LECTURE 05

Modeling

October 23, 2023

PBHLTH 198, Fall 2023 @ UC Berkeley

Andrew O'Connor

Class Outline

- Recap of hw 4
- High level introduction to modeling from ML perspective
- Types of modeling methods

Modeling

Modeling

- Why do we want to build models? As far as data scientists and statisticians are concerned, there are two reasons, and each implies a different focus on modeling
- Two main reasons

1. **To explain complex phenomena occurring in the world we live in**
2. **To make accurate predictions about unseen data**

Modeling in Biostatistics

- In biostatistics and epidemiology, data are collected by researchers under **strict experimental design**
- Machine learning techniques are *not* traditional tools in biostatistics and epidemiology
- Why learn about it?
    - Embracing machine learning can empower you to explore new horizons in public health research and data analysis, contributing to the ever-evolving field of epidemiology and biostatistics
    - By understanding the potential applications of machine learning in biostatistics and epidemiology, you are better prepared to leverage its strengths when needed, even within traditional frameworks
    - New Frontiers: Big data, Machine learning for causal inference

Basic Scenario

- How can we train a **model** to **predict an outcome** based on **existing data**

Terminology

- **Dataset**: The existing data you have
- **Model**: The abstract goal of what you're trying to predict
- **Algorithm**: The math/statistics behind what you're trying to achieve
- **Metric**: A measure or statistic we can use to determine how well our model predicts an outcome
- **Feature/Attributes**: Columns/categories in your dataset
- **Training set**: A subset of the full dataset we use to train our model on
- **Test set**: A subset of the full dataset we use to test out model on

Training set + Test set = Full dataset

Modeling Process

Step 1: Determine what kind of question you're asking. What is your expected output?

Step 2: Split your dataset (train-test split, validation/holdout)

Step 3: EDA, Feature selection, Feature engineering

Step 4: Train your model on the training set, Predict values on the test set

Step 5: Evaluate your model

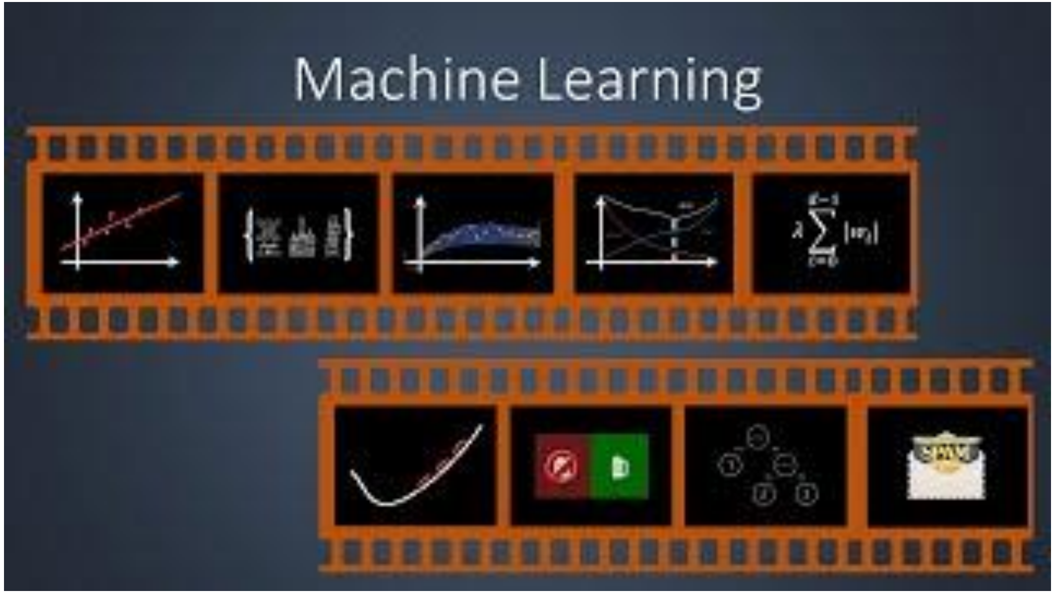
Step 1: What are we doing?

