Mini Project 2

**Unsupervised Learning Report**

by Zhenfei Yu

1. Introduction

Unsupervised learning is a type of self-organized Hebbian learning that helps find previously unknown patterns in data set without pre-existing labels. It is also known as self-organization and allows modeling probability densities of given inputs. It is one of the main three categories of machine learning, along with supervised and reinforcement learning.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

1. Current application of unsupervised learning

Unsupervised learning is more challenging than other strategies, due to the absence of labels. However, they are very significant in machine learning since they can do very complex tasks efficiently.

2.1 Clustering

Clustering is the process of grouping the given data into different clusters or groups. Unsupervised learning can be used to do clustering when we don’t know exactly the information about the clusters. Elements in a group or cluster should be as similar as possible and points in different groups should be as dissimilar as possible. Unsupervised learning can be used to do clustering when we don’t know exactly the information about the clusters.

It is used for analyzing and grouping data which does not include pre-labeled classes or class attributes. Clustering can be helpful for businesses to manage their data in a better way. For example, you can go to Walmart or a supermarket and see how different items are grouped and arranged there. Also, e-commerce websites like Amazon uses clustering algorithms to implement the user specific recommendation system.

These are some of the commonly used clustering algorithms:

* K-Means
* Expectation Maximization
* Hierarchical Cluster Analysis (HCA)

2.2 Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration by getting a set of principal variables.

Many machine learning problems contain thousands of features for each training instance. This will make the training slow as well as it will be difficult to obtain a good solution to the problem.

In dimensionality reduction, the objective is to simplify the data without losing too much information. There can be a lot of similar information in your data. One way to do dimensionality reduction is to merge all those correlated features into one. This method is also called feature extraction.

It is always a good practice to try to reduce the dimensionality of your training data using an algorithm before you feed the data to another machine learning algorithm. This will make the data less complex, much faster, the data may take up less memory and it will perform better at some times.

Reducing the dimensionality may lose some information. So, even if this will speed up the training, most of the times, it may also make your system perform slightly worse. So, use dimensionality reduction only if the training is too slow. Otherwise, try to use the original data.

These are some of the most common dimensionality reduction algorithms in machine learning:

* Principal Component Analysis (PCA)
* Kernel PCA
* Locally-Linear Embedding

## 2.3 Anomaly Detection

**Anomaly detection is the identification of rare items, events or observations which brings suspicions by differing significantly from the normal data.**

In this case, the system is trained with a lot of normal instances. So, when it sees an unusual instance, it can detect whether it is an anomaly or not.

One important example of this is credit card fraud detection. You might have heard about a lot of events related to credit card frauds.

2.4 Visualization

Visualization is the process of creating diagrams, images, graphs, charts, etc., to communicate some information. This method can be applied using unsupervised machine learning.

One example of a visualization algorithm is t-distributed Stochastic Neighbor Embedding (t-SNE).

1. Pros and cons

The pros and or cons of unsupervised machine learning depend on what exactly unsupervised learning algorithms you need to use.

Despite that, there are some common benefits and advantages for the whole group of unsupervised machine learning algorithms.

Advantages:

* Less complexity in comparison with supervised learning. Unlike in supervised algorithms, in unsupervised learning, no one is required to understand and then to label the data inputs. This makes unsupervised learning less complex and explains why many people prefer unsupervised techniques.
* Takes place in real time such that all the input data to be analyzed and labeled in the presence of learners. This helps them to understand very well different models of learning and sorting of raw data.
* It is often easier to get unlabeled data — from a computer than labeled data, which need person intervention. This is also a key difference between supervised and unsupervised learning.

Disadvantages:

* You cannot get very specific about the definition of the data sorting and the output. This is because the data used in unsupervised learning is labeled and not known. It is a job of the machine to label and group the raw data before determining the hidden patterns.
* Less accuracy of the results. This is also because the input data is not known and not labeled by people in advance, which means that the machine will need to do this alone.
* The results of the analysis cannot be ascertained. There is no prior knowledge in the unsupervised method of machine learning. Additionally, the numbers of classes are also not known. It leads to the inability to ascertain the results generated by the analysis.

1. Recommendations

Despite we outlined the benefits and the disadvantages of unsupervised learning, it’s not right to say that unsupervised methods are always well-behaved or imperfect.

The basic tasks and problems you can resolve with machine learning methods are different. When to use unsupervised learning depends on your needs and the problems you have to solve.

1. Conclusion

Unlike [supervised machine learning](https://www.datarobot.com/wiki/supervised-machine-learning/), unsupervised machine learning methods cannot be directly applied to a [regression](https://www.datarobot.com/wiki/regression/) or a [classification](https://www.datarobot.com/wiki/classification/) problem because you have no idea what the values for the output data might be, making it impossible for you to [train](https://www.datarobot.com/wiki/training-validation-holdout/) the algorithm the way you normally would. Unsupervised learning can instead be used to discover the underlying structure of the data.

Each algorithm has its own purpose. Some algorithms are suitable for anomaly detection. Clustering will be the application of some others. Some of the algorithms may be perfect for visualization, finding associations, predicting numerical results, etc.

All these algorithms perform differently for different applications and we need to choose the right algorithm for the right type of application.

6. Refences

[1] http://intellspot.com/unsupervised-vs-supervised-learning/

[2] https://pythonistaplanet.com/applications-of-unsupervised-learning/

[3] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

[4] https://blog.algorithmia.com/introduction-to-unsupervised-learning

[5] https://www.datarobot.com/wiki/unsupervised-machine-learning/