Machine Learning Algorithms

Technology White Paper

Serge-Paul Carrasco

Abstract

The goal of Machine Learning (ML) is to develop algorithms for making predictions from an existing data set. First, you start by collecting the data and analyzing it. Second, you build a model for it and select a ML algorithm. Third, you evaluate and tune the model and the algorithm. Last, you deploy it and make live predictions.

There are three major approaches to ML:

- Regression finds a function to which a new instance belongs to
- Classification establishes the category (or class) to which a new instance belongs to
- Clustering groups/clusters the instances into categories (or classes)

Introduction

The goal of Machine Learning (ML) is to develop algorithms for making predictions from an existing data set. First, you start by collecting the data and analyzing it. Second, you build a model for it and select a ML algorithm. Third, you evaluate and tune the model and the algorithm. Last, you deploy it and make live predictions. The data instances xi of the model are represented as vectors xi = (xi1, xi2,...xin). The xi are called the features of the data (or the object). The goal of the ML algorithm is to learn from the training data in order to make predictions about future data instances.

There are three major approaches to ML:

- Regression finds a function to which a new instance belongs to
- Classification establishes the category (or class) to which a new instance belongs to
- Clustering groups/clusters the instances into categories (or classes)

Regression and classification problems are categorized as supervised learning. In supervised learning, we know what is the output for a given input since we have established a clear function between input vectors xi = (xi1, xi2,...xin) and output vectors yi = (yi1, yi2,...yin).

In a regression problem, we are trying to predict numerical results within a continuous output: yi= f(xi); f is continuous and $yi \in R$. In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input instances into discrete output categories: $yi \in \{\text{finite set}\}$. The yi are called label data and in supervised learning, data is said to be labeled.

Clustering is categorized as unsupervised learning. In unsupervised learning, we do not know what is the correct output for a given input. The output data is said to be unlabeled. However, we can predict the output for a new input but establishing some patterns into the data set.

In a clustering problem, we are trying to predict results based on the relationships among the existing instances in the data set.

Machine learning applications are growing and include:
Adaptive Websites, Affective computing, Bioinformatics, Brain-machine interfaces, Cheminformatics, Classifying DNA sequences, Computational advertising, Computational finance, Computer vision/Object recognition, Detecting credit card fraud, Game playing, Information retrieval, Machine perception, Medical diagnosis, Natural language processing, Recommender systems, Robot locomotion, Search engines, Sentiment analysis (or opinion mining), Sequence mining, Software engineering, Speech and handwriting recognition, Stock market analysis, Structural health monitoring, Syntactic pattern recognition...and many others.

Supervised learning algorithms

In a supervised learning, we are building a data set and can predict an output after having

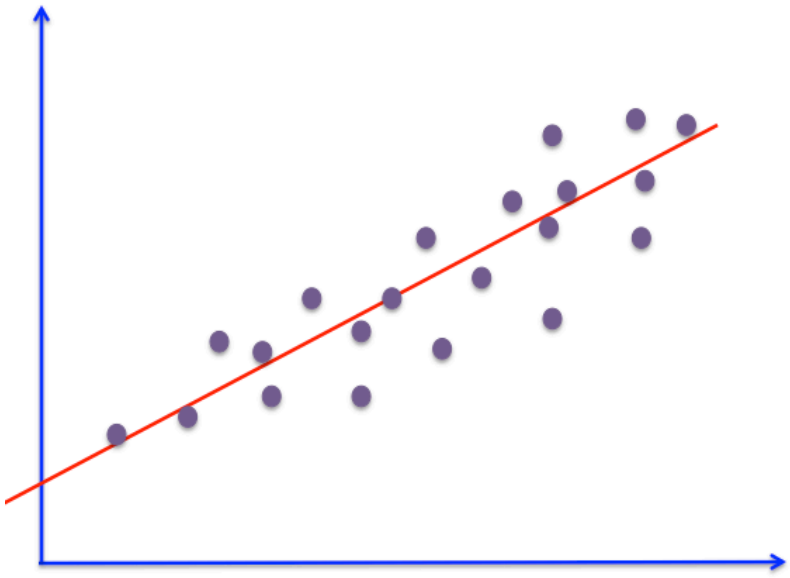
established the relationships that exist between a set of input and output training data.

In supervised learning, the training set is: {(x1,y1), (x2,y2), (x3,y3), (x4,y4),...(xn,yn)}.

