Introduction to Machine Learning

Parimal Parag
Department of ECE
Indian Institute of Science

DRDO Center for Airborne Systems
Bangalore

Nov 30, 2023

What is Machine Learning?

Machine learning

Machine learning is a collection of computational methods to improve performance or make predictions using *experience*.

Experience

Experience is the past information available to the learner. Information may be readily available as digitized human-labeled training sets, or can be obtained via interaction with the environment.

Learning techniques

Data-driven methods with relations to computer science, statistics, probability, and optimization.

Why ML?

Practical objectives

- accurate predictions of unseen items, and
- design of efficient, robust, and scalable prediction algorithms.

Measurement of quality of ML algorithms

- time complexity: running time of the algorithm,
- space complexity: memory requirements of the algorithm, and
- sample complexity: sample size required for the algorithm to learn a family of concepts

When will ML succeed?

The success of prediction depends on the size and the quality of data instances.

Learning guarantees

The theoretical learning guarantees depend on

- complexity of concept class, and
- size of the training sample.

Fundamental questions

- Which concept families can be learned, and under what conditions?
- How well can these concepts be learned computationally?

Input space

Examples and features

Examples are items or data instances. An example is typically represented by a vector $x \in \mathcal{X}$, where the components of an example are its features.

Input space

The set of all possible examples is called the input space and is denoted by $\mathcal{X}$. If an example has N attributes and all of them can be represented by real numbers, then the feature set $\mathcal{X} = \mathbb{R}^N$ with $N \geqslant 1$.

Feature extraction

Features are set of data attributes often represented as a vector associated with an example.
Feature extraction from examples is a domain-specific task done by the experts and is critical to the successful prediction.

Ouput and prediction space

Labels

Labels or targets are values of categories assigned to examples. Labels are discrete for classification and real-valued for regression.

Output space

The set of all possible labels is called the output space and is denoted by $\mathcal{Y}$. For binary classification, output space could be $\mathcal{Y} = \{0, 1\}$ or $\{-1, 1\}$.

Prediction space

The set of all predicted labels is called the prediction space and is denoted by $\mathcal{Y}'$. The set of predictions $\mathcal{Y}'$ may not necessarily be equal to the set of labels $\mathcal{Y}$.

Concept

Concepts

A mapping from input space to output space is called a concept and is denoted by $c : \mathcal{X} \to \mathcal{Y}$. The set of concepts is denoted by C.

Examples

- For binary classification an output space is $\mathcal{Y} = \{0, 1\}$, and any concept c can be identified by the set $A_c = \{x \in \mathcal{X} : c(x) = 1\}$ such that $c(x) = \mathbb{1}_{\{x \in A_c\}} = \mathbb{1}_{A_c}(x)$.
- The set of all triangles, rectangles, circles, and lines in the plane are all examples of concept classes.

Generalization

Possible to generalize to random concepts defined as $c : \Omega \times \mathcal{X} \to \mathcal{Y}$ defined over a probability space $(\Omega, \mathcal{F}, P)$ where the concept output distribution is a conditional distribution for every input x.

Hypothesis set

Hypothesis

The set of all possible candidate concepts that map features to predicted labels is called the hypothesis class and is denoted by $H \subseteq (\mathcal{Y}')^{\mathcal{X}}$.

Let $c : \mathcal{X} \to \mathcal{Y}$ be the true concept, and $h : \mathcal{X} \to \mathcal{Y}'$ be a hypothesis, then $y = c(x)$ is the label for an example x and $y' = h(x)$ is the predicted output for a hypothesis h.

Consistent

A consistent hypothesis set contains the concept to learn, and an inconsistent hypothesis set doesn't contain it. That is, H is consistent iff $c \in H$.

Sample

Assumption

All examples in $\mathcal{X}$ are independently and identically distributed (*i.i.d.*) with a fixed but unknown underlying distribution $D \in \mathcal{M}(\mathcal{X})$.

