Machine Learning through the lens of Probability Theory

Mike Medved, Benny Chen, Daniel Paliulis

November 14th, 2022

1 Introduction

Machine learning is an application of probability and algorithms used to train computers to learn from large data sets. It is currently used in almost every field, such as finance, healthcare, and public safety to name a few. Thus, with it's almost limitless applicability, machine learning (and it's various subsets) is quickly becoming one of the most important technological innovations of the 21st century.

In this short project, we will discuss both the mathematics required to understand how the basis of machine learning training works, as well as a few real life examples of how machine learning is currently being used in the real world.

2 Preliminaries

In order to fully understand the basics behind machine learning, we will first need to understand the basics of both linear regressions, as well as probability theory.

2.1 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is used to predict the value of a dependent variable based on the value of the independent variable(s). In the case of machine learning, the dependent variable is the output of the model, and the independent variables are the inputs to the model. The goal of linear regression is to find the best fit line for the data, which is represented by the equation:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

where $\hat{y}$ is the predicted value, θ_0 is the y-intercept, θ_1 is the slope of the line, and $x_1, x_2, ..., x_n$ are the independent variables. The goal of linear regression is to find the values of $\theta_0, \theta_1, ..., \theta_n$ that minimize the error between the predicted value and the actual value. This is done by minimizing the sum of the squared errors, which is represented by the equation:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

We can measure the error of the model by using the root mean squared error (RMSE), which is the square root of the mean of the squared errors. The RMSE is represented by the equation:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

2.2 Probability Theory

In order to be able to understand the mathematics behind machine learning, we will need to understand the basics of probability theory. We will start by defining a random variable, which is a variable whose value is the outcome of a random phenomenon.

Definition 1. A random variable is a variable whose value is the outcome of a random phenomenon.

For example, if we were to roll a die, the random variable X would be the number that is rolled. The random variable would be discrete, which means that it can only take on a finite number of values, since the only possible values, otherwise known as the Image of X, $Im\ X = \{1, 2, 3, 4, 5, 6\}$.

Definition 2. A random variable is discrete if it can only take on a finite number of values, likewise, a random variable is continuous if it takes numbers in an interval.

This is an important distinction to make when evaluating training data, as each training parameter is represented as a random variable. For example, if you are taking in the number of bedrooms, square footage, and age of a house, each of these values would be represented as a random variable. The number of bedrooms would be discrete, since it can only take on a finite number of values, whereas the square footage and age would be continuous, since they can take on any value in an interval.

Definition 3. A probability distribution (pdf) is a function that gives the probability that a random variable is equal to some value.

$$P(X = x) = \text{Probability that } X \text{ takes the value } x$$

This is important in the context of a machine learning model as it allows us to determine the probability that a given input will result in a given output. For example, if we were to train a model to predict the price of a house based on the number of bedrooms, square footage, and age, we would need to know the probability that a house with 3 bedrooms, 2000 square feet, and 20 years old would sell for \$200,000. This is represented by the equation:

$$P(Price = 200000 \mid Bedrooms = 3, Square\ Footage = 2000, Age = 20)$$

Definition 4. A cumulative distribution function (cdf) is a function that gives the probability that a random variable is less than or equal to some value.

$$P(X \leq x) = \text{Probability that } X \text{ is less than or equal to } x$$

Similarly to the pdf, this is useful in the context of a machine learning model as it allows us to determine the probability relative to certain inputs. However, instead of determining the probability for an exact input, we can determine the probability given an input is less than a certain threshold.

3 Walkthrough

Let us walk through an example of how a machine learning model is setup mathematically.

Example. Let us say we are trying to train a model to predict the price of a house based on the number of bedrooms, square footage, and age. We will assume that the price of a house is a linear function of the number of bedrooms, square footage, and age, and that the price is normally distributed. We will also assume that the number of bedrooms, square footage, and age are all independent of each other.

Firstly, we must setup a linear regression that will model the above problem, and for which we will attempt to maximize the parameters $\theta_0, ..., \theta_3$. We will assume that the price of a house is represented by the equation:

$$\begin{aligned}\text{Price} = \theta_3 &+ \textit{bedrooms} \\ &+ \theta_1 \cdot \textit{age} \\ &+ \theta_2 \cdot \textit{sqft} \\ &+ \theta_0 + \epsilon\end{aligned}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the error term, which is assumed to be normally distributed. We will also assume that the error term is independent of the number of bedrooms, square footage, and age.