Regression

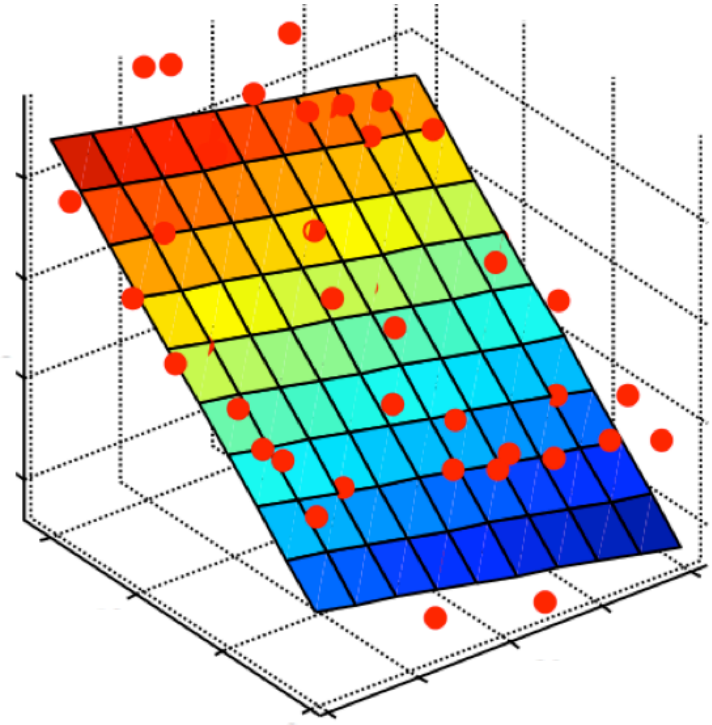
In a regression problem, we are trying to predict real-valued outputs. The input data can be mapped into a continuous linear function.
Wikipedia: Regression

Univariate linear regression: predicting an output value y = f (x) from an input value x (having one feature), having established f for a given of y and x data sets.
Example: What is the price of a house in San Francisco knowing its square feet?
Wikipedia: Linear regression



Linear Regression with one variable

Multivariate linear regression: predicting an output value y = f (x1, x2,...,xn) from multiple input value xi (having multiple features), having established a given of y and x data sets.
Example: What is the price of a house in San Francisco knowing its square feet, number of bedrooms/bathrooms, date of the construction...?
Wikipedia: Linear regression

