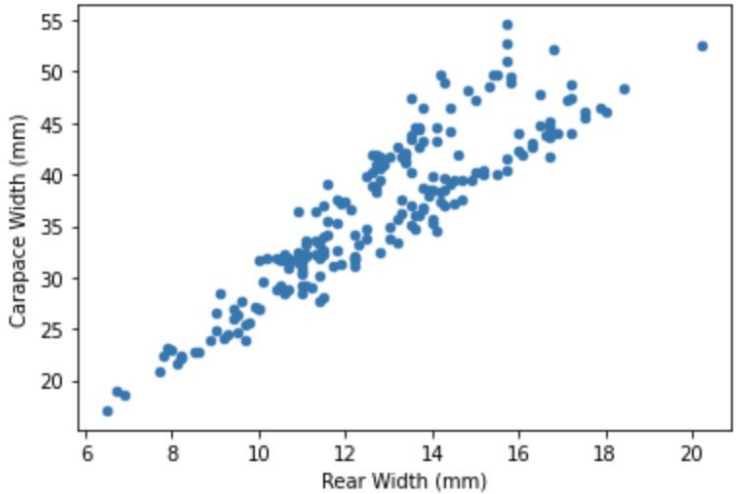
What are we working with?

Let’s plot some features! Do you see strong relationships among features? (trick question!)

```
ax = crab_df.plot.scatter(x='FL', y='CL', marker='o')
ax.set_xlabel("Frontal Lobe Size (mm)")
ax.set_ylabel("Carapace Length (mm)")
```




```
ax = crab_df.plot.scatter(x='RW', y='CW', marker='o')
ax.set_xlabel("Rear Width (mm)")
ax.set_ylabel("Carapace Width (mm)")
```

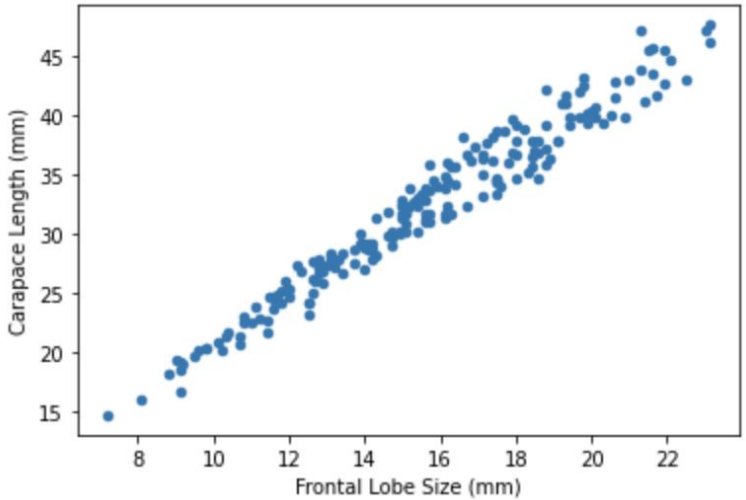


Data is from “Crab body metrics” in Kaggle

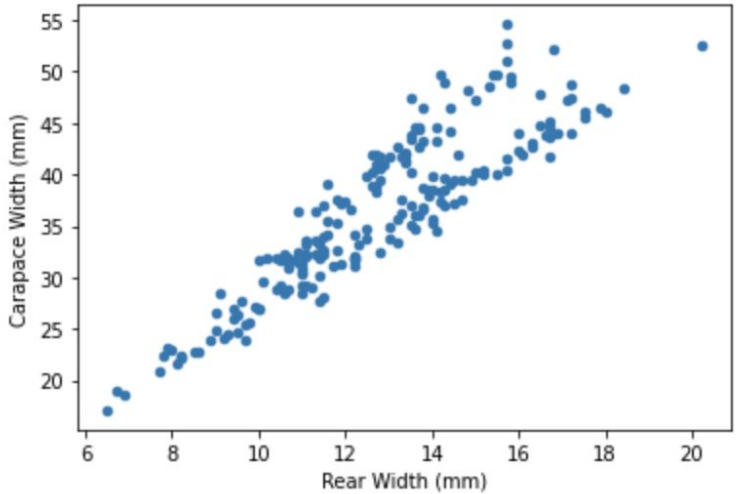
What are we working with?

