Unsupervised Learning

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

In real world, supervised learning involves teachers and a student. Teachers teach the students and measure their performance. They will then adjust the curriculum to get better performance. We experience this type of learning in our daily lives. Consciously or otherwise we also are engaged in unsupervised learning. I personally learned the rules, the play calls, position of the players in the game of football by watching. Unsupervised learning doesn't involve teachers, only students who learn from observation. Our brain is doing more unsupervised learning that supervised.

In machine learning, problems can also be solved either by supervised or unsupervised learning. We call them machine learning. In both cases, the outcome is expected to be in the same way as in real life. In this paper, we will look at unsupervised machine learning, its applications, and the underlying algorithms that power its engine.

Keywords:

Supervised Machine Learning, Unsupervised Machine Learning

Unsupervised Learning

Unsupervised Machine Learning and Algorithms

In supervised machine learning the algorithm is "trained" to fit a model and produce an output that is known or expected. If the model starts showing variations in the output, the model is trained again. The training dataset is usually a sample of a larger dataset and is used to "train" the algorithm. The algorithms in supervised learning are generally regression algorithms such as simple linear regression, multiple linear regression, polynomial, decision trees and other regression algorithms. For discrete variables, classification or categorization algorithms are used. Examples of the classification algorithms are Logistic Regression, K-nearest Neighbors(K-NN), Support Vector Machine (SVM), Kernel SVM, Naïve Bays, Decision Tree, and Random Forest classification.

In unsupervised machine learning the algorithm is not "taught" but rather it "learns". There is no training dataset with labels or columns. The models in unsupervised learning are given data with no labels such as "name" or "salary", etc. and therefore, cannot perform any measures of accuracy on the result. In effect, the data is highly unstructured, and the outcome could be anything the data has at the time. To address this, the model attempts to classify the unlabeled data. For example, if the data is the height and weight of individuals only and without their age or gender, the model would attempt to make classification of the data. If charted as a scatter plot with weight and height on the x and y axis respectively, we may see patterns that resemble clusters. The clusters of data could be separated by shorter people with lower weight and taller people with heavier weight. Based on other data the model could decipher more by making more clusters.

In both supervised and unsupervised learning, the desired outcome is either discrete or continuous. Discrete is when the outcome is either binary (yes/no, like/dislike, exit/stay), nominal (Black, Blue, Yellow, Red), or ordinal (first, second, third). Continuous is when the interval between two numbers is infinity. In unsupervised learning, we address the discrete outcome with classification such as logistic regression, and regression models to address the continuous outcome. In unsupervised learning, we use clustering to address discrete outcome and dimensionality reduction to address continuous outcome.

K-means clustering is a popular method for unsupervised learning. It would create K clusters out of the data. In the example above, we would be able to visualize the two clusters in a simple scatter plot. However, in larger data sets with a many more unlabeled data, K-mean clustering algorithm can be used to not only create clusters from the data but also learns where those clusters should be. It is an iterative algorithm that creates clusters that are closest to what is called centroids. The following steps illustrate how the algorithm works:

1. Choose number of clusters(K)
2. Select at random K points, the centroid (this is a random value not necessarily from the dataset) (i.e. any value from the scatter plot)
3. Assign each data point to the closest centroid (this creates k clusters)
4. Compute and place the new centroid of each cluster
5. Reassign each data point to the new closest centroid.
6. If any reassignment took place in 5, go to step 4; otherwise done.

Its application in image recognition is widespread. It is also used in fraud detection, network intrusion and other fields.

Dimensionality reduction algorithms deal with reducing the number of features in the dataset. In supervised learning, the features have labels and in unsupervised they do not. In supervised, the algorithms select features whereas in unsupervised, they extract features. Some of the techniques used are listed below:

1. Principle Component Analysis (PCA)
2. Linear Discriminant Analysis (LDA)
3. Kernel PCA
4. Quadratic Discriminant Analysis (QDA)

Conclusion

Unsupervised machine learning is used in artificial intelligence at the tip of the iceberg level. It is used in recurrent neural network (RNN) for natural language processing. I is used in in convolutional neural network (CNN) for Image recognition From Robotics to self-driving cars, machine learning will play the key role in their evolution.

References

1. A beginner's guide to dimensionality reduction in Machine Learning - https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e
2. - https://www.forbes.com/sites/bernardmarr/2017/03/16/supervised-v-unsupervised-machine-learning-whats-the-difference/#1da0383f485d
3. Unsupervised Machine Learning - https://www.datarobot.com/wiki/unsupervised-machine-learning/

4. Supervised vs. Unsupervised Learning - https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d
5. Unsupervised Learning: Crash Course AI #6 - https://www.youtube.com/watch?v=JnnaDNNb380
6. Unsupervised Learning explained - https://www.youtube.com/watch?v=lEfrr0Yr684
7. Bengfort, Benjamin; Bilbro, Rebecca; Ojeda, Tony. Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning (p. 107). O'Reilly Media. Kindle Edition.