Untitled

James Scott

1/13/2021

Outline

1. Introduction
2. Parametric vs nonparametric models: KNN
3. Measuring accuracy
4. Out-of-sample predictions
5. Train/test splits
6. Bias-variance tradeoff
7. Intro to classification

Introduction to predictive modeling

The goal is to predict a target variable (y) with feature variables (x).

- Zillow: predict price (y) using a house's features ($x =$ size, beds, baths, age, . . .)
- Citadel: predict next month's S&P (y) using this month's economic indicators ($x =$ unemployment, GDP growth rate, inflation, . . .)
- MD Anderson: predict a patient's disease progression (y) using his or her clinical, demographic, and genetic indicators (x)
- Etc.

In data mining/ML/AI, this is called "supervised learning." We've already seen a simple example (OLS with one x feature)

Introduction to predictive modeling

A useful way to frame this problem is to think that y and x are related like this:

$$y_i = f(x_i) + e_i$$

where: - y_i is a scalar *outcome* or *target* variable
- $x_i = (x_{i1}, x_{i2}, ...x_{iP})$ is a vector of features - f is an unknown function

Our main purpose is to *learn* $f(x)$ from the observed data.