Definitions and Classes of Machine Learning

Lecture # 2

Classes of Learning Problems

For different applications of ML, different techniques are used that include:

Classification: This assigns a category to each item. For example, document classification may assign items with categories such as politics, business, sports or weather. Image classification may assign items like landscape, portrait or animal. The number of categories is often small but they could be large in case of, for example, text classification or speech recognition.

Regression: Predict a real value for each item. Examples include prediction of stock values or variation of economic variables. The penalty for an incorrect prediction depends on the magnitude of the difference between the predicted and the true values This is in contrast with the classification where there is no notion of closeness between various categories.

Ranking: Orders items according to some criterion. The examples include returning web pages relevant to a search query.

Clustering: Partitions items into homogeneous regions. This is often used to analyze very large data sets. Examples include identification of communities within large groups of people (in the context of social network analysis).

Dimensionality Reduction: This transforms an initial representation of items into a lower dimensional representation of these items, while preserving some properties of the initial representation. Examples include processing digital images in computer vision tasks.

Applied Machine Learning

The main practical objective of ML is to generate accurate predictions of unseen items and of designing efficient and robust algorithms to produce these predictions, even for large-scale problems. The question is: which concept families can be learned and under what conditions? How well could these concepts be learned computationally?

ML Terminology

Here we present some basis definitions used in ML. These can be used as a reference in future lectures. To demonstrate the definitions, we use spam detection as an example to demonstrate the use of ML algorithms in practice. Spam detection is a problem of learning to automatically classify email messages as either spam or non-spam.

Examples: Items or instances of data used for learning or evaluation. In our spam example, this corresponds to the collection of email messages we will use for learning and testing.

Features: The set of attributes, often represented as a vector, associated to an example. In case of spam detection example, some relevant features could include the length of the message, the name of the sender, characteristics of the header, the presence of certain keywords in the body of the message.

Labels: Values of categories assigned to examples. In classification problems, examples are assigned different categories. For example, the SPAM and non-SPAM Categories in our binary classification problem are labels.

Training sample: Examples used to train a learning algorithm. In our spam example, the training sample consists of a set of email examples along with their associated labels- e.g. the keywords that lead to classification to a spam category of some emails.

Validation sample: Examples used to tune the parameters of a learning algorithm when working with labeled data. Learning algorithms have typically one or more free parameters and the validation sample is used to select appropriate values for these model parameters.

Test sample: Examples used to evaluate the performance of a learning algorithm. The test sample is separate from the training and validation data and is not made available at the learning stage. In the spam problem, the test sample consists of a collection of email examples for which the learning algorithm must predict labels based on features. These predictions are then compared with the labels of the test sample to measure the performance of the algorithm.

Loss Function: A function that measures the difference, or "loss", between a predicted label and a true label. Let the set of all labels be denoted as *Y* and the set of possible predictions as *Y'*, a loss function L is a mapping $L: Y \times Y' \rightarrow R$

Hypothesis: A set of functions mapping features (feature vectors) to the set of labels Y. In the spam example, these may be a set of functions mapping email features to *Y = {SPAM, non-SPAM}*. Hypothesis may be functions mapping features to a different set Y'. They could be linear functions mapping email feature vectors to real numbers interpreted as scores *(Y' = R)*, with higher score values more indicative of SPAM than lower ones.

ML Scenarios

There are different ML scenarios depending on the type of training data available to the learner, the order and method by which training data is received and the test data used to evaluate learning algorithm.

Supervised Learning: The learner receives a set of labeled examples as training data and makes predictions for all unseen points. This is the most common scenario associated with regression, ranking and classification problems. The SPAM detection problem is an instance of supervised learning.

Unsupervised Learning: The learner receives unlabeled training data and makes predictions for all unseen points. Since no labeled example is available, it is difficult to quantitatively evaluate the performance of a learner. Clustering and dimensionality reduction are examples of unsupervised learning.

Semi Supervised: The learner receives a training sample consisting of both labeled and unlabeled data and makes predictions for all the unseen points. Semi-supervised learning is common where unlabeled data is easily accessible but labels are expensive to obtain. The distribution of unlabeled data can help to achieve a better performance than in the supervised setting.