MACHINE LEARNING: METHODS, MODELS, USES, ADVANTAGES & DISADVANTAGES

Assignment Report – Predictive Analytics and Machine Learning

by

Aryan Tripathi

A Report

Submitted to

the

INSTITUTE OF ACTUARIAL AND QUANTITATIVE STUDIES

and

MR. SUJITH GOPINATHAN

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

in

Actuarial Science and Quantitative Finance

March 2022

Table of Contents

Machine Learning: An Overview

Machine learning is about extracting knowledge from data. It is a research field at the intersection of statistics, artificial intelligence, and computer science and is also known as predictive analytics or statistical learning. The application of machine learning methods has in recent years become ubiquitous in everyday life. From automatic recommendations of which movies to watch, to what food to order or which products to buy, to personalized online radio and recognizing your friends in your photos, many modern websites and devices have machine learning algorithms at their core. When you look at a complex website like Facebook, Amazon, or Netflix, it is very likely that every part of the site contains multiple machine learning models.

Outside of commercial applications, machine learning has had a tremendous influence on the way data-driven research is done today. The tools introduced in this book have been applied to diverse scientific problems such as understanding stars, finding distant planets, discovering new particles, analysing DNA sequences, and providing personalized cancer treatments.

Supervised vs Unsupervised Learning

Supervised Learning

Supervised machine learning is one of the most commonly used and successful types of machine learning. Supervised learning is used whenever we want to predict a certain outcome from a given input, and we have examples of input/output pairs. We build a machine learning model from these input/output pairs, which comprise our training set. Our goal is to make accurate predictions for new, never-before-seen data. Super- vised learning often requires human effort to build the training set, but afterward automates and often speeds up an otherwise laborious or infeasible task.

There are two major types of supervised machine learning problems, called classification and regression. In classification, the goal is to predict a class label, which is a choice from a predefined list of possibilities. Classification is sometimes separated into binary classification, which is the special case of distinguishing between exactly two classes, and multiclass classification, which is classification between more than two classes. You can think of binary classification as trying to answer a yes/no question. For regression tasks, the goal is to predict a continuous number, or a floating-point number in programming terms (or real number in mathematical terms). Predicting a person's annual income from their education, their age, and where they live is an example of a regression task. When predicting income, the predicted value is an amount, and can be any number in a given range.

Unsupervised Learning

The second family of machine learning algorithms that we will discuss is unsupervised learning algorithms. Unsupervised learning subsumes all kinds of machine learning where there is no known output, no teacher to instruct the learning algorithm. In unsupervised learning, the learning algorithm is just shown the input data and asked to extract knowledge from this data.

Unsupervised transformations of a dataset are algorithms that create a new representation of the data which might be easier for humans or other machine learning algorithms to

understand compared to the original representation of the data. Clustering algorithms, on the other hand, partition data into distinct groups of similar items.

The Difference
The main distinction between the two approaches is the use of labeled datasets. To put it simply, supervised learning uses labeled input and output data, while an unsupervised learning algorithm does not.
In supervised learning, the algorithm "learns" from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. While supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately. Unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabeled data. Note that they still require some human intervention for validating output variables.

In supervised learning, the goal is to predict outcomes for new data. You know up front the type of results to expect. With an unsupervised learning algorithm, the goal is to get insights from large volumes of new data. The machine learning itself determines what is different or interesting from the dataset.

Supervised learning models are ideal for spam detection, sentiment analysis, weather forecasting and pricing predictions, among other things. In contrast, unsupervised learning is a great fit for anomaly detection, recommendation engines, customer personas and medical imaging.

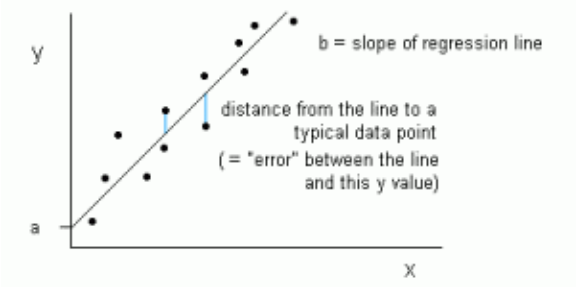
Supervised learning is a simple method for machine learning, typically calculated through the use of programs like R or Python. In unsupervised learning, you need powerful tools for working with large amounts of unclassified data. Unsupervised learning models are computationally complex because they need a large training set to produce intended outcomes.

A Deeper Look at … (Supervised Model)

Linear Regression is a supervised ML algorithm that helps find a suitable approximate linear fit to a collection of points.

At its core, linear regression is a linear approach to **identifying the relationship between two variables** with one of these values being a dependent value and the other being independent.

The idea behind this is to understand how a change in one variable impacts the other, resulting in a relationship that can be positive or negative.



Linear Regression is represented as a line in the form of y = a + bx.

This line is known as the regression line and represented by a linear equation Y= a *X + b.

In this equation:

- Y — Dependent Variable
- a — Slope
- X — Independent variable
- b — Intercept

This algorithm applied in cases where the predicted output is continuous and has a constant slope, such as:

- Estimating sales
- Assessing risk
- Weather data analysis
- Predictive Analytics
- Customer survey results analysis
- Optimizing product prices

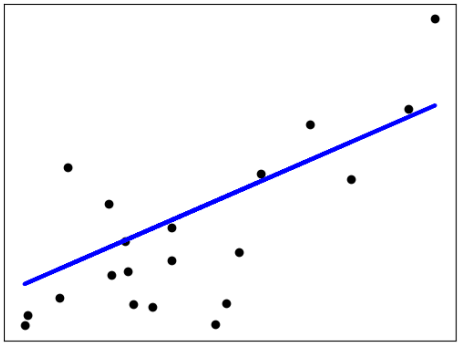
The following are a set of steps intended for regression in which the target value is expected to be a linear combination of the features. In mathematical notation, if y^ is the predicted value.