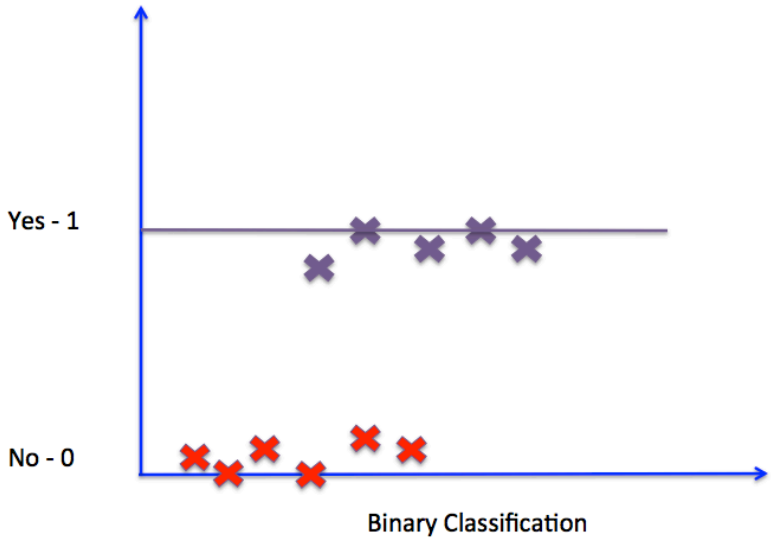


Linear Regression with multiple variables

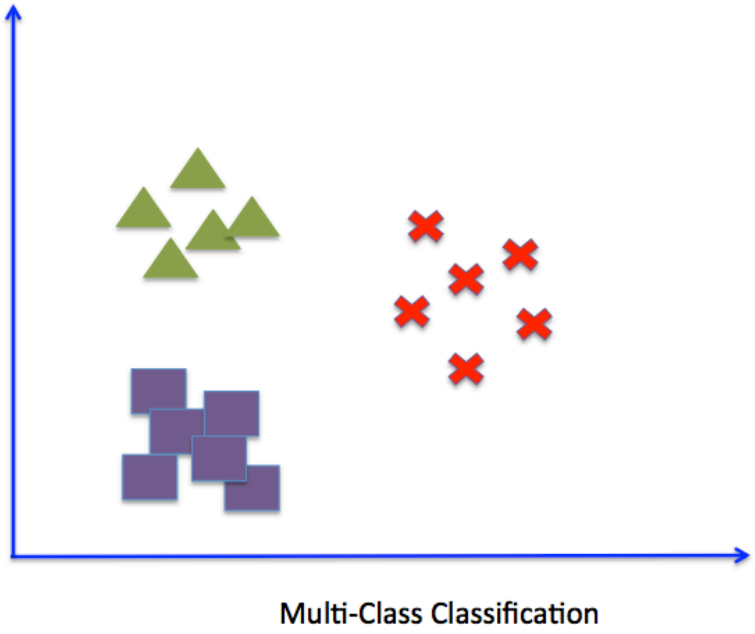
Classification

In a classification problem, we are trying to predict discrete-valued outputs. The input data can be mapped into discrete linear or non-linear categories.

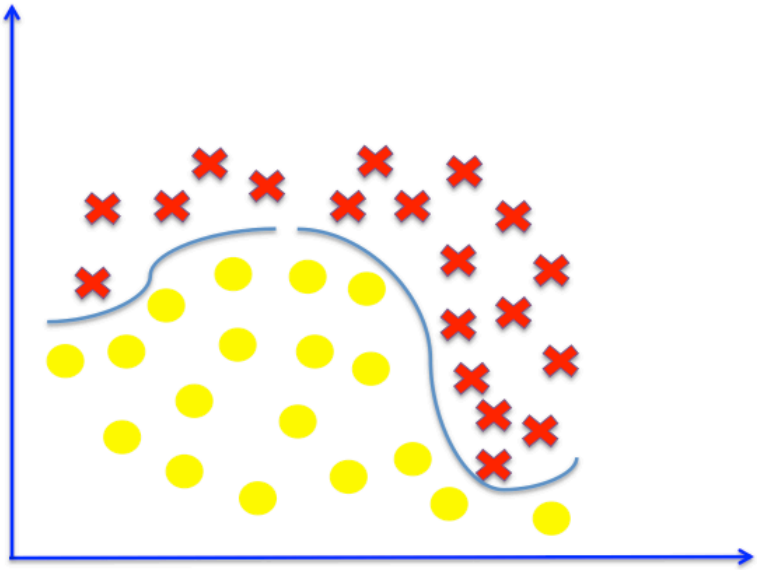
Linear classification (also called Logistic Regression):
Binary classification: predicting if the output y is either 0/no or 1/yes for a given input x, having established a given of y and x data sets.
Example: Is that e-mail is spam or not?
Wikipedia: Binary Classification



Multiclass classification: predicting if the output y is into more than two instances {0,1, 2,...n} for a given input x, having established a given of y and x data sets.
Example: Is that e-mail from my family, my friends, or for my personal hobbies or interests?
Wikipedia: Multiclass classification

