GLOBAL RISK INSTITUTE

FINANCIAL INNOVATION SERIES

Machine Learning: Unsupervised Learning in Finance

Author: Alexey Rubtsov,
PhD, Senior Research Associate, Global Risk Institute

INTRODUCTION

The last decade has witnessed a large-scale adoption of machine learning tools in finance. According to the latest report by Refinitiv, the number of data science teams in financial services firms have risen by more than 260 per cent since 2018 (see Refinitiv 2020). This extraordinary growth stems largely from recent revolutionary applications of machine learning (e.g., Google Neural Machine Translation, AlphaGo) and reveals the potential of machine learning to transform almost all aspects of the financial services industry. Evidently, it has become critical for financial executives to be able to effectively communicate with data science professionals. For instance, in one of its reports, J.P. Morgan's quantitative investing and derivatives strategy team wrote (J.P. Morgan 2017):

> *Regardless of the timeline and shape of the eventual investment landscape, we believe that analysts, portfolio managers, traders and CIOs will eventually* ***have to become familiar with Big Data and Machine Learning approaches to investing.***

In its "Financial Innovation" series, the Global Risk Institute provides non-technical reviews of Machine Learning (ML) tools, its financial applications, and associated risks that executives should be aware of when developing ML solutions in their organizations. In this paper we discuss Unsupervised Learning (UL), one of the four main categories of ML.* Financial applications that we consider include: understanding country risk for foreign investment, trading, model risk management, fraud detection, assessment of companies' financial situations, financial regulation, identification of complex relationships in stock markets, and early warning models for financial crises.

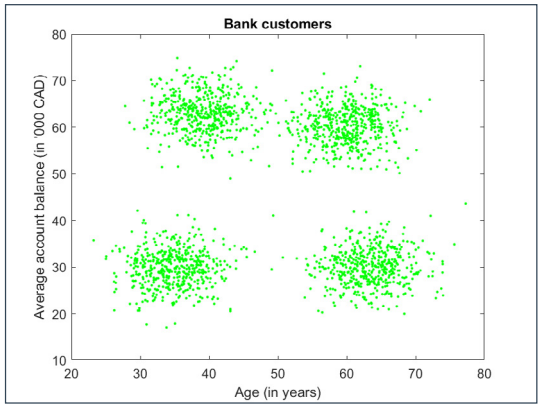
UL is used to draw inferences from data. The main goal is not to predict a certain variable, but rather to understand the structure of the data. Methods of UL can often be categorized as either *clustering* (splitting the data into groups also called clusters) or *factor analyses* (identifying the main factors that best describe the data).

* The other three categories of Machine Learning are Supervised, Semi-Supervised, and Reinforcement Learning.

UNSUPERVISED LEARNING (UL)

In its simplest form an UL algorithm attempts to find similarities in data. To illustrate, consider the following bank customers data shown in Figure 1.

(a)



(b)

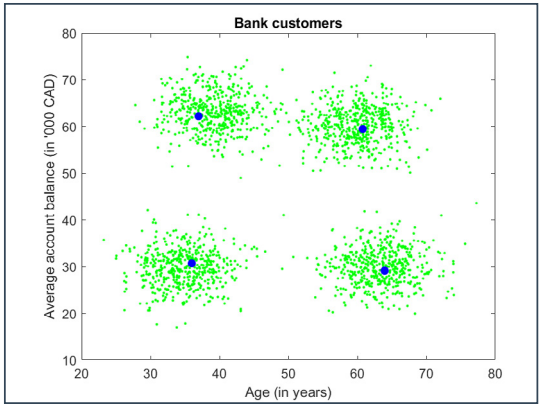


***Figure 1**. Age and average account balance of 2,000 bank customers. Part (a) data; part (b) specification of four groups that has the smallest sum of squared distances of observations in the group from the group average/centre (large dots)*

It is clear from Figure 1(a) that the customers can be regarded as consisting of four groups (or clusters): Junior/Low Balance, Junior/High Balance, Senior/Low Balance, and Senior/High Balance. An important observation is that it is distances among the points in Figure 1(a) that allowed us to conclude that there are four groups of bank customers. A group can be defined as a set of points that are relatively close to each other. We also note that clustering is the main tool used in UL.

One of the most popular clustering algorithms is the k-means algorithm. The algorithm determines each group by minimizing the sum of squared distances of observations in the group from the group average, also called the 'centre of the group'.[†] In other words, among all possible group specifications the algorithm attempts to find the specification that has the smallest sum of squared distances of observations in the group from the group average (see Figure 1(b)). Initially the centres are arbitrarily specified and then updated by the algorithm until the best match to the data structure is found. Future *classification* of new data is based on proximity of the data to group averages. Intuitively, among all possible ways to split the data in Figure 1 into four groups, the algorithm selects the one that is most consistent with the data structure.

The major difference from Supervised Learning is that there is no a priori information regarding what data belong to what group, that is, there is no "supervisor" who tells the algorithm how to classify the data. It is the algorithm that identifies the groups based on the similarities among the observations.

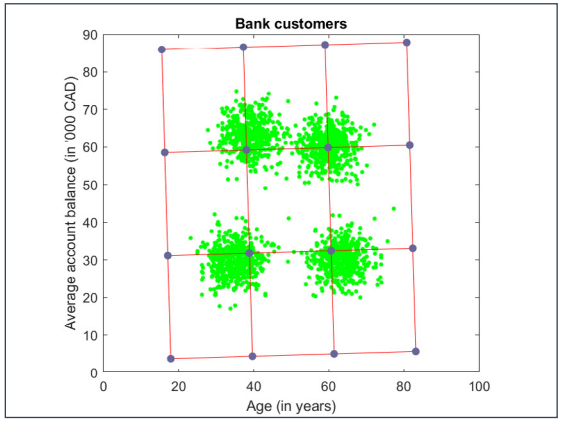
[†] See also Hull, J., Mishra, N., Rubtsov, A.: Machine Learning: The Benefits and Pitfalls. Financial Innovation series, Global Risk Institute in Financial Services.