Let's plot some features! Do you see any strong relationships? (trick question!)

```
ax = crab_df.plot.scatter(x='FL', y='CL', marker='o')
ax.set_xlabel("Frontal Lobe Size (mm)")
ax.set_ylabel("Carapace Length (mm)")
```




```
ax = crab_df.plot.scatter(x='RW', y='CW', marker='o')
ax.set_xlabel("Rear Width (mm)")
ax.set_ylabel("Carapace Width (mm)")
```



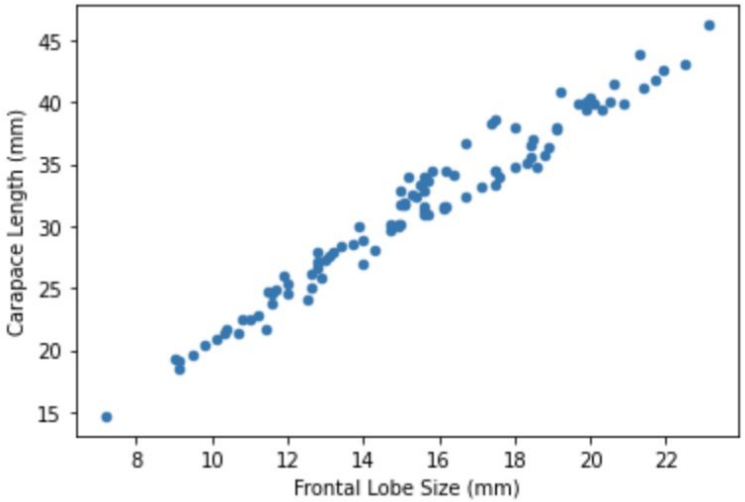
There are two overlapping linear relationships!

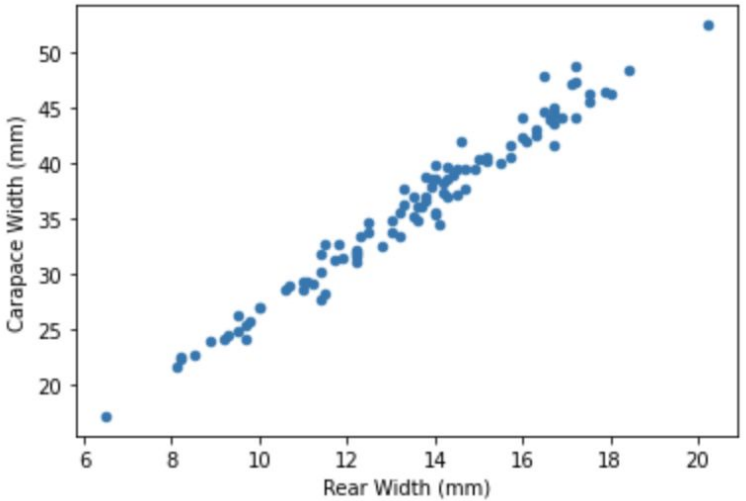
Data is from "Crab body metrics" in Kaggle

What are we working with?

What if we only look at female crabs?

```
crab_f_df = crab_df[crab_df['sex'] == 'F']
```





Great! We might be able to recover a strong relationship!

Data is from “Crab body metrics” in Kaggle

Outline

A model is useless if we cannot evaluate how well it performs. The main goal of these slides is to teach you how to evaluate the models you build.

a) Training and Testing
   i) Getting to Know Your Dataset
   ii) Cross Validation
   iii) Bias and Variance

Measures of Success (for regression problems)

Let's say the model is Support Vector Regression (SVR; a variant of SVMs from the previous lecture) with initialized with default parameters

How do you know the model does well for this problem?

