Why Machine Learning

This lesson will focus on why we need machine learning models for predictions.

WE'LL COVER THE FOLLOWING

Issues with Regression

In the previous chapter, we learned how to use linear and logistic regression models for making predictions from data. But there are some issues with the regression framework. We will look at these issues one by one.

Non-linear relationships

By design, linear regression explores linear relationships between the dependent and independent variables. It assumes that there is a straight-line relationship between the variables and tries to find the line that best fits the data. Sometimes, it is not the case that variables follow a linear relationship. For instance, the relationship between age and income is not linear. Income rises exponentially during the early years and then grows almost linearly at the later stages.

Linear Regression parameters converge on the mean of the predicted variable

Linear regression finds the mean of the dependent variable since the error at the mean is relatively less to all points as compared to some other value. That is why in our **tips** example, the parameter chosen was close to the mean of the percentage tip. However, looking at extremes is also important.

Sensitivity to outliers

Linear regression is sensitive to outliers since it looks at the mean of the data. The best fit line can change direction to try to fit outliers.

Independent data assumption

Linear regression assumes that there is no significant relationship between the dependent variables. However, that is not always the case. Correlated independent variables affect the performance of linear regression models. This problem is also known as **multi collinearity** in statistics. Although there are ways to handle this, the performance of linear regression is not satisfactory when the dependent variables have relationships among themselves.

Because of these issues present in typical data, the performance of linear regression is often not very good. Linear regression works best when there are linear relationships between the dependent and independent variables, and the data contains no outliers. Therefore, we need another framework of predictive models that perform better. This is where *machine learning* comes in.

What is Machine Learning?

Machine Learning is the branch of computer science that deals with algorithms and systems performing specific tasks using patterns and inference, rather than explicitly programmed instructions. One use case of machine learning is predictive models. Most machine learning tasks can be categorized in the following two types:

- Supervised Learning
- Unsupervised Learning

Supervised Learning

Supervised Learning

In **supervised learning**, we make predictive models using data that has *labels*. When we have the target variable available in our data, we call the values of target variable **labels**. Until now, when we made linear regression models, we had labels available. Some supervised machine learning algorithms to make predictive models are:

- Decision Trees
- Neural Networks
- Random Forests
- Support Vector Machines (SVM)

We will look at some of these later in this chapter.

Unsupervised Learning

Unsupervised learning is when we use unlabeled data to allow a model to learn relationships between data observations and pick up on underlying patterns. Most data in the world is unlabeled, which makes unsupervised learning a very useful method of machine learning. The most common algorithms for unsupervised learning are *clustering algorithms*. We will look at some of these later in this chapter.

Machine Learning vs. Artificial Intelligence vs. Data Science

People often use the terms *machine learning*, *artificial intelligence*, and *data science* interchangeably. In reality, machine learning is a subset of artificial intelligence and overlaps heavily with data science. Artificial intelligence deals with any technique that allows machines to display *intelligence*, similar to humans. Machine learning is one of the main techniques used to create artificial intelligence, but other non-ML techniques are also widely used in AI.

On the other hand, data science deals with gathering insights from datasets. Traditionally, data scientists have used statistical methods, such as regression, for gathering these insights. However, as machine learning continues to grow, it has also penetrated into the field of data science.

In the next lesson, we will look at the Machine Learning pipeline.