- Are you looking to predict a **numeric value**?
  - **Regression** (Simple linear regression, Multiple linear regression, Random Forest regressor, Decision Tree Regressor, Poisson Regression)
- Are you looking to **group** up individuals into **categories**?
  - **Clustering** (K-nearest neighbors, K-means clustering, DBSCAN clustering)
- Are you looking to determine if something is **either** X or Y?
  - **Classification** (Decision Trees, Random Forests, Neural Networks, Naive Bayes)

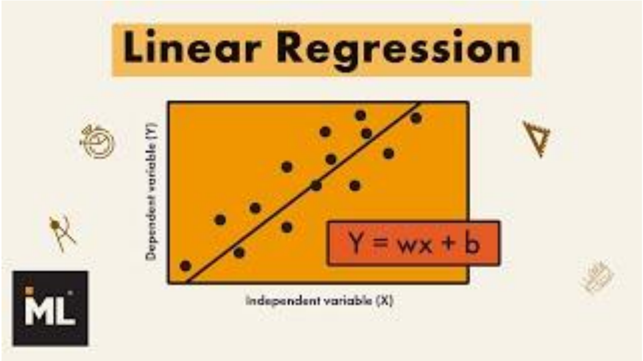
EX: {Feature 1, log(Feature 2), log(Feature 4 * Feature 3)} -> Predict numerical value (Regression)

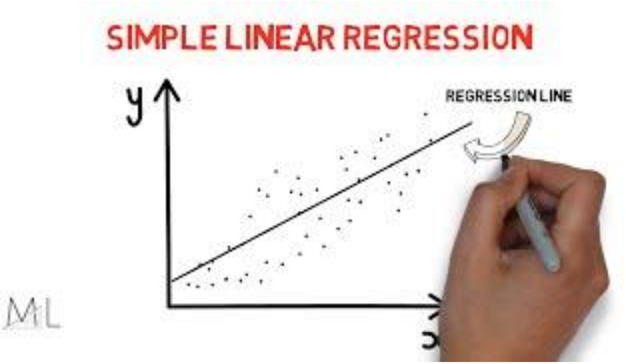
EX: {Feature 2, Feature 3, Feature 7} -> Predict if datapoint is 0 or 1 (Binary Classification)

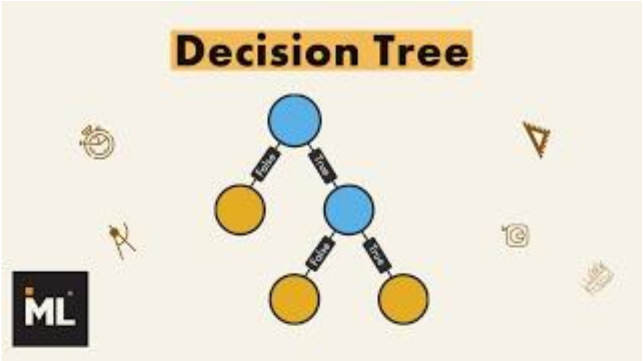
Step 1: What are we doing?



Step 1: What are we doing?







For the following scenarios, what kind of model would you use to predict new values?

1. Predicting patient blood pressure
2. Predicting whether or not someone survived on the titanic
3. Predicting what group a patient belongs to
4. Predict whether a person is "Healthy" or "Diabetic"

Exercise: Model Type

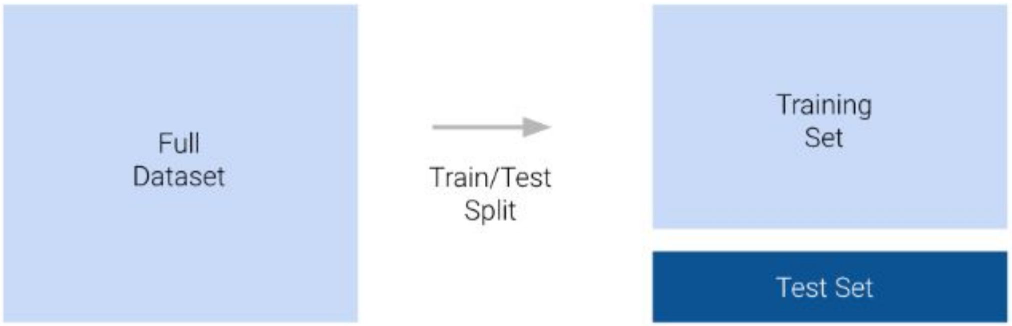
Exercise: Model Type

For the following scenarios, what kind of model would you use to predict new values?

