K-Means Clustering

with an implementation of Insurance fraud detection

Vincent Chinedu Obigwe
Electronic Engineering
Hochschule Hamm Lippstadt
Hamm, Germany
vincent-chinedu.obigwe@stud.hshl.de

***Abstract*—K-means clustering is a major machine learning data clustering algorithm that comes with the capability of local optimization. K-means clustering is a numerical unsupervised, repetitive method that which the outcome cannot be determined. It is usually regarded to be fast and effective depending on the usecase it is been applied on. However, this can be argued because there are complaints in terms of performance and the need for improvement. This paper is meant to serve as an introductory guide to the K-means clustering method while pointing out some high and low points. It will be concluded with an implementation of the algorithm in an insurance fraud detection usecase.**

***Index Terms*—component, formatting, style, styling, insert**

I. Motivation

In the olden days, humans have applied various kinds of methodologies in performing tasks in a more easy way. Due to the intelligence of the human brain that helps in developing tools needed to solve these problems. These machines that were invented by humans has made our lives easy by helping us to do many tasks such as laundry, traveling etc. Machine learning therefore is one of such inventions [9]. Intelligent application those days systems that are called intelligent were coded with a bunch of ”if” and ”else”, that helps process the data and make decision on the output based on the users input. For example a spam filter that has a function of filtering out spam emails based on some already blacklisted words and mark them as spam. Usually decision was made using a bunch of decision process to filter out email, but as things progress it gets difficult because the spa word changes, it is therefore important that a system which learns and updates itself is been used. This does not mean that all systems should be based on machine learning, manually creating a bunch of decisions is goods, especially for systems that we have a very good understanding on how we could model the process to generate a good output [10].

A. Machine Learning

Machine learning have served as a solution to life long question of how computers can be built such that it can learn from itself, based on its experience and improve its actions. It is considered to be one of the highest technical field that is at the cross road between computers and statistics. In the machine learning of today, the changes we now see are majorly caused by the development of new machine learning algorithm and theories, also it can be traced to the mass availability of data online coupled with the low cost of computation. Machine learning has so been accepted that it can be seen in all spheres of life from medicine to commerce, that helps us in taking decisions from our past events [7]. Machine learning is an ever changing scope of computational algorithms that is developed to be human like in terms of intelligence, in such a way that it can learn from its environment, this is whx they are considered as the driver of the era of big data [8].

Since machine learning is all about gaining knowledge from data, its application is limitless. We have seen machine learning application from recommendation seen on sites like Spotify to image recognition algorithm used by majority of the social media companies. Apart from the industrial application of machine learning, it further shapes the way in which data driven research is done. These days machine learning is not only applied to high-scale applications, but also to even small applications [10].

Machine learning does not occur in isolation, there are tools and programming language that powers it. Machine learning is mostly done with two programming languages Python, Matlab and R. This is because those languages are considered to be data intensive meaning they work well with data and there are thousands of open source developers that develop modules that is been used within the software. One of such module is the Scikit.learn. Scikit-learn is an open source python module that contains hundred of machine learning algorithm that can be used in solving both supervised and unsupervised machine learning problems. Scikit-learn makes it easy for one to use machine learning because of the wide range of support, performance and documentations available [11].

Machine learning algorithms can mostly be classified as either supervised or unsupervised machine learning, even though some authors introduced reinforcement learning as the third type which learns data and pattern for the purpose of reacting to the environment, we will mostly be dealing with supervised and unsupervised machine learning. The major difference between the two is the presence of label in the training data [12]. From 1 it indicates not just the division of machine learning, but also the presence of data and cloud. Data serves as the most important aspect of machine learning because it is the beginning and substance of the process, while the cloud serves as a platform for the machine learning model to be deployed.

