Machine Learning

João Gama
University of Porto, Portugal

André C P L F de Carvalho
University of São Paulo, Brazil

INTRODUCTION

Machine learning techniques have been successfully applied to several real world problems in areas as diverse as image analysis, Semantic Web, bioinformatics, text processing, natural language processing,telecommunications, finance, medical diagnosis, and so forth.

A particular application where machine learning plays a key role is data mining, where machine learning techniques have been extensively used for the extraction of association, clustering, prediction, diagnosis, and regression models.

This text presents our personal view of the main aspects, major tasks, frequently used algorithms, current research, and future directions of machine learning research. For such, it is organized as follows: Background information concerning machine learning is presented in the second section. The third section discusses different definitions for Machine Learning. Common tasks faced by Machine Learning Systems are described in the fourth section. Popular Machine Learning algorithms and the importance of the loss function are commented on in the fifth section. The sixth and seventh sections present the current trends and future research directions, respectively.

BACKGROUND

Machine learning can be seen as a subfield of artificial intelligence (Bratko, 1984) and is influenced by works on statistics (inference and pattern recognition [Duda & Hart, 1973; Fukunaga, 1990]), databases (analytical and multivariate databases [Berson & Smith, 1997]).

Machine learning is strongly linked to search, optimization, and statistics. Several models present optimization mechanisms, like support vector machines. Others are based on statistics inference, for instance, Bayesian classifiers.

Machine learning models have been extensively used in data mining. Data mining is concerned with the discovery of useful information in large databases. Very often, the observations need to be collected, selected, and preprocessed before machine learning techniques can be employed. It is important to mention that data mining relies not only on machine learning, but also on statistics, artificial intelligence, databases, and pattern recognition.

WHAT IS MACHINE LEARNING?

Informally speaking, the main goal of machine learning is to build a computational model from past experience of what has been observed. For such, machine learning studies the automated acquisition of domain knowledge looking for the improvement of systems performance as result of experience.

In the beginning of the 1980s, Michaslky, Carbonell, and Mitchell (1983) presented one of the first definitions of machine learning "Self-constructing or self-modifying representations of what is being experienced for possible future use" (p. 10).

The focus of this definition is on programs that modify themselves in response to feedback from their environment. This definition reflects the main research lines at that time: expert systems (Weiss & Kulikowski, 1991), automatic programming, and reinforcement learning (Sutton, 1998).

A more recent definition appears in (Hand, Mannila, & Smyth, 2001) "Analysis of observational data to find unsuspected relationships and to summarize the data in novel ways that is both understandable and useful for the data owner" (p. 1).

An even more recent definition is due to (Alpaydin, 2004), where machine learning is defined as "Programming computers to optimize a performance criterion using example data or past experience" (p. 3).

Clearly, the task here is much closer to a data analysis task, enlarging the range of practical applications, mainly industrial and commercial, where machine learning is frequently employed. In any case we can define machine learning as the acquisition of a useful (understandable) representation of a data set from its extensional representation.

MACHINE LEARNING TASKS

In a basic learning task, observations take the form of pairs. $\{\vec{x}, y\}$ The elements of the vector $\vec{x}$ are named *independent*

variables or *attributes* and the dependent variable $y=f(\vec{x})$ is an unknown function. The learning task is to obtain a predictive model or an approximation function $\hat{f}$ able to predict $\hat{y}$ for future observations of the independent variables $\vec{x}$ (Mitchell, 1997). In this framework, we can consider two different problems: *classification problems,* whenever y takes values in a finite set of unordered values (e.g., $y \in \{C_1,...,C_n\}$), and *regression problems,* when y takes values in a subset of R. In machine learning theory, these observations are assumed to be independent and generated at random, according to a stationary probability distribution. This task is referred as *predictive or supervised learning*, because the value of the target variable y in the observations or training set is known.[1] When there is no clear target variable in the training set, we have an unsupervised learning task.