1. Predicting patient blood pressure
   a. **Regression**
2. Predicting whether or not someone survived on the titanic
   a. **Classification**
3. Predicting what group a patient belongs to
   a. **Clustering**
4. Predict whether a person is "Healthy" or "Diabetic"
   a. **Classification**

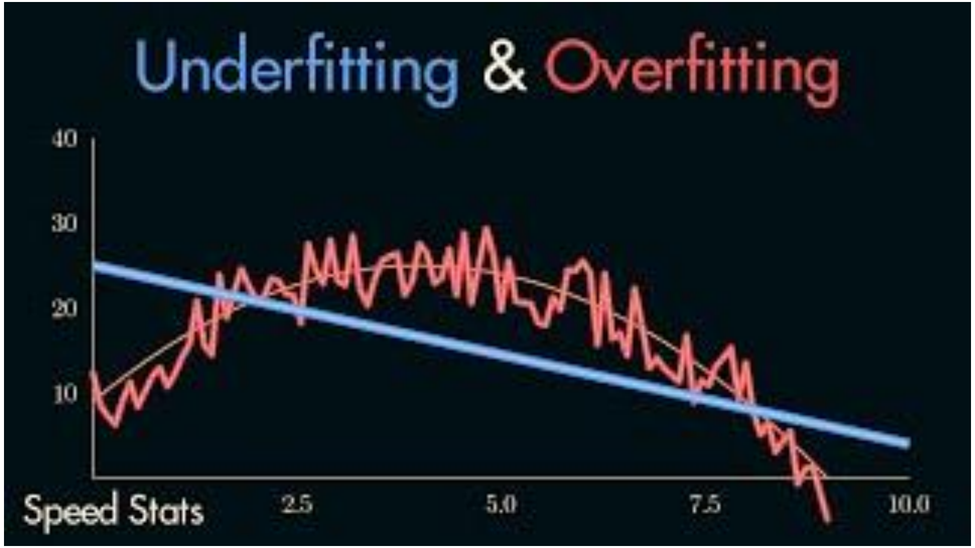
Step 2: Splitting the Dataset

- Why do we need to split our dataset? Can't we just train a model using the full dataset?
  - No, then your model will **overfit** the training data
  - Splitting the dataset into 2 sets will help avoid our model becoming too "familiar" with the training dataset

