Machine Learning Concepts

A Quick Overview

Machine Learning

Machine learning is a set of methods that can automatically detect patterns in data, and then use the **uncovered patterns to predict future data**, or to perform other kinds of decision making under **uncertainty**.

Uncertainty comes in different forms like which is the best choice of method, output or representation of the data. Probability theory can be used to model these kinds of uncertainty.

Types of Machine Learning

Machine learning is generally divided into two major types: **Supervised or predictive** and **Unsupervised or descriptive.**

Supervised Learning

In the predictive or supervised learning approach, the goal is to learn a mapping from inputs x to outputs y, given a labeled set of input-output pairs $D= \{x_i,y_i\}$, where D is called a training set.

Each x_i is represented by a n-dimensional vector of numbers called features /attributes /covariates. Similarly the output variable y_i is called response variable. The response variable can be nominal (continuous value) or categorical (class label).

Classification

When y_i is categorical, the learning problem is called **classification** or pattern recognition. Based on the number of classes, the classification task could be binary (2 classes) or multi class classification (more than 2).

Examples: Spam detection, Face recognition, Handwritten character recognition, etc

Regression

When y_i is nominal, the problem is called regression. Here, the task is to predict a continuous value, based on the training data.

Examples: stock market prediction, house rent calculation, etc

Unsupervised Learning

In descriptive or unsupervised learning, we are only given inputs. The goal is to find interesting patterns in the data. Since there are no prior class labels, there is no fixed evaluation metrics as misclassification rate in supervised learning.

Clustering

Here, the task is to cluster the data into groups in such a way that the data points within the cluster have high similarity (high intra cluster similarity) and low similarity between data points belonging to different clusters (high inter cluster dissimilarity). Similarity is commonly measured by computing the distance between data points.

Example:
Cluster of customers who have similar purchase tendency
Cluster of genes associated with common proteins/disease

Dimensionality reduction

Handling high dimensional data is a difficult task. Unsupervised learning can be used to project data into a low dimensional subspace which captures the essence of the data. These low dimensional features are referred to as latent factors. Methods like PCA, LDA and topic models use unsupervised learning to extract the latent representation.

Graph clustering

The correlation between data points can be represented as a graph which then yields answers to questions like which data points are most correlated. Identifying cliques (connected components) or hub nodes from the graph have applications in interpreting graph structured data.

[Optional: Mathematical foundation of graph clustering by Gilbert Strang, is available at https://www.youtube.com/watch?v=cxTmmasBiC8]

Matrix completion

Unsupervised learning techniques can be used to fill the missing values in the matrix, often represented by NaN. Matrix completion methods have great applications in

- Image imputation (inpainting): creating plausible images from low-resolution images or images with missing data Eg: images from astronomy, medical images etc.
- Collaborative filtering : automatic predictions about the interests of a user by collecting taste information from many users [Optional : https://en.wikipedia.org/wiki/Netflix_Prize]

[Optional Reading:
https://www.stat.cmu.edu/~ryantibs/convexopt-F16/scribes/acceleration-scribed.pdf]

Association mining

Association mining is a machine learning method for discovering interesting relations between variables in large databases. One of its commercial applications is market basket analysis, used for frequent itemset mining. (We will discuss more on association mining in MLDM course).

Basic Concepts

Parametric and non-parametric models

A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model. No matter how much data you throw at a parametric model, it won't change its mind about how many parameters it needs. They make use of the prior knowledge about the data.

For example, in linear regression, the learning is limited to fixed set parameters (slope and intercepts). In neural networks also the parameter set depends on the algorithm (not the dataset).

Algorithms that do not make any strong assumptions on the form of mapping functions are called non-parametric models. Nonparametric methods are good when you have a lot of data and no prior knowledge.

For example, in SVM, the kernel parameters are learned from pairwise distance of data points. So, there is no fixed set of parameters, but determined by the dataset.
Read More:
https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/

Overfitting

The goal of a good machine learning model is to generalize well from the training data to any data from the problem domain. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model.

Read more:
https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Model selection

When we have a variety of models with different efficiencies we need to compare their performance. In supervised learning, we can compare their performance by misclassification rate. In that case, the objective is to select the model which has lower misclassification error. The errors at training and testing (generalization error) are also important. A common procedure is to use a validation set along with training to measure the generalisation power of the model. Cross validation is a simple popular technique used among the ML community to systematically compare the model performance.

For example in K-fold cross validation, the training set is divided into k equal subsets. Then for each fold from 1..k, we train the model with (k-1) set, except k^{th} set, and k^{th} set is used for validation.

References

1. K. Murphy, Machine Learning: Probabilistic Perspective (Chapter-1)
2. Han et al, Data Mining: Concepts and Techniques (Chapter 1)