greatlearning
Learning for Life

Introduction to Machine Learning

Machine Learning

Learning from Data

- Can we learn about the world around us using data?
- Model building from data
  - Take data as input
  - Find patterns in the data
  - Summarize the pattern in a mathematically precise way
- Machine learning automates this model building.

The Challenge

- Data unfortunately contains noise. If not, machine learning would be trivial!
- Think of Data = Information + Noise
- The challenge is to identify the information content and distill away the noise.
- To help do this, machine learning uses a train and test approach.

Over fitting Vs under fitting

- If the model we finish with ends up
  - modeling the noise as well, we call it “over fitting” - bad for prediction!
  - not modeling all the information, we call it “under fitting” - bad for prediction!
- The hope is that the model that does the best on testing data manages to capture/model all the information but leave out all the noise.

Machine Learning tasks

1. **Supervised learning:** Building a mathematical model using data that contains both the inputs and the desired outputs (ground truth).

    - Examples:

        - Determining if an image has a horse. The data would include images with and without the horse (the input), and for each image we would have a label (the output) indicating if there is a horse in that image.

        - Determining is a client might default on a loan

        - Determining if a call center employee is likely to quit

    - Since we have desired outputs, model performance can be evaluated by comparisons.

Machine Learning Tasks

2. **Unsupervised learning:** Building a mathematical model using data that contains only inputs and no desired outputs.

- Used to find structure in the data, like grouping or clustering of data points. To discover patterns and group the inputs into categories.

- Example: an advertising platform segments the population into smaller groups with similar demographics and purchasing habits. Helping advertisers reach their target market with relevant ads.

- Since no labels are provided, there is no specific way to compare model performance in most unsupervised learning methods.

Tools and techniques

- Supervised learning
  - Regression: desired output is a continuous number
  - Classification: desired output is a category
- Unsupervised learning
  - Clustering: Grouping data
  - Dimensionality reduction: Compressing data
  - Association rule learning: If X then Y