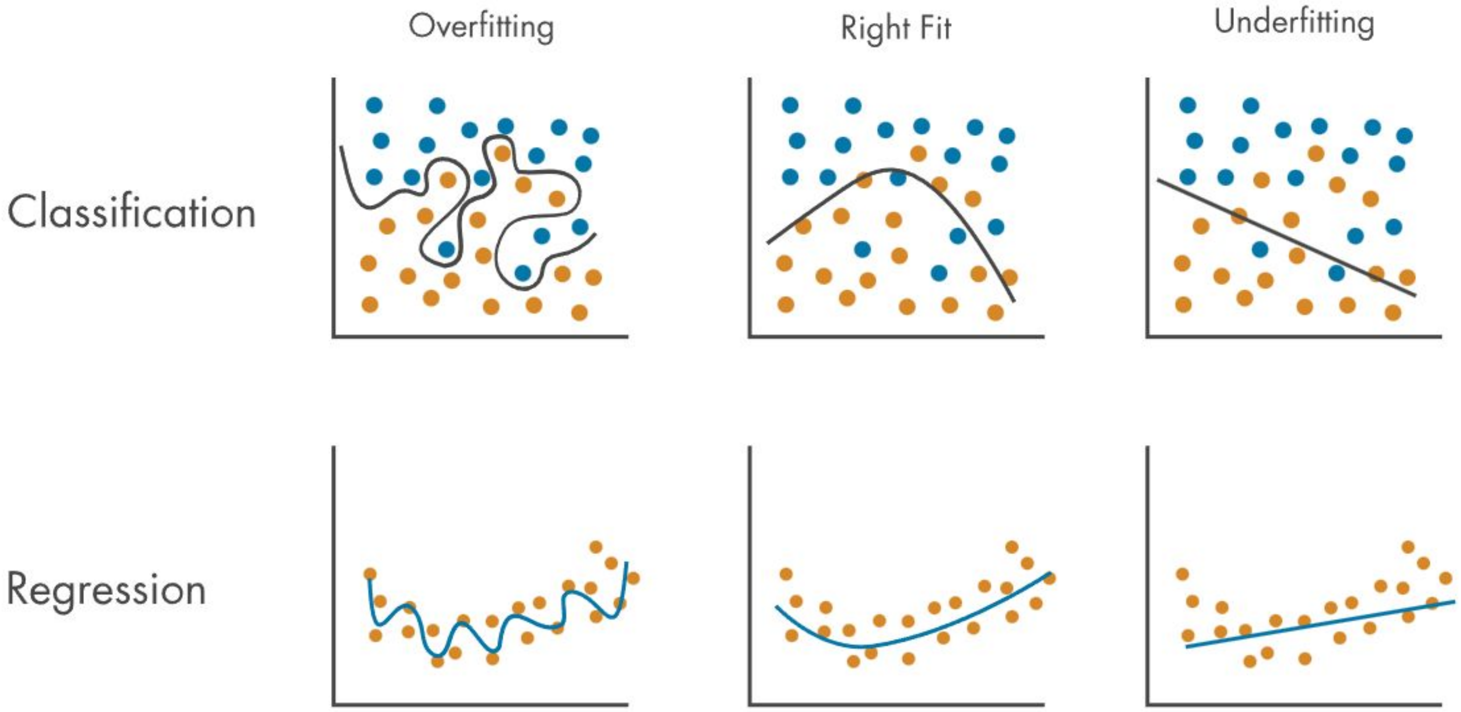


Overfitting & Underfitting

- Remember, our goal is to predict an outcome **based** on existing data
- Our data is (almost always) just a sample of the population
  - Samples don't always represent the population well…



Overfitting & Underfitting



Step 3: Feature Selection/Engineering

- EDA: Exploratory Data Analysis
  - Explore your dataset; What attributes are you working with?
- Feature Selection
  - What features/columns does your model rely on?
    - EX: Suppose we want to predict height based on age, weight and sex; Consider, does having race, gender, religion, political affiliation help our model?
  - Too many features in your model -> Increase in model complexity -> **Overfitting**
  - Not enough important/significant features -> decrease in model complexity -> **Underfitting**
- Feature Engineering
  - Sometimes we need to transform values due to the nature of our **loss function,** sometimes we have categorical variables that need to be **one-hot-encoded**, New features may have meaning when they are multiplied, added, divided -> **interaction terms**

Exercise: Feature Selection

Suppose we want to predict **Height** from a dataset with the following features:

Height, Religion, Age, Weight, Sex, Gender, Political Affiliation, CountryOfBirth, GPA, Student ID, WearsGlasses, BloodType

Which features would you use? Why?

