Grokking Artificial intelligence Algorithms
by Rishal Hurbans

Section 8 Machine Learning

Machine learning can seem like a daunting concept to learn and apply, but with the right framing and understanding of the process and algorithms, it can be interesting and fun.

What is it?

Machine learning aims to find patterns in data for useful applications in the real world. We could spot the pattern in this small dataset, but machine learning spots them for us in large, complex datasets. Typically, data is represented in tables. The columns are referred to as features of the data, and the rows are referred to as examples. When we compare two features, the feature being measured is sometimes represented as y, and the features being changed are grouped as x.

Problems applicable to machine learning

Machine learning is useful only if you have data and have questions to ask that the data might answer. Different categories of machine learning algorithms use different approaches for different scenarios to answer different questions. These broad categories are supervised learning, unsupervised learning, and reinforcement learning.

First we have supervised learning which deals with looking at data, understanding the relationships and patterns to predict the results with other datasets of the same format. This can be seen in search autocomplete programs, as well as music applications that recommend new music to its users. There are 2 sub categories, regression and classification. Regression involves drawing a line through the set of data points to most closely fit the overall shape of the data. This can be used for applications such as trends between marketing initiatives and sales. As well as determine factors that affect something. On the other hand classification aims to predict categories of examples based on features.

Then there is unsupervised learning that involves finding underlying patterns in data that may be difficult to find by inspecting the data manually. This is useful for clustering data that has similar features and uncovering features that may be important in the data and not as obvious. For example on an e-commerce site, items may be clustered based on the consumers purchasing behavior.

Finally there is reinforcement learning which is inspired by behavioral psychology and operates by rewarding or punishing the algorithm based on its actions in an environment.That way the better ti behaves the more it gets rewarded.

Machine learning workflow

Machine learning is not all just about algorithms, rather it is often about the context of the data, its preparation, and the questions that are asked. These questions can be found by questioning the validity of the data being collected. As well as is the data

applicable in other environments. Understanding the context of the data is very important and data needs to be gathered from various systems and combined to be more effective. When we have a dataset it is important to look at the information since they may be coming from different places with different standards. If values are missing we can remove examples that have missing data. To fill in missing data we can use the median or mean to guess what it could be but this isn't fool proof since we may be overlooking certain correlations. We can also use the most frequent value since it has a better chance of being more accurate. We can also use neural networks or k-nearest neighbors. Or some algorithms can handle missing data without any preparation. Some data may also be ambiguous such as data being in 2 different measurements. It helps to standardize the data to avoid any confusion. When values are not numbers and instead are strings such as good or fair we can represent them using numerical values. With one-hot encoding we can think of them as switches where all are off except one. With label coding we represent each category as a number between 0 and the number of categories. However, this approach is really only for ratings. Regression can be used to predict a continuous value, where continuous means the value can be any number in a range. When drawing the line of regression we want it to pass through the intersection of the mean of x and the mean of y. To see how accurate the line is we can use the method of least squares whichs aims to create a line that minimizes the distance between the line and among all the points in the dataset. We want to find a line thats the closest to the data. We find the differences between the actual data values and the predicted data values, differences will vary with some being large, small, positive, and negative. We can square the differences and sum them. This will take into consideration all differences for all data points. Once the line of regression is calculated we can make predictions for other values. We can measure the performance of the line with new examples that we know the actual price of. Training and testing data are usually split 80/20. When training a model on data the models will often not perform as well as desired. To improve the model we can collect more data, prepare the data differently, choose different features in the data, use a different algorithm, or you may be dealing with false-positive tests. False-positive tests usually occur from overfitting which is when the model is too closely aligned with the training data and is not flexible for dealing with new data with more variance. If the model did not show anything useful it may be necessary to ask a different question

Classification with decision trees

Classification may also deal with discrete values that are categorical features such as color/clarity. Decision trees are structures that describe a series of decisions that are made to find a solution to a problem. FOr decision tree learning there are many algorithms one of which is CART Classification and Regression Tree. For the decision tree learning life cycle when we build a tree we test all possible questions to determine which one is the best question to ask at a specific point in the decision tree. To test a

question, we use the concept of entropy, or the measurement of uncertainty of a dataset.

Important Figures:

Machine learning categories

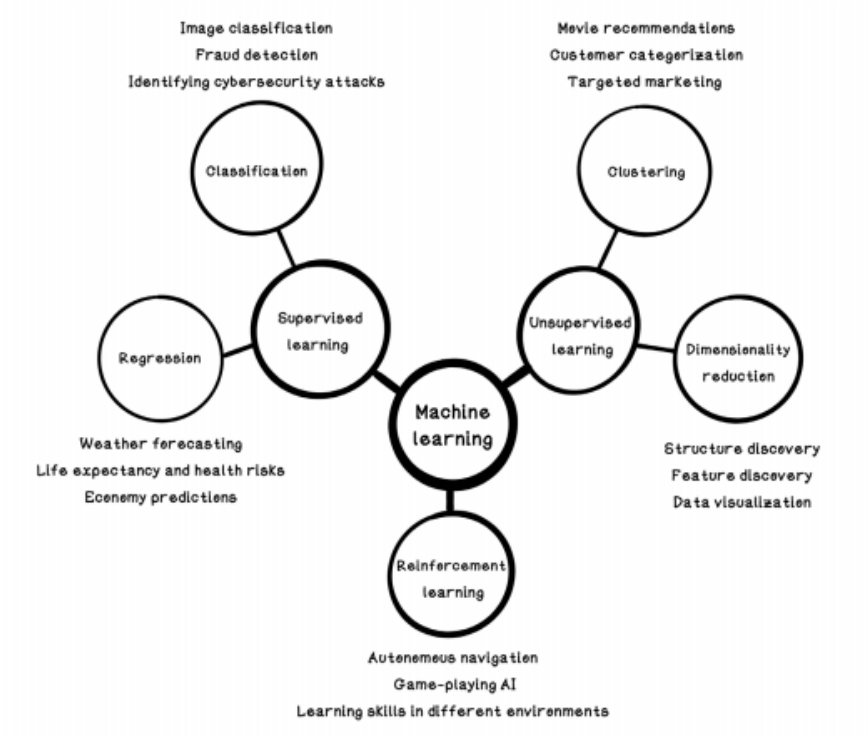
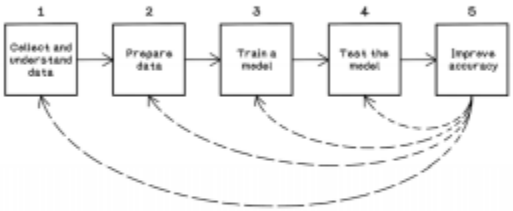


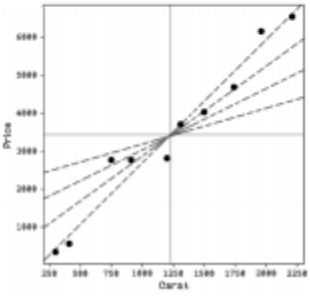
Figure 8.3 Categorization of machine learning and uses

Machine learning is more about context, understanding data, and asking the right questions than algorithms.



The life cycle of ML projects is iterative and experimental.

Linear regression involves finding the best line to fit the data, which means minimizing the error to each data point.



Possible regression lines

Decision trees split data using questions until the dataset is perfectly split into categories. The key concept is reducing uncertainty in the dataset.