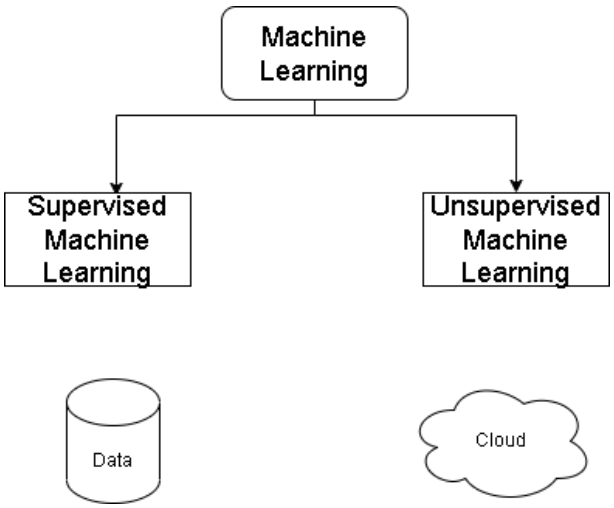

Fig. 1. Classification of Machine Learning

B. Supervised Machine Learning

Supervised machine learning is a type of machine learning that has the ability to generate patterns and theories based on some external examples it to predict future occurrence [13]. Supervised machine learning can also further be classified as a machine learning that maps a function to an input and provides an output but based on the previous sample of the input-output provided. This function mapped is inferred from training a labelled data. In the supervised machine learning, there is need for external assistance and support. Usually in a supervised machine learning, the input data set is divided into two (train and test) data-sets. The training data will contain all the variables including the input and output, while the testing data set does not contain the output to enable one determine its accuracy. All types of supervised machine learning are built in a way that they learn some patterns in training data and then use the pattern learnt to engage in prediction or classification. The 2 explains the flow of machine learning from data set to production in a real industrial scenario. Some of the most popular supervised machine learning models includes decision tree, Naive Bayes, Support Vector machine etc.

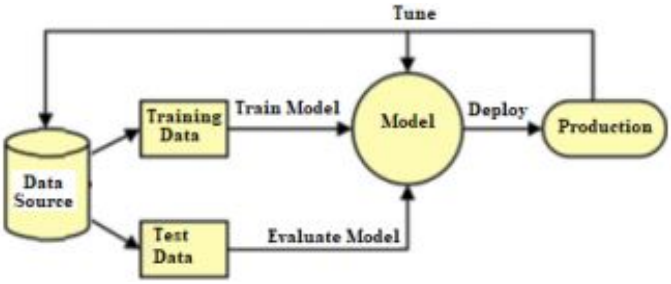

Fig. 2. flow of supervised machine learning [9]

C. Unsupervised Machine Learning

Even though we have a situation where most of the application we have in today's machine learning is composed of supervised machine learning, we still have a high number of data that are unlabelled we would still need to make sense of. Though unsupervised machine learning can be used on both labelled and unlabelled data, it is mostly remembered when we see unlabelled data [6].

It is referred to as an unsupervised learning, due to the fact that unlike in supervised learning, there is no expected output. The unsupervised machine learning algorithm are left on their own to find out patterns on the data presented to them. After the initial pattern learning, when new data is introduced, it will use the initial pattern to classify the new data. It is mostly used for clustering and feature reduction [9].

1) Feature Reduction: Feature reduction is a feature of unsupervised machine learning where an algorithm helps to make create a new version of data in a way that is easier for either humans or other machine learning algorithms to understand and make use of. a very common application is the reduction of data to two dimensions for visualization [10].

2) Clustering: This is usually used to separate data into several groups of similar items. let's take an example with social media websites. When you upload pictures to social media, to better organize your pictures, the site might decide to cluster your pictures wit certain features together, thereby creating several groups from your picture [10].

II. K-MEAN CLUSTERING

K-means clustering is sometimes referred to as the easiest type of unsupervised machine learning in terms of implementation of algorithms that deals with clustering. The procedure adopted by K-means clustering is such that it deals with classification of data in various clusters. The major thing is define K which will serve as the center for each cluster. Care is always when determining the center, because it affects the result [9]. In K-means algorithm, n been the number of available data points are divided into K clusters with regards to the similar attributes that the data points posses. The algorithm is very fast and therefore it is one of the most used clustering algorithm. Application of K-means clustering includes vector quantization, Custer analysis etc [14].

K means clustering works in such a way that after clustering the data, it locates the center of the cluster that is assumed to represent a particular section of the data. This is done by switching between two steps: mapping each data point to the cluster that is nearest to it and setting the center of each cluster created as the mean of all the data that belongs to that particular cluster. The algorithm can be considered done when there is no longer changes in the assignment of the clusters. Figure 3 explains how K-mean nearest works on a synthetic data set. In the figure, the clusters are shown as triangles while the circles represents each data point that is part of a cluster. In K-means clustering, one can determine how many clusters they require from the data set, but care should be taken to avoid over clustering or under clustering a particular cluster because this will determine the insight you might derive from the data [10]. It is important to note that we can assign each data instance to a particular cluster, this is called *hard clustering*, however

when scores are given to the distance between the data point and the center of the cluster, this is called *soft clustering* [6].