Exercise: Feature Selection

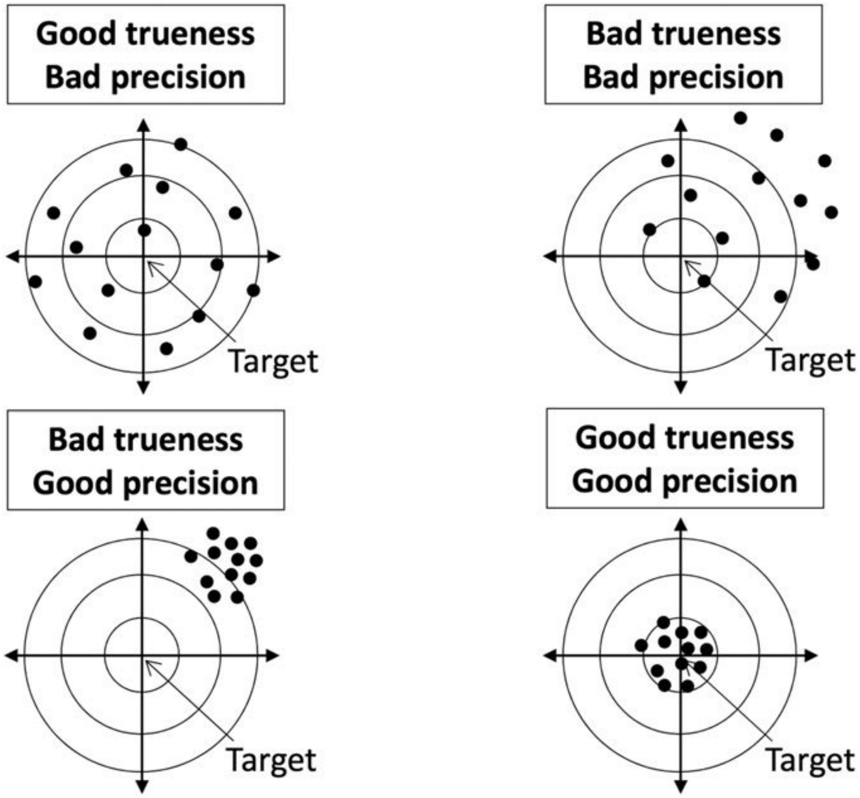
Suppose we want to predict **Height** from a dataset with the following features:

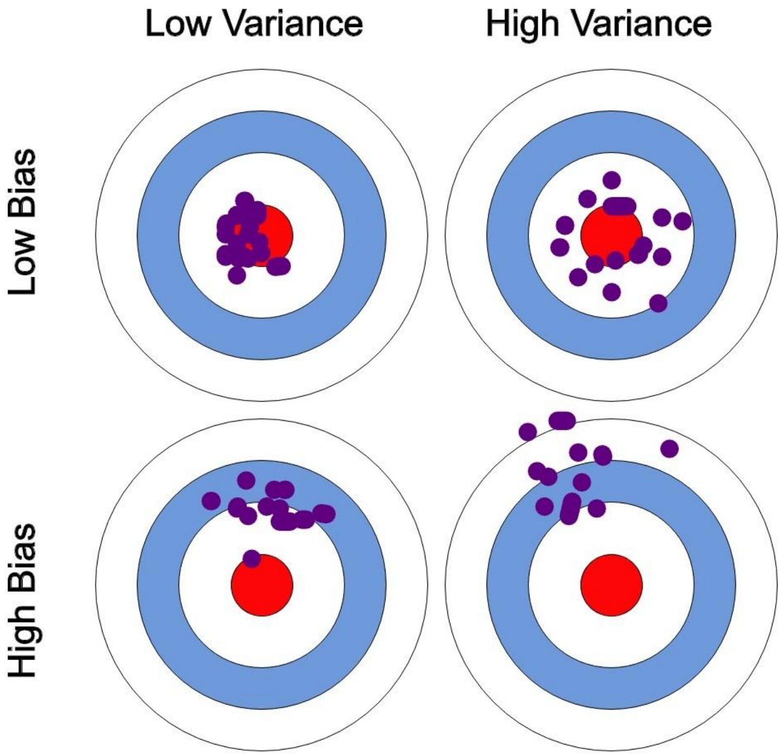
Height, Religion, Age, Weight, Sex, Gender, Political Affiliation, CountryOfBirth, GPA, Student ID, WearsGlasses, BloodType

Which features would you use? Why?