Idea 1: Correlation- won't generalize to nonlinear models

Idea 2: Prediction error

- **Mean Squared Error (MSE)** $\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$
- Root Mean Squared Log Error (RMSLE) $\sqrt{\frac{1}{N}\sum_{j=1}^{N}(log(y_j) - log(\hat{y}_j))^2}$
- ... and more

Measuring Prediction Error when Iterating on a Model

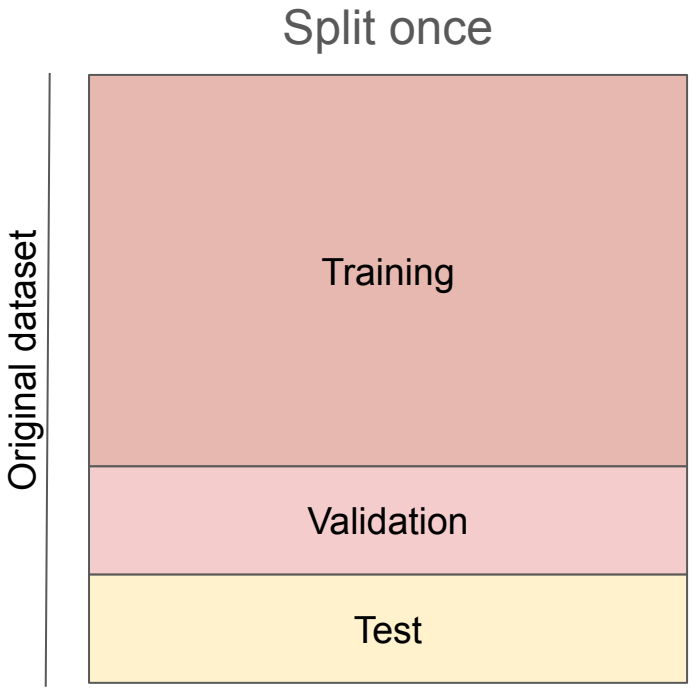
If we use the same training data to fit and assess, that assessment will overestimate how well the model does at test time.

This is often called “data incest”

Instead, we should section off a **validation set** from the training set.

Splitting a Dataset

Training, Validation, and Test sets



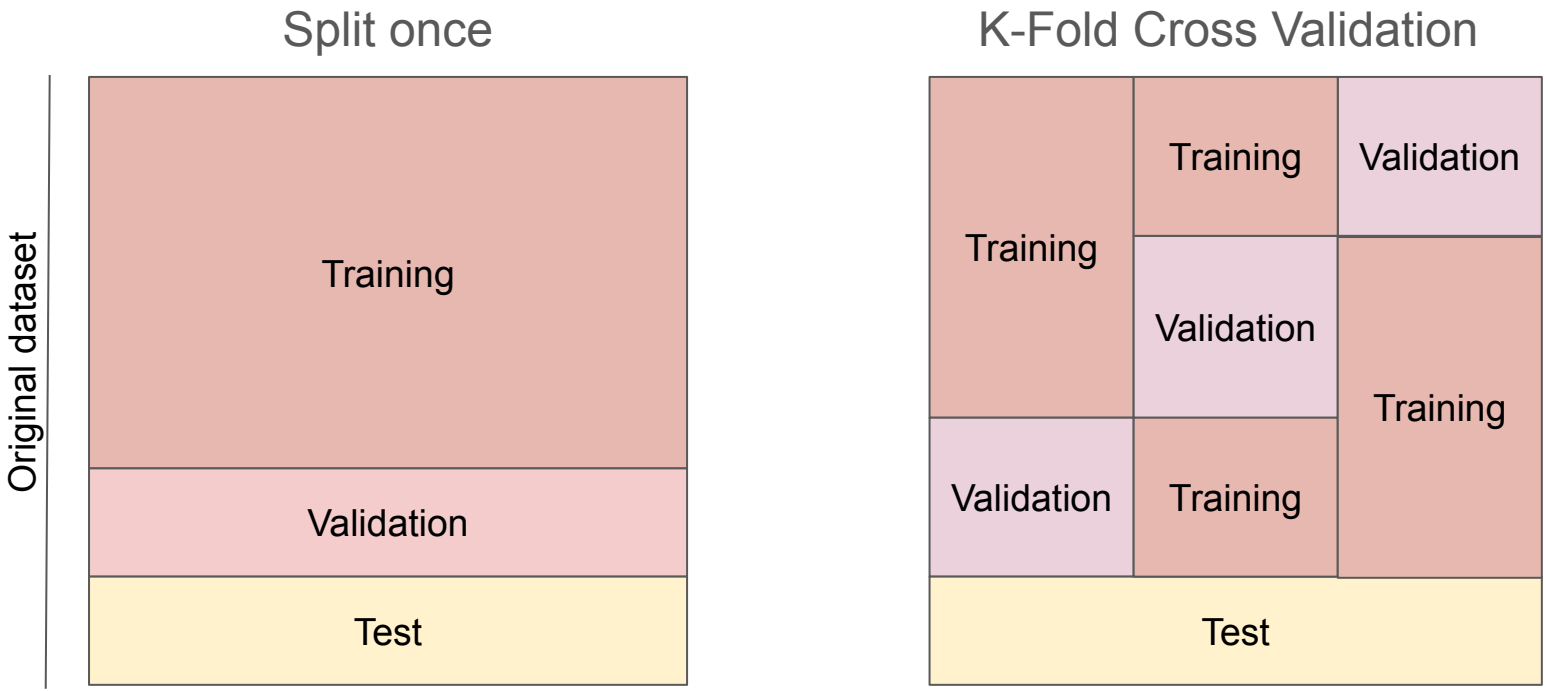
Training- find good weights for the model