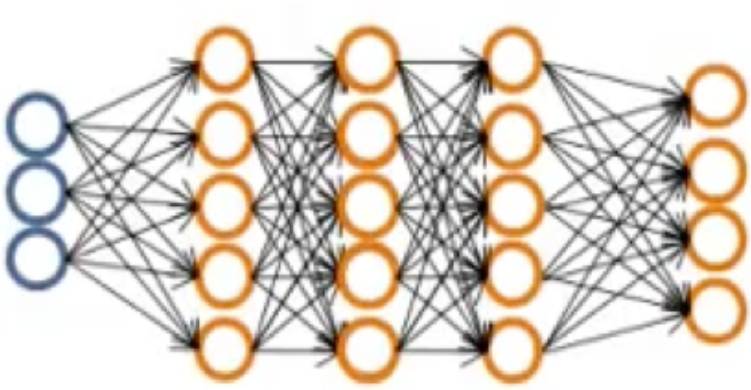


Non-linear classification



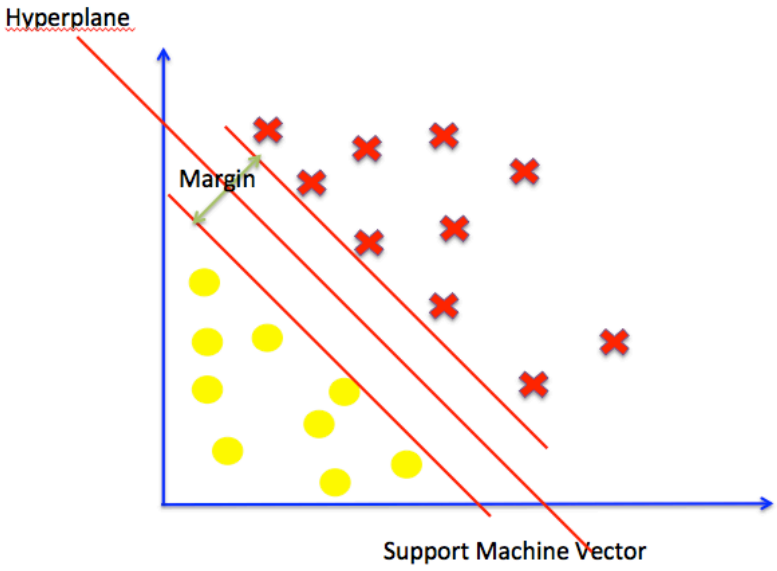
Non-Linear Classification

Neural networks classification (back propagation algorithm): predicting if the output y is into more than two instances {0,1,2,...n} for a given input x = {0,1, 2,...n}, having established a given of y and x data sets.
Example: Is that image a person, a car, or an animal?
Wikipedia: Backpropagation



Neural Networks

Support Vector Machine (SVM): SVM constructs an hyperplane from the data sets that can separate the various categories of data. SVM can be used both for linear and non-linear hypotheses. SVMs are probably the most efficient supervised learning algorithms.
Wikipedia: Support vector machine

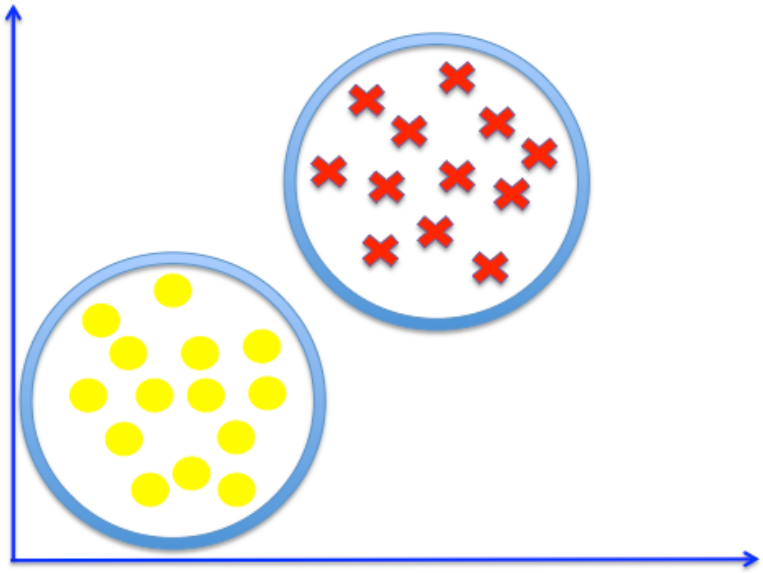


Unsupervised learning algorithms

In unsupervised learning, we are building a data set and attempt to establish how the data is structured and what are the patterns that the data seems to follow.

In unsupervised learning, the training set is: {x1, x2, x3, x4,...xn}. There is no label data: yi.

K-Means Clustering: aims to partition the data set into K clusters. Clusters are defined by a centroid that is determined in an iterative way to fit all the data into the cluster.
Wikipedia: K-means clustering



Clustering

Dimensionality Reduction: attempts to reduce the dimension of the features by for instance merging correlated features and visualizing them in 2D instead of 3D or 1D instead of 2D.

Anomaly Detection: the goal is to find out if new data appears to be an anomaly from previous data sets. To that end, we define a probably for the data to be not anomalous P(x) and a threshold to separate anomalous from normal data.

Recommender Systems

Produces a list of recommendations- through collaborative or content-based filtering or both. Collaborative filtering approaches build a model from a user's past behavior. Content-based filtering approaches utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties.
Wikipedia: Recommender system

Building a Machine Learning System

There are a lot of know-how in building a machine learning system that includes: bias/variance, regularization, evaluating the appropriate features, evaluating the learning algorithms, learning curves, error analysis, and ceiling analysis.

Regularization: attempts to address the challenge of overfitting that comes with both linear and logistic regression. Underfitting is when the prediction function maps poorly to the dataset. Overfitting is when the prediction function maps well to the existing dataset but does not predict well new data.

References

The best to learn more about the details of the algorithms is to take Stanford University Professor Andrew Ng's class on Machine Learning available on YouTube, iTunes and Coursera.

Other Materials:
Machine Learning algorithms on Wikipedia
Patricia Hoffman's Machine Learning class @ UC Santa Cruz Extension