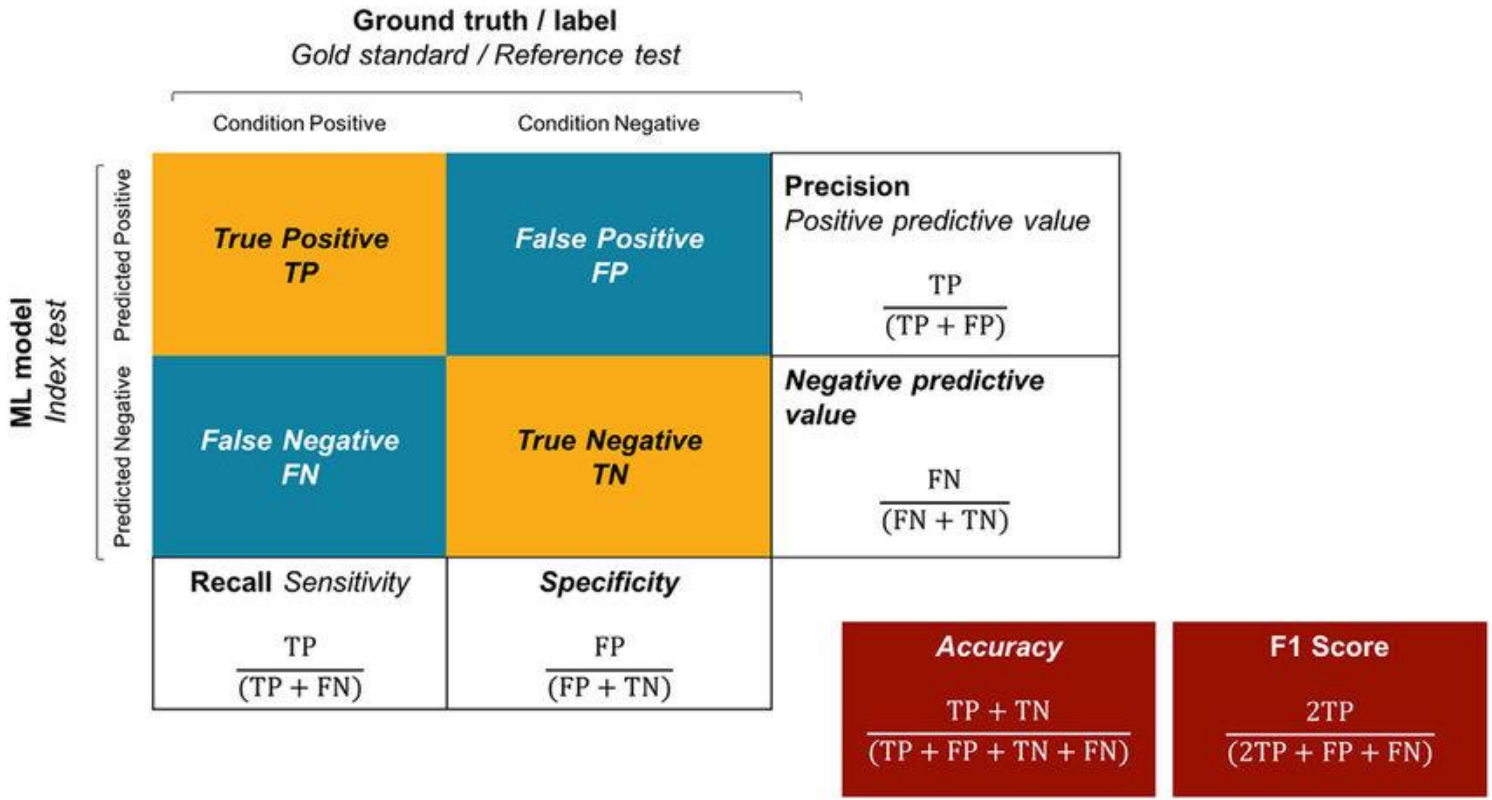
Consider the bias-variance trade off. If we increase model complexity, what does that mean for our model? Are all these features really necessary to determine height?

Step 5: Evaluate Model





Evaluation Metrics: Classification



hw 5 time