Lecture-01: Introduction

1 What is machine learning?

Machine learning is computational methods to improve performance or make predictions using *experience*. Experience is the past information available to the learner. Information maybe readily available as digitized human-labeled training sets, or can be obtained via interaction with environment.

Two main practical objectives of machine learning are:

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

The quality of machine learning algorithms is measured by

- 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

The success of prediction depends on size and quality of data instances. The theoretical learning guarantees depend on

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

Fundamental algorithmic and theoretical questions that arise are

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

Learning techniques are data-driven methods with relations to computer science, statistics, probability, and optimization.

2 Learning Problems

Learning problems can be broadly classified into following major classes.

1. Classification: assign a category to each item. Applications include document classification, text classification, image classification where number of categories are small. Other applications where there are large or unbounded categories are optical character recognition and speech recognition.
2. Regression: assign a real value to each item. Applications include prediction of stock values or other economic variables, or prediction of physical processes such as temperature, humidity etc.
3. Ranking: assign order to items. Applications include recommendation systems, web search, and natural language processing.
4. Clustering: partition items into homogeneous regions. Clustering is typically used for large unlabeled data sets. Applications include community detection in large data sets.
5. Dimensionality reduction: Transform an initial representation of items to a low dimensional representation preserving some properties of the initial representation. Applications include machine aided compression, preprocessing of digital images.

3 Definitions and terminology

Following are the terms we will use frequently.

1. Examples: items/data instances.
2. Features: set of data attributes often represented as a vector associated to an example. Feature extraction from examples is domain-specific task done by the experts, and is critical to the successful prediction. Typically represented by $x \in \mathcal{X}$. If an example has N attributes and all of them can be represented by real numbers, then the feature set $\mathcal{X} = \mathbb{R}^N$.
3. Labels: values of categories assigned to examples. Labels are discrete for classification and real-valued for regression. The set of labels is typically denoted by $\mathcal{Y}$.
4. Sample: collection of examples together with their labels is called a sample. There are three kinds of samples.
   (a) Training sample: examples/samples to train a learning algorithm.
   (b) Validation sample: samples to tune the free parameters of the learning algorithm.
   (c) Test sample: samples to evaluate the performance of the learning algorithm.
5. Loss function: measures the difference or loss between predicted and the true label. Set of predictions are denoted by $\mathcal{Y}'$ not necessarily equal to the set of labels $\mathcal{Y}$. Then the loss function $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}_+$ is not necessarily bounded. Two popular examples are:
   (a) when $\mathcal{Y} = \mathcal{Y}'$ and $L(y, y') = 1_{\{y \neq y'\}}$.
   (b) when $\mathcal{Y} = \mathcal{Y}' = I \subset \mathbb{R}$, and $L(y, y') = (y - y')^2$.
6. Hypothesis set: is the set $H \subseteq \mathcal{Y}^{\mathcal{X}}$ of functions mapping features to the set of labels.

Learning stages for a given sample (collection of labeled examples).

1. Randomly partition into training, validation, and test sample.
2. Associate features to examples.
3. Fix free learning parameters and pick a hypothesis.
4. Pick the hypothesis with best performance on validation sample.
5. Predict labels of the test examples.
6. Evaluated the algorithm using the test labels.

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.

4 Learning scenarios

1. Supervised learning: The learner receives a sample for training and validation, and makes prediction for all unseen points. This is common scenarios for classification, regression, and ranking.
2. 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.
3. Semi-supervised learning: The learner receives a training sample consisting of both labeled and unlabeled data, and make predictions for all unseen points.
4. Transductive inference: The learner receives a labeled training sample along with a set of unlabeled test points, and make predictions for only these test points.
5. 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.

6. 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.

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