Sample

We have an example $x \in \mathcal{X}^m$ of size m generated *i.i.d.* according to the distribution D. For a concept $c : \mathcal{X} \to \mathcal{Y}$,we have a labeled sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ such that $z_i = (x_i, c(x_i))$.

Sample types

- Training: to train a learning algorithm.
- Validation: to tune the free parameters of the learning algorithm.
- Test sample: to evaluate the performance of the learning algorithm.

Loss function

Loss function

The loss function measures the difference or loss between the predicted and the true label, and is denoted by $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}_+$.

Hamming loss function

When $\mathcal{Y} = \mathcal{Y}'$ is discrete, the Hamming loss function $L_H(y, y') \triangleq \mathbb{1}_{\{y \neq y'\}}$ is bounded.

Euclidean loss function

When $\mathcal{Y} = \mathcal{Y}' \subseteq \mathbb{R}$, the Euclidean loss function $L_E(y, y') = (y - y')^2$ is unbounded for unbounded label sets.

Generalization risk

Generalization error

Given a hypothesis $h \in H$, the target concept $c \in C$, and an underlying distribution D from which an example $X \in \mathcal{X}$ is generated *i.i.d.* ,the generalization error/risk of hypothesis h is defined as $R(h) \triangleq \mathbb{E}L(c(X), h(X))$.

Hamming loss function $L_H(y, y') = \mathbb{1}_{\{y \neq y'\}}$

The generalization risk is
$R(h) = \mathbb{E}\mathbb{1}_{\{c(X) \neq h(X)\}} = P\{c(X) \neq h(X)\}$.

Risk minimizing hypothesis

If the sample distribution D and concept c are known, then the optimal hypothesis h^* that minimizes the generalization error is given by $h^* = \arg\min_{h \in H} R(h)$.

Model-free learning

Difference between ML and DE

The generalization error of a hypothesis is not directly accessible to the learner since both the distribution D and concept c are unknown. However, one can measure the *empirical error* of a hypothesis on the labeled sample z.

Empirical error

For a hypothesis $h \in H$, a target concept $c \in C$, and an labeled sample $z \in (\mathcal{X} \times \mathcal{Y})^m$ The *empirical error* is defined as

$$\hat{R}_z(h) = \frac{1}{m} \sum_{i=1}^{m} L(c(x_i), h(x_i)).$$

Supervised learning

Supervised learning

The *supervised learning* is the selection of a hypothesis $h_z \in H$ to minimize the empirical error with respect to loss function L. That is,

$$h_z = \arg\min_{h \in H} \hat{R}_z(h).$$

Empirical vs generalization risk

The empirical error of a hypothesis is the average error over the sample x, while the generalization error is the expected error based on the distribution D.

- $\mathbb{E}\hat{R}_z(h) = R(h)$
- $\hat{R}_z(h) \approx R(h)$ with high probability for large m

Hamming loss function $L_H(y, y') = \mathbb{1}_{\{y \neq y'\}}$

The empirical risk $\hat{R}_z(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{h(x_i) \neq c(x_i)\}}$ is an empirical average of m *i.i.d.* Bernoulli random variables.

Learning stages

- Randomly partition into training, validation, and test sample.
- Associate features to examples.
- Fix free learning parameters and pick a hypothesis set.
- Pick the hypothesis with the best performance on the validation sample.
- Predict labels of the test examples.
- Evaluate the algorithm using the test labels.

Consistent learning algorithm

- A learning algorithm is called *consistent* if there are no errors on the training data.
- A consistent algorithm may perform very poorly on test data if the learning class is highly complex. This is the difference between memorization and generalization.

Learning Problems

Types of problems

- Classification: assign a category to each item.
- Regression: assign a real value to each item.
- Ranking: assign order to items.
- Clustering: partition items into homogeneous regions.
- Dimensionality reduction: transform an initial representation of items to a low dimensional representation preserving some properties of the initial representation.

