GENERAL ASSEMBLY

DATA SCIENCE

MID-COURSE REVIEW

Data Science vs Machine Learning

DATA SCIENCE

- The art of extracting knowledge from data

- *Could be a visualization*

- *Could be a statistical hypothesis test*

MACHINE LEARNING

- Algorithms that are self-correcting and self taught

- *Machine Learning is a part of data science*

Supervised vs Unsupervised Learning

SUPERVISED LEARNING

- What is the point?
- *How do we evaluate it?*

UNSUPERVISED LEARNING

- What is the point?
- *How do we evaluate it?*

Regression vs Classification

REGRESSION VS. CLASSIFICATION

- Regression is predicting a continuous variable

- Classification is predicting a categorical variable

- That’s it!

Cross Validation

Why do we do Cross Validation?

To prevent overfitting!!

- in any dataset we have the signal and the noise. A great model is only capturing the signal while an overfit model is also trying to predict the noise

- It’s like bringing a very powerful microphone into a recording studio that hears the band playing AS WELL AS the background noise in the room. We really only want the band.

- Bonus question: why do you think R squared is a bad metric to detect overfitting?

- Over-complicating models (too many unnecessary predictors) is a great way to overfit a model.
  - Use a combination of cross validation and EDA to prevent overfitting

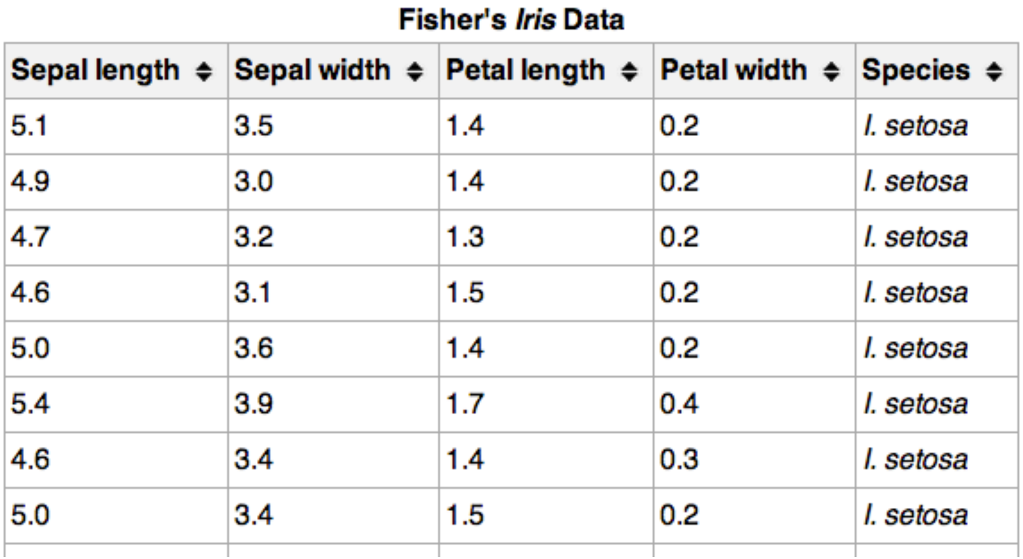
Bias vs Variance

- Bias is the measure of how off the model is (residuals)
  - Low bias models tend to have low training error
  - High bias models tend to have high training error

- Variance is a measure of how random sampling affects our models
  - Low variance models tend to be more “Stable” on random samples
  - High variance models tend to be less “Stable” on random samples (less reliable in the wild)

These are called

______



Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

These are called ______

This is called

______