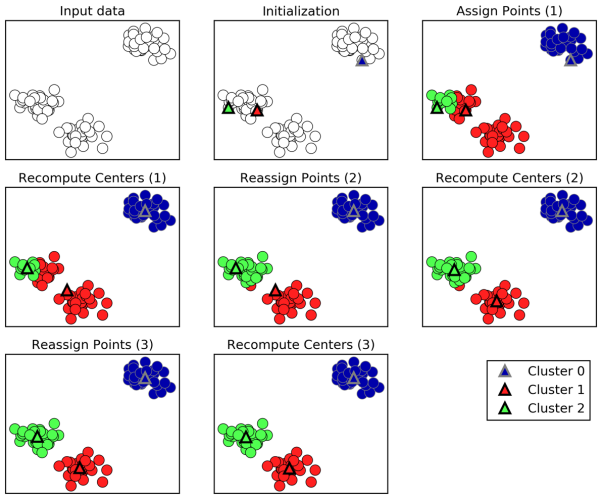

Fig. 3. Three Steps of K-Means algorithm [10]

In figure 3 we indicated that we are searching for three clusters from our data set. The K-means clustering algorithm starts by first initializing three data points randomly to serve as the center of the three clusters we requested for, this can be seen in the initialization box. After the first process, the algorithm now assigns each data point to the closest cluster, this can be seen in the assign point (1) box. Proceeding this step, the cluster center point are recalculated using the mean of the data points assigned, the process does not stop until after two or more iteration when the center of the cluster is no longer changing.

Figure 4 shows the shows the boundaries of the cluster center that we just created. It is quite note able that applying K-means clustering to scikit-learn is very direct and does not involve much complications. The documentation and supports are fully available online.

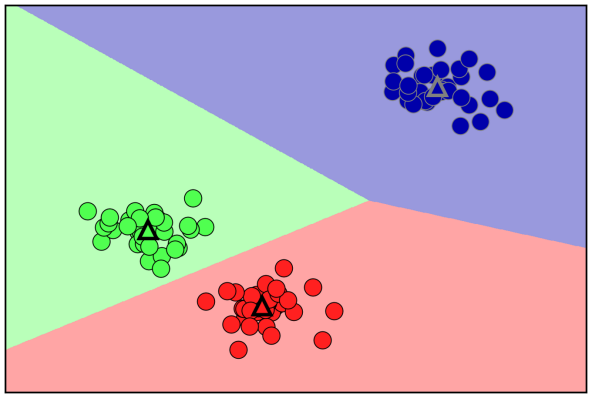

Fig. 4. Diagram indicating the center of each cluster and boundaries [10]

As can be seen from what we have done so far, K-means clustering can be likend to classification algorithm of machine learning. This is because at the end of the whole process for both of them, each item gets a label and are grouped into various groups/clusters. The only difference in terms of labeling is that the labels in K-means clustering dont necessary have a meaning unlike what we have in classification [10].

To calculate the distance of each datapoint to the center of the cluster this formula is used:

$$\mathrm{D}(\mathrm{X}_p, \mathrm{C}_j) = \sqrt{\sum_{i=1}^{d} (Xpi - Cji)^2}$$

[14]

while to recalculate the new center of the cluster we use:

$$\mathrm{Cj} = \frac{1}{NJ} \sum \forall xp \ \in \ cj \ Xp$$

[14]

One of the major shortfalls of K-means clustering is that there is always a problem providing the number of clusters needed. To be able to do this almost accurately, one need to have a perfect knowledge of clustering.

III. K-Means Clustering Implementation: Customer Segmentation

IV. Conclusion

This section concludes our write-up while still providing a context for further research that could still be carried out in this domain.

References

[1] Jin X., Han J. (2011) K-Means Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.

[2] Na, S., Xumin, L., Yong, G. (2010, April). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In 2010 Third International Symposium on intelligent information technology and security informatics... (pp. 63-67). Ieee.

[3] Lloyd, S. P. (1957). Least squares quantization in PCM. Technical Report RR-5497, Bell Lab, September 1957.

[4] Steinley, D., Brusco, M. J. (2007). Initializing k-means batch clustering: A critical evaluation of several techniques. Journal of Classification, 24(1), 99–121.

[5] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam J. Neyman (Eds.), Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, pp. 281–297). California: University of California Press.

[6] Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".

[7] Jordan, M. I., Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[8] El Naqa, I., Murphy, M. J. (2015). What is machine learning?. In machine learning in radiation oncology (pp. 3-11). Springer, Cham.

[9] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

[10] Müller, A. C., Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[12] Berry, M. W., Mohamed, A., Yap, B. W. (Eds.). (2019). Supervised and unsupervised learning for data science. Springer Nature.

[13] Singh, A., Thakur, N., Sharma, A. (2016, March). A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). Ieee.

[14] Rajeswari, K., Acharya, O., Sharma, M., Kopnar, M., Karandikar, K. (2015, February). Improvement in K-means clustering algorithm using data clustering. In 2015 International Conference on Computing Communication Control and Automation (pp. 367-369). IEEE.

V. Appendix

Our code and some diagrams that could not be captured in the original write-up appears here.

Declaration of Originality

I, Vincent Obigwe, herewith declare that I have composed the present paper and work by myself and without the use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The paper and work in the same or similar form have not been submitted to any examination body and have not been published. This paper was not yet, even in part, used in another examination or as a course performance. I agree that my work may be checked by a plagiarism checker.