UNIVERSIDAD POLITÉCNICA DE YUCATÁN

MACHINE LEARNING

FUNDAMENTAL CONCEPTS OF MACHINE LEARNING

KAREN VALERIA VILLANUEVA NOVELO
IRC9B
2009146

Fundamental Concepts of Machine Learning

Concept and characteristics of supervised and unsupervised learning

Supervised machine learning requires labelled input and output data during the training phase of the machine learning model lifecycle. This training data is often labelled by a data scientist in the preparation phase, before being used to train and test the model. Once the model has learned the relationship between the input and output data, it can be used to classify new and unseen datasets and predict outcomes.

The reason it is called supervised machine learning is because at least part of this approach requires human oversight. The vast majority of available data is unlabelled, raw data. Human interaction is generally required to accurately label data ready for supervised learning. Naturally, this can be a resource intensive process, as large arrays of accurately labelled training data is needed.

Supervised machine learning is used to classify unseen data into established categories and forecast trends and future change as a predictive model. A model developed through supervised machine learning will learn to recognise objects and the features that classify them. Predictive models are also often trained with supervised machine learning techniques. By learning patterns between input and output data, supervised machine learning models can predict outcomes from new and unseen data. This could be in forecasting changes in house prices or customer purchase trends.

Supervised machine learning is often used for:

- Classifying different file types such as images, documents, or written words.
- Forecasting future trends and outcomes through learning patterns in training data.

Unsupervised machine learning is the training of models on raw and unlabelled training data. It is often used to identify patterns and trends in raw datasets, or to cluster similar data into a specific number of groups. It's also often an approach used in the early exploratory phase to better understand the datasets.

As the name suggests, unsupervised machine learning is more of a hands-off approach compared to supervised machine learning. A human will set model hyperparameters such as the number of cluster points, but the model will process huge arrays of data effectively and without human oversight. Unsupervised machine learning is therefore suited to answer questions about unseen trends and relationships within data itself. But because of less human oversight, extra consideration should be made for the explainability of unsupervised machine learning.

The vast majority of available data is unlabelled, raw data. By grouping data along similar features or analysing datasets for underlying patterns, unsupervised learning is a powerful tool used to gain insight from this data. In contrast, supervised machine learning can be resource intensive because of the need for labelled data.

Unsupervised machine learning is mainly used to:

- Cluster datasets on similarities between features or segment data.
- Understand relationship between different data point such as automated music recommendations.

- Perform initial data análisis.

[1] “Supervised vs Unsupervised Learning Explained,” Seldon, Oct. 16, 2021. https://www.seldon.io/supervised-vs-unsupervised-learning-explained

Probabilistic model

Probabilistic Models are one of the most important segments in Machine Learning, which is based on the application of statistical codes to data analysis. This dates back to one of the first approaches of machine learning and continues to be widely used today. Unobserved variables are seen as stochastic in probabilistic models, and interdependence between variables is recorded in a joint probability distribution. It provides a foundation for embracing learning for what it is. The probabilistic framework outlines the approach for representing and deploying model reservations. In scientific data analysis, predictions play a dominating role. Their contribution is also critical in machine learning, cognitive computing, automation, and artificial intelligence.

These probabilistic models have many admirable characteristics and are quite useful in statistical analysis. They make it quite simple to reason about the inconsistencies present across most data. In fact, they may be built hierarchically to create complicated models from basic elements. One of the main reasons why probabilistic modeling is so popular nowadays is that it provides natural protection against overfitting and allows for completely coherent inferences over complex forms from data.

[2] Simplilearn, “What Are Probabilistic Models in Machine Learning? | Simplilearn,” Simplilearn.com, Oct. 18, 2022. https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-are-probabilistic-models

Differences between supervised and unsupervised learning.

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. For example, a supervised learning model can predict how long your commute will be based on the time of day, weather conditions and so on. But first, you’ll have to train it to know that rainy weather extends the driving time.

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. For example, an unsupervised learning model can identify that online shoppers often purchase groups of products at the same time. However, a data analyst would need to validate that it makes sense for a recommendation engine to group baby clothes with an order of diapers, applesauce, and sippy cups.

Other differences:

- **Goals:** 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.
- **Applications**: 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.
- **Complexity:** 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.
- **Drawbacks**: Supervised learning models can be time-consuming to train, and the labels for input and output variables require expertise. Meanwhile, unsupervised learning methods can have wildly inaccurate results unless you have human intervention to validate the output variables.

[3] J. Delua, "Supervised vs. Unsupervised Learning: What's the Difference?," IBM Blog, Mar. 12, 2021. https://www.ibm.com/blog/supervised-vs-unsupervised-learning/

Difference between regression and classification

Regression finds correlations between dependent and independent variables. Therefore, regression algorithms help predict continuous variables such as house prices, market trends, weather patterns, oil and gas prices (a critical task these days!), etc.

The Regression algorithm's task is finding the mapping function so we can map the input variable of "x" to the continuous output variable of "y."

On the other hand, Classification is an algorithm that finds functions that help divide the dataset into classes based on various parameters. When using a Classification algorithm, a computer program gets taught on the training dataset and categorizes the data into various categories depending on what it learned.

Classification algorithms find the mapping function to map the "x" input to "y" discrete output. The algorithms estimate discrete values (in other words, binary values such as 0 and 1, yes and no, true or false, based on a particular set of independent variables. To put it another, more straightforward way, classification algorithms predict an event occurrence probability by fitting data to a logit function.

Classification algorithms are used for things like email and spam classification, predicting the willingness of bank customers to pay their loans, and identifying cancer tumor cells.

[4] J. Terra, "Regression vs. Classification in Machine Learning for Beginners | Simplilearn," Simplilearn.com, Apr. 28, 2022. https://www.simplilearn.com/regression-vs-classification-in-machine-learning-article