Machine Learning

Nicholas Lambert

6/25/2021

Definition

"The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience."[1]

Origin

Computer Science:

- how to manually program computers to solve tasks

Statistics:

- what conclusions can be inferred from data

Machine Learning:

- intersection of computer science and statistics
- how to get computers to program themselves from experience plus some initial structure
- effective data capture, store, index, retrieve and merge
- computational tractability[2]

Types of machine learning

Machine learning approaches are divided into two main types

- **Supervised**
  - training of a "predictive" model from data
  - one (or more) attribute of the dataset is used to "predict" another attribute e.g., classification
- **Unsupervised**
  - discovery of descriptive patterns in data
  - commonly used in data mining e.g., clustering

Supervised

- Training dataset
  - input attribute(s)
  - attribute to predict
- Testing dataset

[1] Mitchell, T. (1997). Machine Learning. McGraw Hill.

[2] Mitchell, T.M., 2006. The discipline of machine learning (Vol. 9). Pittsburgh, PA: Carnegie Mellon University, School of Computer Science, Machine Learning Department.

    - input attribute(s)
    - attribute to predict
- Type of learning model
- Evaluation function
    - evaluates difference between prediction and output in testing data

Unsupervised

- Dataset
    - input attribute(s) to explore
- Type of model for the learning process
    - most approaches are iterative
    - e.g., hierarchical clustering
- Evaluation function
    - evaluates the quality of the pattern under consideration during one iteration

Semi-supervised learning

Supervised learning requires “labelled data”

- which can be expensive to acquire

Semi-supervised learning

- combines a small amount of labelled data with a larger un-labelled dataset
    - train on small labelled dataset
    - apply model to larger unlabled dataset generating “pseudo-labels”
    - re-train the model with all data (including “pseudo-labels”)
    - assumptions: continuity, cluster, and manifold (lower dimensionality)

Reinforcement learning

Based on the idea of training agents to learn how to

- take actions
    - which affect: agent state, environment
- to maximize reward
- balancing:
    - exploration (new paths/options)
    - exploitation (of current knowledge)

Overfitting

- creating a model perfect for the training data but not generic enough to be useful for prediction

- An issue for machine learning e.g., regression n predictors can generate a line fitting the data exactly n cases Occam’s razor one in ten rule 10 cases per predictor

Algorithmic bias

Assumptions and training dataset quality still matter!

- garbage in, garbage out

Joy Buolamwini and Timnit Gebru’s work on facial recognition

- black women were 35% less likely to be recognised than white men.
- Buolamwini, J. and Gebru, T., 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91).
- see also, Facial Recognition Is Accurate, if You're a White Guy by Steve Lohr (New York Times, Feb. 9, 2018)

Summary

Machine Learning

What's Machine Learning? Types Limitations Next: Artificial Neural Networks

Logistic regression Artificial neural networks Deep learning