Key Terms Used in Machine Learning

By Sunil Ghimire - “**Be Unique, Be Identifiable, Be You**”

Before starting tutorials on machine learning, I came with the idea of providing a brief definition of key terms used in Machine Learning. These key terms will be regularly used in our coming lectures, tutorials, and workshops on machine learning and will also be used in further higher courses. So let’s start with the term Machine Learning:

A. MACHINE LEARNING

Machine learning is the study of computer algorithms that comprises algorithms and statistical models that allow computer programs to automatically improve through experience. It is the science of getting computers to act by feeding them data and letting them learn a few tricks on their own without being explicitly programmed.

B. CLASSIFICATION

Classification, a sub-category of supervised learning, is defined as the process of separating data into distinct categories or classes. These models are built by providing a labeled dataset and making the algorithm learn so that it can predict the class when new data is provided. The most popular classification algorithms are **Decision Tree, Support Vector Machine (SVM)**. I will study these algorithms in the coming articles.

C. REGRESSION

While classification deals with predicting discrete classes, regression is used in predicting continuous numerical valued classes. Regression is also falls under supervised learning generally used to answer “How much?” or “How many?”. Regressions create relationships and correlations between different types of data. **Linear Regression** is the most common regression algorithm.

D. CLUSTERING

Cluster is defined as groups of data points such that data points in a group will be similar or related to one another and different from the data points of another group. And the process is known as clustering. The goal of clustering is to determine the intrinsic grouping in a set of unlabelled data. Clustering is a form of unsupervised learning since it doesn’t require labeled data.

E. DIMENSIONALITY

The dimensionality of a data set is the number of attributes or features that the objects in the dataset have. In a particular dataset, if there are a number of attributes, then it can be difficult to analyze such a dataset which is known as the curse of dimensionality.

F. CURSE OF DIMENSIONALITY

Data analysis becomes difficult as the dimensionality of the data set increases. As dimensionality increases, the data becomes increasingly sparse in the space that it occupies.

1. For classification, there will not be enough data objects to allow the creation of a model that reliably assigns a class to all possible objects.
2. For clustering, the density and distance between points that are critical for clustering become less meaningful.

