Machine Learning

Statistical Learning

Irving Gómez Méndez

Introduction

In various fields of science, technology and humanities, experts aim at predicting a phenomenon based on past observations or measurements.

- Meteorologists try to forecast the weather for the next days or weeks from the climatic conditions of the previous days.
- In medicine, clinical information is collected for diagnosing the condition of patients.

Regression and Classification

When the possible values taken by the response variable belong to a discrete set of values, the prediction of the response from observed predictor variables is known as **classification**:

- For example when a patient is diagnosed from a set of possible diseases based on its clinical information.

When the possible values of the response belong to a continuous space, then the prediction of its value from observed predictor variables is known as **regression**:

- For example when the temperature of the next days is predicted from previous climatic conditions.

Statistical Learning

Statistical methods for regression and classification have been used for centuries to help researchers and practitioners with these problems of prediction. However, **the increasingly speed at which data is generated and the variety of its type has surpassed the limits of standard statistical methods**.

Fortunately alongside the progress in data generation, new techniques and algorithms from the field of machine learning have been developed as powerful tools for the analysis of complex and large data.

When scientists learn about nature, the environment can be thought as a passive agent - apples drop, stars shine, and the rain falls without regard the needs of the scientists. We model such learning scenarios by postulating that data is generated by some random process.

The task of statistical models is to process such randomly generated examples toward drawing conclusions for the phenomenon from which these examples are picked.

Statistical learning uses the strengths and special abilities of computers to perform tasks that fall way beyond human capabilities. For example, the ability to scan and process huge databases allows the statistical learning program to detect meaningful patterns that are outside the scope of the scientists' perception and may have been missed by the human observer.

Phases of Learning

Machine learning can be defined as the study of systems that can learn from data without being explicitly programmed, this process typically has two phases:

1. **Training phase**: In this phase, a model is constructed from a training data set, where usually the response variable is known along with the predictor variables (labeled data), and both of them are used to improve the performance of the model. Hence, the model is trained by minimizing some loss function that measures the difference between the prediction made by the model and the true value of the response variable.

2. **Testing phase**: This phase is characterized by the use of a testing data set which has not made available in the training phase. During the testing phase the constructed model must predict the response variable of the observations in the testing data set. These predictions are then compared with the true value of the response to measure the performance of the algorithm.

We can distinguish different learning scenarios based on the types of data, the order and the method by which the training data and the testing data are received, as well as their use to evaluate the learning algorithms.

Learning Scenarios

- **Supervised learning**: It describes a scenario in which the training data set contains labeled observations. The model is then train to predict the label based on the predictor variables. In such cases we can think of the environment as a teacher that "supervises" the model by providing that extra information (the response variable). This is the most common scenario associated with classification, regression, and ranking problems.

- **Unsupervised learning**: In this case the model exclusively receives unlabeled training data, that is, observations without the value for the response variable, or even without the existence of such labels. The model processes the input data with the goal of coming up with some summary, or compressed version of that data. Since there is an unclear distinction between training and testing data, it can be difficult to evaluate the performance of the model. Clustering and dimensionality reduction are examples of unsupervised learning problems.

- **Semi-supervised learning**: The model receives a training sample consisting of both, data with the response variable (labeled data) and without it (unlabeled data), and makes predictions for all unseen points.
  Semi-supervised learning is common in settings where unlabeled data is easily accessible but labels are expensive to obtain. Various types of problems arising in applications, including classification, regression, or ranking tasks, can be framed as instances of semi-supervised learning. The hope is that the distribution of unlabeled data accessible to the model can help it achieve a better performance than in the supervised setting.

- **Online learning**: In contrast with the previous scenarios, the online scenario involves multiple rounds where training and testing phases are intermixed. In this case the model has to respond online, throughout the learning process, instead of engage the acquired knowledge only after having a chance to process large amount of data. At each round, the model receives a training point without the value of the response variable, makes a prediction, receives the true value, and incurs a loss. The objective in the online setting is to minimize the cumulative loss over all rounds or to minimize the regret, that is the difference of the cumulative loss incurred and that of the best model in hindsight. The model becomes an expert over time, but might have made costly mistakes in the process. In contrast, in the previous scenarios the model has large amount of training data to play with before having to output conclusions.

- **Reinforcement learning**: The training and testing phases are also intermixed in reinforcement learning. To collect information, the model actively interacts with the environment and in some cases affects it, and receives an immediate reward for each action. The objective of the model is to maximize the reward over a course of actions and iterations with the environment. Hence, the model is faced with the exploration versus exploitation dilemma, since it must choose between exploring unknown actions to gain more information versus exploiting the information already collected.