Classification

Classification

Assign a category to each item.

Applications

- Small number of categories: Document classification, text classification, and image classification
- large or unbounded categories:Optical character recognition, speech recognition

Regression

Regression

Assign a real value to each item.

Applications

- Prediction of stock values or other economic variables
- Prediction of physical processes such as temperature, humidity etc.

Ranking

Ranking

Assign order to items.

Applications

- Recommendation systems
- Web search
- Natural language processing

Clustering

Clustering

Partition items into homogeneous regions. Clustering is typically used for large unlabeled data sets.

Applications

- Community detection in large data sets.

Dimensionality reduction

Dimensionality reduction

Transform an initial representation of items to a low dimensional representation preserving some properties of the initial representation.

Applications

- Machine-aided compression
- Pre-processing of digital images

Learning scenarios

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Transductive inference
- Online learning
- Reinforcement learning
- Active learning

Unsupervised learning

Unsupervised learning

- The learner receives unlabeled examples for training and makes predictions for all unseen points.
- Difficult to quantitatively evaluate the performance of a learner.
- Clustering and dimensionality reduction are examples of unsupervised learning.

Semi-supervised learning

Semi-supervised learning

The learner receives a training sample consisting of both labeled and unlabeled data and makes predictions for all unseen points.

Transductive inference

Transductive inference

The learner receives a labeled training sample along with a set of unlabeled test points and makes predictions for only these test points.

Online learning

Online learning

- At each round, the learner receives an unlabeled training example, makes a prediction, receives the true label, and incurs a loss.
- The objective is to minimize the cumulative loss over all rounds.

Reinforcement learning

Reinforcement learning

- The learner actively interacts with the environment and receives an immediate reward for each action.
- The objective is to maximize reward over a course of actions and iterations with the environment.

Active learning

Active learning

- The learner adaptively/interactively collects training samples by querying an oracle for new samples.
- The goal is to achieve comparable performance to supervised learning with fewer samples.

Applications

Perception building

Perception engine building

Object identification from multiple camera and radar feeds, and perception building will increasingly rely on machine learning, either locally or in a distributed fashion.

Massive MIMO

- MIMO systems rely on spatial diversity, spatial multiplexing, and beamforming.
- 5G NR has included support for higher number of antennas at the base station, to help focus energy for achieving improved throughput and energy efficiency.
- Both the network and mobile devices need to implement more complex designs to coordinate MIMO operations.
- Such systems rely on dynamic feedback about signal quality, while controlling the 3D beamforming and path diversity to simultaneously communicate with multiple users in the same channel.

Parameter selection

Radio resource control

To orchestrate the transmission based on the network requirement or requests.

Radio link control

Decide modulation or power, based on signal quality

Network slicing

- Network slicing allows creating isolated virtualised slices on the same physical network infrastructure to support applications demanding different QoS levels.
- The network parameters and resource allocation needs to be managed for large number of slices.

Configuration of a large number of devices

Futuristic use cases

- Industry 4.0, digital twins, augmented reality, etc. are expected to be enabled.
- Cellular systems are expected to support massive scale device-to-device communications with stringent QoS requirements.
- Further, with targeted operations in mm-wave spectrum, large number of base stations will be required.
- Managing such large number of entities in such deployment through manual configurations or heuristics will be very difficult.
- Machine learning based approaches such as Reinforcement Learning are being considered for handling this scale.

Large scale viturализation

Complexity due to virtualization of network functions

- Many network functions will be disaggregated and provisioned through VMs/containers, and their placement and configuration will also need to be automated.
- AI-assisted Orchestration of virtualised functions is being considered, not just for providing data services, but for provisioning compute and AI as a service through cloud-edge-device continuum.

References

- Foundations of machine learning, Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar.
- Understanding machine learning, Shai Shalev-Shwartz and Shai Ben-David.