$\hat{y}(w,x)=w_0+w_1x_1+...+w_px_p$

Across the module, we designate the vector w=(w1,...,wp) as coef_ and w0 as intercept_.

Ordinary Least Squares:
LinearRegression fits a linear model with coefficients w=(w1,...,wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. Mathematically it solves a problem of the form:

$\min_w ||Xw-y||_2^2$



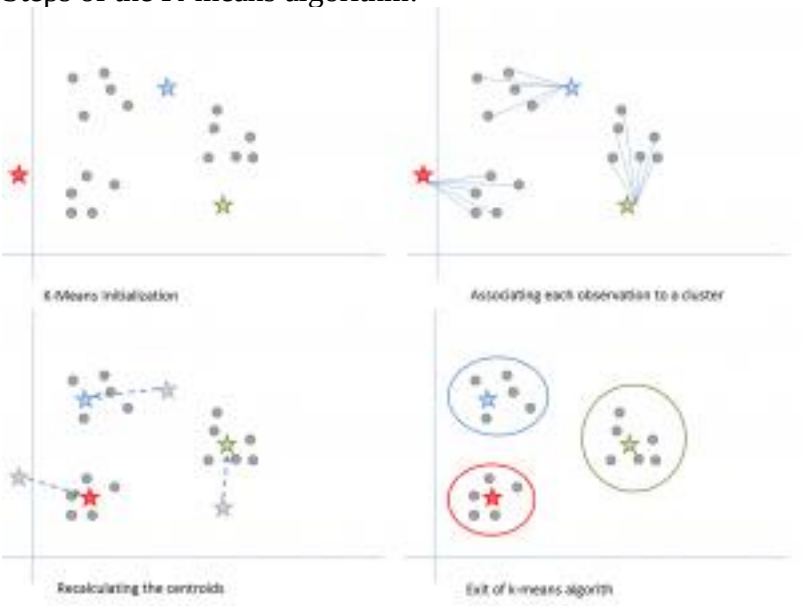
LinearRegression will take in its fit method arrays X, y and will store the coefficients w of the linear model in its coef_member: The coefficient estimates for Ordinary Least Squares rely on the independence of the features. When features are correlated and the columns of the design matrix X have an approximately linear dependence, the design matrix becomes close

to singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed target, producing a large variance. This situation of *multicollinearity* can arise, for example, when data are collected without an experimental design.

A Deeper Look at … (Unsupervised Model)

k-means clustering is an **iterative unsupervised learning algorithm** that partitions n observations into k clusters where each observation belongs to the nearest cluster mean.

Steps of the K-means algorithm:



In simpler terms, this algorithm aggregates a collection of data points based on their similarity. Its applications range from clustering similar and relevant web search results, in programming languages and libraries such as **Python**, **SciPy**, **Sci-Kit Learn**, and **data mining**.

The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.
Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

Differences in the Usage of Supervised and Unsupervised Machine Learning

An algorithm of supervised learning gets inputs from training data that should be labeled, allowing you to project unforeseen data results.

- In regression and classification problems, supervised learning is perfect, such as identifying to which group a news story refers to or forecasting the sales volume for a defined date in the future. Learning provides the information of the algorithm that can be used to produce new unknown data observations.
- To improve the performance of the algorithm, expertise also helps. The algorithms of supervised learning also help in solving real-world computations.

Applications Of Supervised Learning:

Bioinformatics

This is one of Supervised Learning's most popular implementations, and most of us use it in our everyday lives. Bioinformatics is the preservation of our humans' biological knowledge such as fingerprints, iris texture, earlobe, etc. Today's mobile phones are smart enough to learn our biological information and can then authenticate us to increase the system's security.

Speech Recognition

It's the kind of application where you express your voice to the algorithm and it will be able to spot you. Digital assistants like Google Assistant and Siri, which will awaken to the keyword only with your voice are the most well-known real-world devices.

Spam Detection

This program is used to block unreal or machine-based messages and e-mails. G-Mail has an algorithm that learns various keywords that may be incorrect. The One plus Messages App gives the user the task of letting the application know which keywords must be blocked and the keyword would block those messages from the app.

Object-Recognition For Vision

When you need to define something, this sort of software is being used. You have a large dataset that you use to teach your algorithm and a new instance can be recognized using this. A very well example is Raspberry Pi algorithms that detect objects

Conversely, ***unsupervised learning*** refers to inferring underlying patterns from an unlabeled dataset without any reference to labeled outcomes or predictions.

Some use cases for unsupervised learning — more specifically, clustering — include:

- Customer segmentation, or understanding different customer groups around which to build marketing or other business strategies.
- Genetics, for example clustering DNA patterns to analyze evolutionary biology.

- Recommender systems, which involve grouping together users with similar viewing patterns in order to recommend similar content.
- Anomaly detection, including fraud detection or detecting defective mechanical parts (i.e., predictive maintenance).

Image Classification and its Applications

Support vector machines (SVM) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. Support vector machines have their unique way of implementation as compared to other machine learning algorithms. They are extremely popular because of their ability to handle multiple continuous and categorical variables. Support Vector Machine model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by support vector machine so that the error can be minimized. The goal is to divide the datasets into classes to find a maximum marginal hyperplane. It builds a hyper-plane or a set of hyper-planes in a high dimensional space and good separation between the two classes is achieved by the hyperplane that has the largest distance to the nearest training data point of any class. The real power of this algorithm depends on the kernel function being used. The most commonly used kernels are linear kernel, gaussian kernel, and polynomial kernel. As a result, labor costs will be reduced.