STATS / DATA SCI 315

Basic Elements of Linear Regression

Kinds of ML problems

- *Supervised learning*: have labeled data
    - Regression: Label is a real number
    - Classification: Label is chosen from a finite set
- *Unsupervised learning*: have only unlabeled data
- *Reinforcement learning*: interact with an environment and receive rewards

Regression

- Modeling the relationship between input variables and a real-valued output
- Input variables also called:
    - Features
    - Covariates
    - independent variables
- Output variable also called:
    - Label
    - Target
    - Dependent variable
- ``Regress y on **x**" means "Run a regression with **x** as input, y as output"

Examples

- predicting prices (of homes, think of Zillow's Z-estimate)
- predicting length of stay (for patients in the hospital)
- demand forecasting (for retail sales)

Linear regression

- A special but important case of regression
- Model the relationship of y, the output variable, as linear in **x**
- Wish to estimate the prices of houses (in dollars) based on their area (in square feet) and age (in years)
- *Linearity assumption*: target (price) can be expressed as a weighted sum of the features (area and age):

$$\text{price} = w_{area} \cdot \text{area} + w_{age} \cdot \text{age} + b$$

Weights and bias

- w_{area} and w_{age} are called *weights*
- b is called a *bias* (also called an *offset* or *intercept*)
- Strictly speaking, our model for price involves an *affine transformation*
- Affine = Linear + bias
- What makes a model good?
- How do we find good values for the weights and bias?

Training Dataset

- Need a dataset where we know the sale price, area, and age for each home
- This is called a *training dataset* or *training set*
- Put one sale info on each row
- Each row is called an *example* (or *data point*, *data instance*, *sample*)
- Each example has
    - A label (price)
    - Features (area, age)