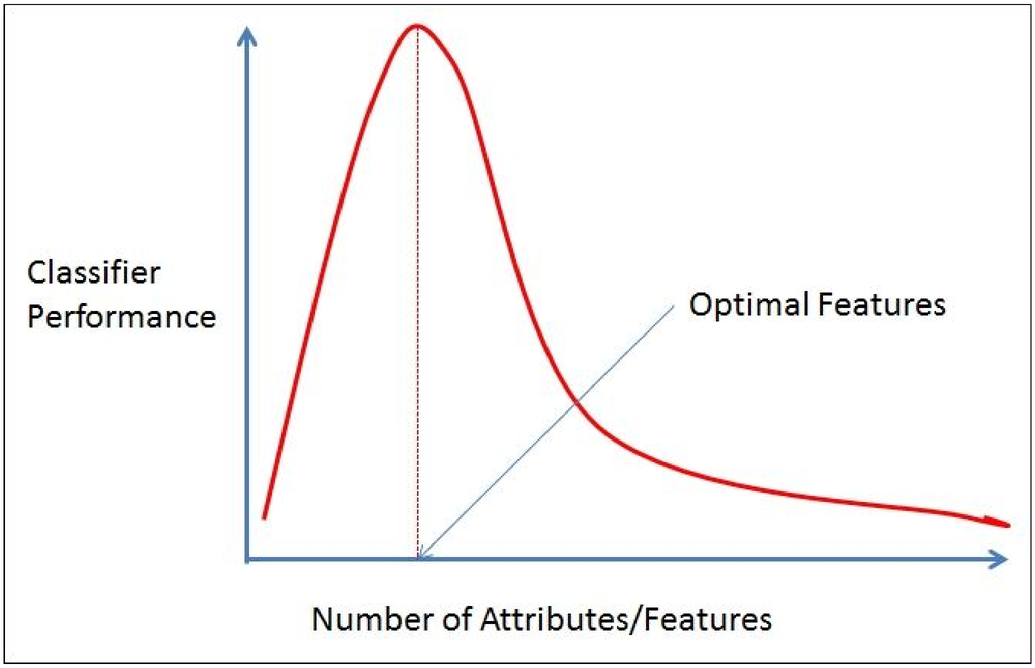

Figure 01: Curse of Dimensionality Graph

G. UNDERFITTING

A machine learning algorithm is said to have underfitting when it can’t capture the underlying trend of data. It means that our model doesn’t fit the data well enough. It usually happens when we have fewer data to build a model and also when we try to build a linear model with non-linear data when using a less complex model.

H. OVERFITTING

When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our dataset. The cause of overfitting is non-parametric and non-linear methods. We use cross-validation to reduce overfitting which allows you to tune hyperparameters with only your original training set. This allows you to keep your test set as a truly unseen dataset for selecting your final model.

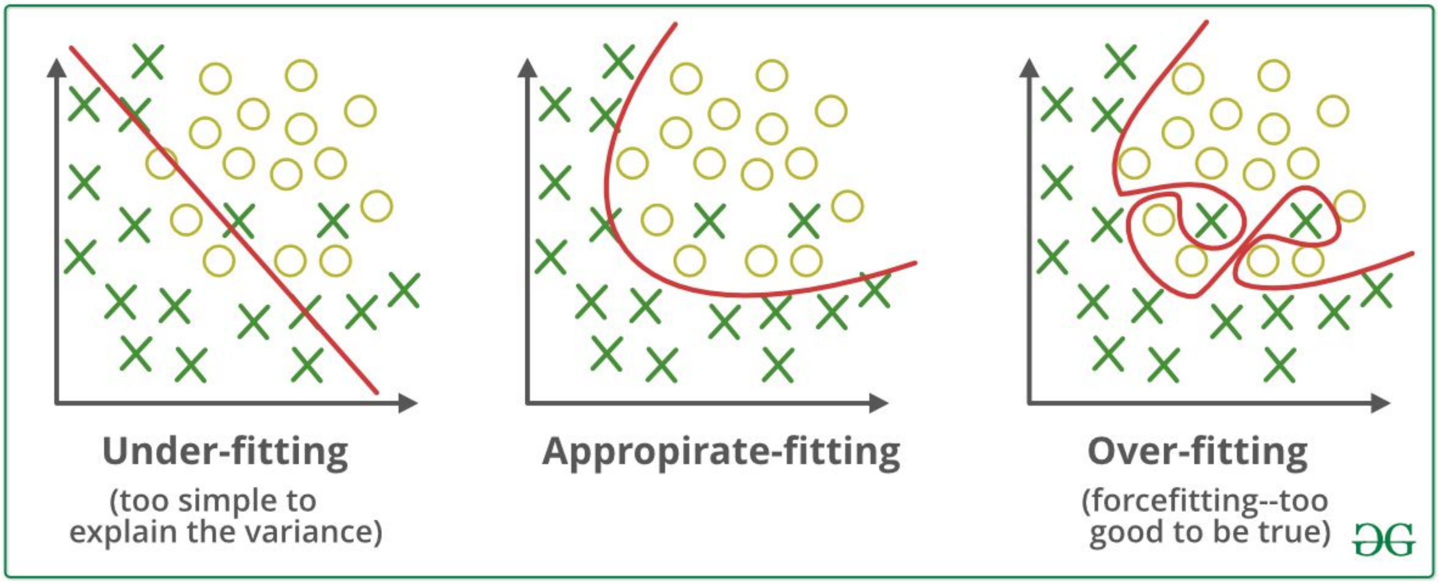
Under-fitting
(too simple to explain the variance)
Appropirate-fitting
Over-fitting
(forcefitting--too good to be true)

Figure 02: Under-fitting, Appropriate-fitting, and over-fitting

Note: There are many key terms used in machine learning other than described above. Other key terms will be discussed later in the article.

☺ Thanks for your time ☺

What do you think of this "**Key Terms Used in Machine Learning**"? (Appreciation, Suggestions, and Questions are highly appreciated).