INTRO TO MACHINE LEARNING (ML) TERMINOLOGY

Matthew Hall

REGRESSION VS CLASSIFICATION

- **Predict Continuous Response Variable**
  - Sales, Time, Stock Price, Returns, Cost, Default Risk Pct
- **Predict Categorical Response Variable**
  - Yes/No, A/B/C/D, Buy/Sell, Default Yes/No, Underwrite Loan Yes/No, Fraud Detection

THE GOOD THING

- **Regression / Classification is Probably Chosen For You Based Upon Dataset**

SUPERVISED VS UNSUPERVISED

- **Requires Labeled Training Data Points**
- **Need The Response Variable, Y!**
- ***Regression Falls Under This Category***

- **Does Not Require Labeled Training Data**
- **Do Not Need To Input Response Variable, Y!**
- ***Clustering Falls Under This Category***

TRAIN – TEST – VALIDATION

- How Do We Train & Evaluate Our ML Models?
- Need To Provide It Data!
- However, We Cannot Train Model On All of Our Data
- Need To Keep Part of Dataset Reserved For Checking If Model is a “Good Fit”

3 SPLITS

- **Training Sample: Data Used to “Train” Model and Detect Patterns**
- **Validation Sample: Data Used to Measure “Accuracy” While Tuning Models for Fit**
- **Testing Sample: Final, Unbiased Sample to Gauge Performance of Model**

CROSS VALIDATION

- **Way To Test Question *"Are My Results The Way They Are Because I Chose a Lucky Sample?"***
- **Cross Validation Takes Randomized Splits In Dataset, Trains a New Model For Each Split**
- **We Can Then Compare Coefficients, "Accuracy," and Model Stability Across Folds**
  - Folds Refer to The Splits in the Dataset

UNDERFITTING VS OVERFITTING

- **Model Can Sometimes Fail To Detect Patterns of Dataset**
- **Misses General Trend in Population As Well**
- **Model Overanalyzes Patterns in Training Sample**
- **Cannot Be Used to Predict Population Characteristics Reliably**

ML IN PYTHON

- Can Use SciKit Learn (SKLearn) Package
- Don’t Need to Worry About “Behind The Scenes” Mathematics & Statistics
- But You Should Know What Each Algorithm Tries to Do, So You Can Properly “Tune” Models


MELIORA