STA130 Winter 2022, Tutorial 8

Ente Kang

University of Toronto

2022-03-20

Machine Learning Demystified

Supervised Learning

- In naive terms: *There is a target variable that we want to predict*

Unsupervised Learning

- In naive terms: *We want to explore structures and groupings within our data*

Terminology

Fill in the blanks

- Train-test split
    -
- Training data
    -
- Test data
    -
- Fitting
    -
- Validation
    -

Classification

- Supervised Learning method
- A **categorical** target variable
- Predictors can be of any type
- Our goal is to minimize the following:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_0 \neq \hat{y}_0)$$

which can be thought of as an average error rate

Classification trees

- They are a subset of the available classification algorithms
- Strengths
  - Easy to interpret
- Weaknesses
  - High variation and risk of overfitting

Classification in our daily lives

- Xbox fitness and Wii sports
- Fraud detection
- Who should we hire?

Be aware of the ethical implications!