Validation- test the model's structure, itself

Test- find final test accuracy (only use ONCE)

Splitting a Dataset with Limited Amounts of Data

When there isn't enough training data but we still want to validate



Splitting a Dataset with Limited Amounts of Data

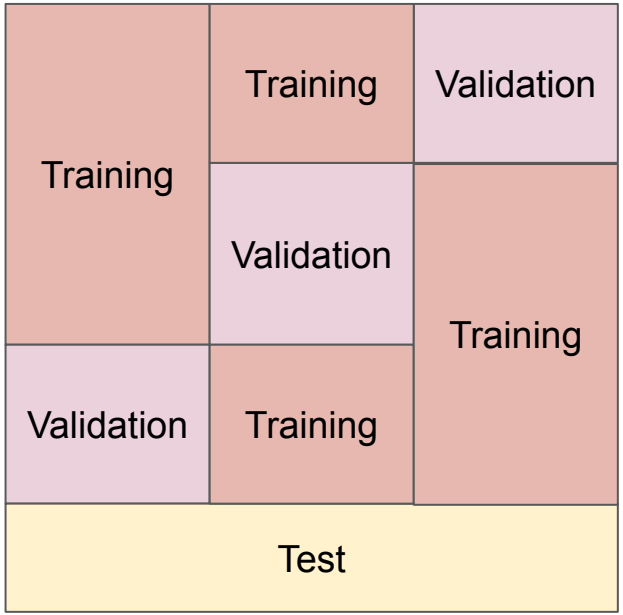
When there isn't enough training data but we still want to validate

For each of the K splits:

- Train with the training data
- Compute accuracy with the validation set

Validation accuracy is the average of all K accuracies

K-Fold Cross Validation

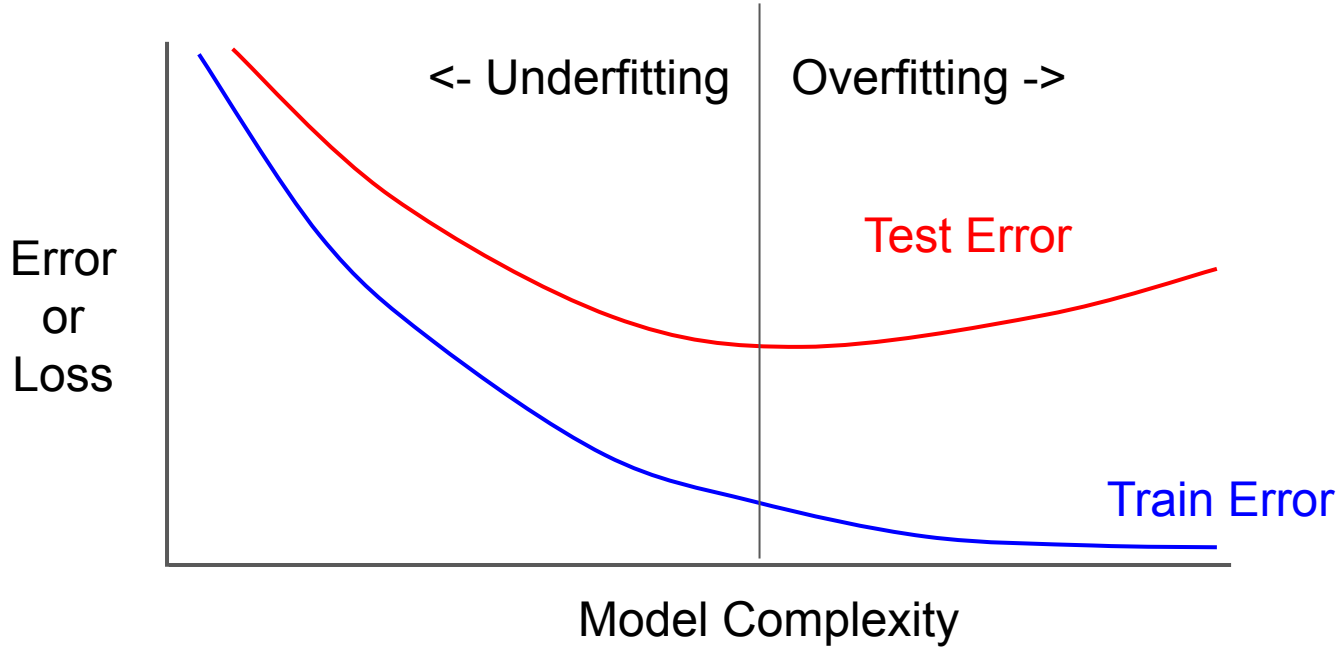


Outline

A model is useless if we cannot evaluate how well it performs. The main goal of these slides is to teach you how to evaluate the models you build.

a) Training and Testing
   i) Getting to Know Your Dataset
   ii) Cross Validation
   iii) Bias and Variance

A Common Trend You May Find



Theoretically, this is described by the bias-variance tradeoff

Bias and Variance Definitions

Bias- Erroneous assumptions in the model

Variance- Sensitivity to small fluctuations in the training set

---

For mean squared error:

$$\text{Total Noise} = \underbrace{(E[h(x|D)] - f(x))^2}_{\text{Bias}} + \underbrace{Var(h(x|D))}_{\text{Variance}} + \underbrace{Var(Z)}_{\text{Irreducible Error}}$$

where h(x|D) is the model's prediction given a training dataset, f(x) is the true label, and Z is the inherent noise in the labels

(a full derivation is in the notes)

Breaking Down the Variance Term for MSE

$$Var(h(x|D))$$

where h(x|D) is the model's prediction given a training dataset, f(x) is the true label, and Z is the inherent noise in the labels

Random variable- has an unknown value that varies by a probabilistic distribution

Since the dataset D is a random variable, the prediction h(x|D) is a random variable

“Variance of the model’s prediction for x given various training datasets”

Breaking Down the Bias Term for MSE

$$(E[h(x|D)] - f(x))^2$$

where h(x|D) is the model's prediction given a training dataset, f(x) is the true label, and Z is the inherent noise in the labels

E[h(x|D)] = **expectation** of the model's prediction over the distribution of datasets D

= $\sum_{i=1}^{N} h(x|D_i)P(D_i)$ (when there are N possible datasets)

$(\hat{f}(x) - f(x))^2 = (error)^2$ squared error

"Squared error between what is expected from model predictions of x and true labels"

Mini Quiz!

What are relative training and test accuracies when the model has...

- High bias and low variance
  - ?
- Low bias and high variance
  - ?

Mini Quiz!

What are relative training and test accuracies when the model has...

- High bias and low variance
  - Training and test accuracies are both low, but about equal
- Low bias and high variance
  - Training accuracy is much higher than test accuracy

We will be implementing an empirical calculation in the project!

Summary: Training and Testing

1. **Visualize** and understand the data
2. **Split** dataset into the train and test sets
3. Formulate a **model** (ex: OLS)
4. Either
    a. Section off part of the train set as a **validation set**, if there are lots of samples
    b. Use **cross validation** to section off certain parts of the train for **validation** at a time, if data is limited
5. Calculate **validation accuracy or loss**. Perhaps also compute the **bias** and **variance** of the model given various training datasets to help you understand how to improve the model.
6. **Repeat** 2-5 until the model does well enough on the validation set
7. Calculate **test accuracy or loss**