Now that we have the linear regression setup for the above scenario, we can move forward with designing the training data. We will assume that we have a dataset of n houses, and that each house has a price, number of bedrooms, square footage, and age. We will also assume that the price of each house is normally distributed, and that the number of bedrooms, square footage, and age are all independent of each other.

Since each training data point is essentially an experiment to minimize the error function, we can represent the pdf, P, of given training data's accuracy given $\theta_0 ... \theta_3$ as inputs. The following equation represents the inputs $\theta_0 ... \theta_3$ for which the probability of the ith training result, y_i approaches a reasonable margin of error of the known output of the test case Y_i:

$$\theta_0, \theta_1, \theta_2, \theta_3 = \max_{\theta_0, \theta_1, \theta_2, \theta_3} P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n \mid \theta_0, \theta_1, \theta_2, \theta_3)$$

We can simplify this equation by vectorizing $\theta_0 ... \theta_3 \to \hat{\theta}^{MLE}$, yielding the following simpler equation:

$$\hat{\theta}_{MLE} = \max_{\theta} P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n \mid \theta)$$

Now, since we know that the training set is independently and identically distributed (IID), we can use the properties of independent random variables in order to simplify, and further transform the above equation:

$$\begin{aligned}\hat{\theta}_{MLE} &\Rightarrow \max_{\theta} \left[P(Y_1 = y_1 \mid \theta) \cdot P(Y_2 = y_2 \mid \theta) \cdot ... \cdot P(Y_n = y_n \mid \theta)\right] \\ \hat{\theta}_{MLE} &\Rightarrow \max_{\theta} \left[f_y(y_1 \mid \theta) \cdot f_{y_2}(y_2 \mid \theta) \cdot ... \cdot f_{y_n}(y_n \mid \theta)\right] \\ \hat{\theta}_{MLE} &\Rightarrow \max_{\theta} \left[\prod_{i=1}^{n} f_{y_i}(y_i \mid \theta)\right]\end{aligned}$$

For a linear regression, we know that $f_y = \mathcal{N}(\hat{y_i}, \sigma^2)$, thus, we can plug this into the above equation:

$$\hat{\theta}_{MLE} = \max_{\theta} \left[(-1) \cdot \sum_{i=1}^{n} (y_i - \hat{y_i})^2\right]$$

The above equation models the residual sum of squares (RSS) error function, which is the error function that is minimized in order to train a linear regression model.

4 Real Life Examples

Machine learning is utilized throughout many different fields, and is used to solve a wide variety of problems. In this section, we will discuss a few real life examples of how machine learning is being used in the real world.

4.1 Finance

A notable example of how machine learning is being used in the real world is in the field of finance. In the past, financial institutions would use a combination of human intuition and algorithms to make decisions on how to invest their money. However, with the rise of machine learning, financial institutions are now able to use machine learning algorithms to make these decisions for them. This has allowed financial institutions to make more accurate predictions, and has also allowed them to make more accurate predictions in a shorter amount of time.

Specifically, stock trading is one of the most common applications of machine learning in the field of finance. With the various parameters at the disposal of large financial institutions, such as trading volume, price, indicators (EMA, MACD, etc.) news, and public sentiment, institutions have been able to scale machine learning models to make predictions on the future price of a stock.

4.2 Healthcare

One notable example of how machine learning has been used to benefit the healthcare industry is in the field of cancer research. In 2018, researchers at the University of California, San Francisco, developed a machine learning algorithm that was able to predict the likelihood of a patient developing cancer based on their medical history. The algorithm was able to predict the likelihood of a patient developing cancer with 90% accuracy, which is significantly higher than the 70% accuracy of the current standard of care.

This is a significant improvement, as it allows doctors to better determine which patients are at a higher risk of developing cancer, and thus, which patients should be screened more frequently.

4.3 Public Safety

A very important example of where machine learning is being applied right now is in the field of public safety. Recently, companies have been developing visual classification programs to help detect weaponry and alert personnel. This can be used to prevent tragedies such as mass shootings, since the program can trigger a lockdown and notify police before any shots can even be fired.