Lab-10: Introduction to Machine Learning

10.1 Objectives:

1. To understand machine learning

10.2 Machine Learning

Our imaginations have long been captivated by visions of machines that can learn and imitate human intelligence. Software programs that can acquire new knowledge and skills through experience are becoming increasingly common. We use such machine learning programs to discover new music that we might enjoy, and to find exactly the shoes we want to purchase online. Machine learning programs allow us to dictate commands to our smart phones, and allow our thermostats to set their own temperatures. Machine learning programs can decipher sloppily-written mailing addresses better than humans, and can guard credit cards from fraud more vigilantly. From investigating new medicines to estimating the page views for versions of a headline, machine learning software is becoming central to many industries. Machine learning has even encroached on activities that have long been considered uniquely human, such as writing the sports column recapping the Duke basketball team's loss to UNC.

10.3 Learning from experience

Machine learning systems are often described as learning from experience either with or without supervision from humans. In supervised learning problems, a program predicts an output for an input by learning from pairs of labeled inputs and outputs. That is, the program learns from examples of the "right answers". In unsupervised learning, a program does not learn from labeled data. Instead, it attempts to discover patterns in data. For example, assume that you have collected data describing the heights and weights of people. An example of an unsupervised learning problem is dividing the data points into groups. A program might produce groups that correspond to men and women, or children and adults. Now assume that the data is also labeled with the person's sex. An example of a supervised learning problem is to induce a rule for predicting whether a person is male or female based on his or her height and weight. We will discuss algorithms and examples of supervised and unsupervised learning in the following chapters.

10.4 Machine learning tasks

Two of the most common **supervised machine learning** tasks are **classification** and **regression**. In classification tasks, the program must learn to predict discrete values for one or more response variables from one or more features. That is, the program must predict the most probable category, class, or label for new observations.

Applications of classification include predicting whether a stock's price will rise or fall, or deciding whether a news article belongs to the politics or leisure sections. In regression problems, the program must predict the values of one more or continuous response variables from one or more features.

Examples of regression problems include predicting the sales revenue for a new product, or predicting the salary for a job based on its description. Like classification, regression problems require supervised learning.

A common **unsupervised learning task** is to discover groups of related observations, called **clusters**, within the dataset. This task, called clustering or cluster analysis, assigns observations into groups such that observations within a group are more similar to each other based on some similarity measure than they are to observations in other groups.

Clustering is often used to explore a dataset. For example, given a collection of movie reviews, a clustering algorithm might discover the sets of positive and negative reviews. The system will not be able to label the clusters as positive or negative; without supervision, it will only have knowledge that the grouped observations are similar to each other by some measure. A common application of clustering is discovering segments of customers within a market for a product. By understanding what attributes are common to particular groups of customers, marketers can decide what aspects of their campaigns to emphasize. Clustering is also used by internet radio services; given a collection of songs, a clustering algorithm might be able to group the songs according to their genres. Using different similarity measures, the same clustering algorithm might group the songs by their keys, or by the instruments they contain.

10.5 Training data, testing data, and validation data

A **training set** is a collection of observations. These observations comprise the experience that the algorithm uses to learn. In supervised learning problems, each observation consists of an observed response variable and features of one or more observed explanatory variables. The **test set** is a similar collection of observations. The test set is used to evaluate the performance of the model using some

performance metric. It is important that no observations from the training set are included in the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it. A program that generalizes well will be able to effectively perform a task with new data.

In addition to the training and test data, a third set of observations, called a **validation or hold-out set**, is sometimes required. The validation set is used to tune variables called hyperparameters that control how the algorithm learns from the training data. The program is still evaluated on the test set to provide an estimate of its performance in the real world. The validation set should not be used to estimate real-world performance because the program has been tuned to learn from the training data in a way that optimizes its score on the validation data; the program will not have this advantage in the real world.

TASK: Define the following terms;

1. Supervised machine learning
2. Unsupervised machine learning
3. Classification problem
4. Regression problem