Several areas of human activity can involve *supervised machine learning*: predicting the use of land based on satellite images; assigning credit to individuals on the basis of financial information; sorting letters on the basis of machine readable post codes; preliminary diagnosis of a patient's disease; and so forth. Several problems in industry, commerce, and science are *decision problems* and require the analysis of complex and extensive data. Most of these problems can be analyzed from a supervized machine learning perspective (Witten & Frank, 2005).

In contrast to predictive learning, *descriptive or unsupervised learning models* provide compact representations for the whole data or for the process generating the data. Examples of such descriptions include models for density estimation, clusters analysis and segmentation, and models describing relations between variables. This task includes data synopsis or signatures (Arasu & Manku, 2004), data visualization, and cluster analysis (Berthold & Hand, 1999; Duda, Hart, & Stork, 2001).

MACHINE LEARNING ALGORITHMS

We can formalize a machine learning problem as either a parameter optimization problem or a hypothesis search problem. In the former, examples are points in a multi-dimensional space associated with a metric. The goal is to minimize a loss function. Illustrative examples following this approach are *k-nearest neighbor* (Aha, 1997), discriminant functions (Duda et al., 2001), and neural networks (Ripley, 1996). In the later, a language used to represent generalizations of examples defines a search space of possible hypothesis. The learning algorithm performs a search in this space. The goal is to find the best hypothesis, a state in the search space that maximizes some objective function. Illustrative examples of this approach are decision trees (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1993), decision rules (Quinlan, 1993), and Bayesian networks (Neapolitan, 2003).

Each approach employs a different representation schema and explores different search strategies. A representation is the decision structure of a certain type (i.e., a decision tree, a set of discriminant equations, a table of conditional probabilities, etc.) used to generalize the examples. The representation used by a learning algorithm restricts the set of hypotheses considered by the algorithm. Some authors call it the *restricted hypotheses space bias* (Mitchell, 1990).

A *search strategy* is the set of methods and heuristics used to explore the search space defined by the set of possible representations. Associated with each search strategy is the *evaluation component.* This component is used to guide the search, by either preferring one hypothesis over others or by ranking the set of possible hypotheses. Both the search strategy and the evaluation component have preferences on the possible set of hypotheses. Such preference is also known in the machine learning literature as *preference bias*, because it imposes a certain preference order on the elements of the hypotheses space. A commonly used heuristic is *to prefer general hypotheses over specialized ones*. For example, if we consider decision trees, this corresponds to the preference of small trees to larger ones.[2] This kind of preference that minimizes the syntactic complexity of the hypotheses representation reduces the chance of the model overfitting the training observations. Overfitting occurs when the model is over-adjusted to the training data. Overfitting decreases the model generalization, that is, its capacity to correctly classify new observations.

There is a strong relation between overfitting and the *Occam's razor* (Blumer, Ehrenfeucht, Haussler, & Warmuth, 1990), which states, "The simpler of two competing hypotheses should always be preferred."

One argument in favor of simplicity is that there are fewer simple hypotheses than complex ones (based on combinatorial arguments). As such, if both fit the data, we should prefer the simpler hypothesis because it is less likely to be a statistical coincidence (Mitchell, 1997). Domingos' (1998a) work presents a thorough discussion about the interpretation of Occam's razor in the context of machine learning. Domingos concludes "if a model with low training-set error is found within a sufficiently small set of models, it is likely to also have low generalization error" (p. 37). Therefore a fully adequate model evaluation is only possible if the search process by which the models are obtained is also taken into account (Domingos, 1998b).

Another general heuristic claim is, "Examples that are near in the instance space correspond to similar concepts." This heuristic is fully exploited in the so-called *instance-based learning*, also known as *lazy learning,* where learning consists in memorizing previous observations, as in k-nearest neighbor and *case-based reasoning* (Aamodt & Plaza, 1994). The term *lazy learning* is used because instead of estimating the target function once for the entire instance space, these algorithms delay learning until they need to output a pre-