After the algorithm has determined the groups, it might be a challenge to explain why the data is structured in a particular way. For instance, in the above example it is not clear why there is no group of customers centered at age 50.

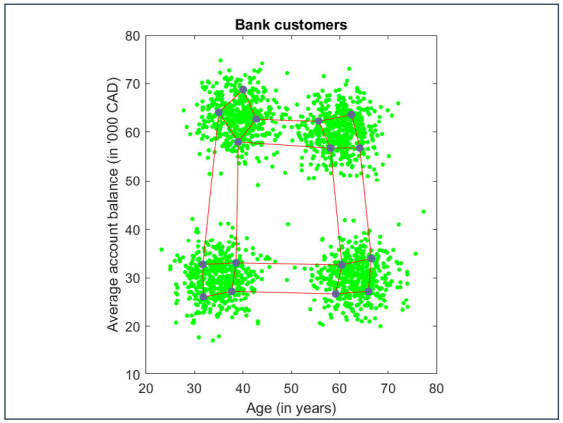
Although in this example it was easy to identify groups visually, it is far less obvious how to do so if each customer is represented by, say, 20 characteristics, also called features in data science speak. Furthermore, understanding what each group represents becomes even more challenging especially if the number of groups is large. In this respect there are quite powerful UL algorithms that can visually represent data with many characteristics. For example, Self-Organizing Map (SOM) is a complementary tool that can be used for initial data analysis. SOM applies a process known as self-organization to the initial data set. For example, when financial data for a group of bank customers is introduced, the data will be self-organized in such a way that those with similar characteristics will be located close to one another on a map. We clarify this concept with an example.

Let us consider using an ML algorithm to split the data in Figure 1 into 16 groups.‡ However, instead of arbitrary specification of group centres, we now specify initial centres as a rectangular grid (see Figure 2(a)).

(a)



(b)



(c)

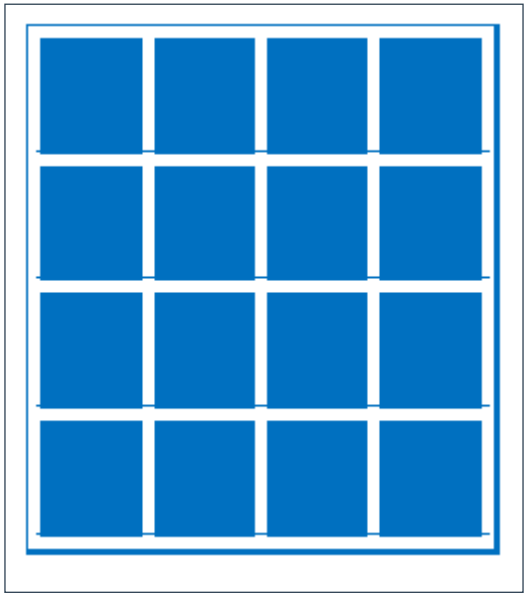


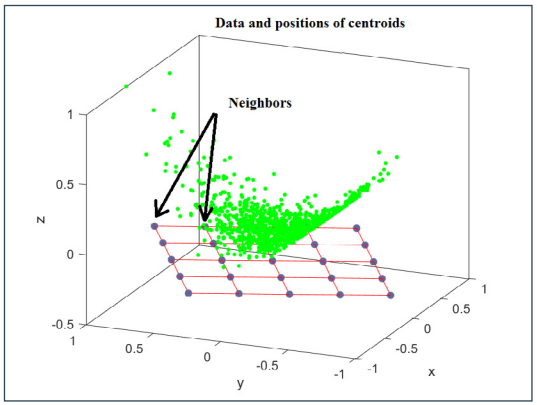
Figure 2. *Age and average account balance of 2,000 bank customers. Part (a) data (green) and SOM with initial specification of 16 group averages (blue); part (b) data (green) with adjusted by ML algorithm group averages (blue); part (c) map of group positions: each square corresponds to a group.*

‡ We chose a larger number of groups just to make the intuition behind SOM clearer.

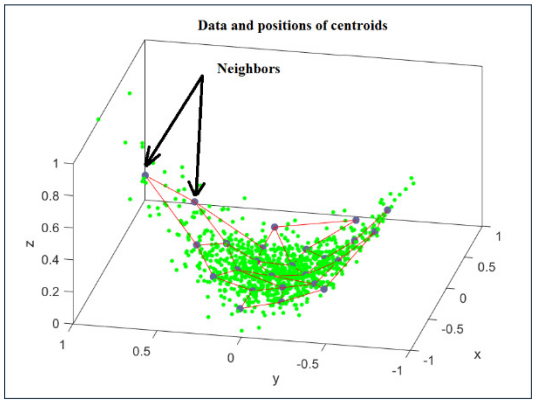
The SOM algorithm adjusts the grid in such a way that the initial order of group centres is preserved even after adjustments (see Figure 2(b)). In particular, the positions of all group centres are the same as their positions in map (see Figure 2(c)). How is this useful? When the data about customers have many characteristics, it is not possible to visualise the data and determine similarity among the groups. On the other hand, if the map in Figure 2(c) is used as initial positions of group centres, then we are sure that these positions are preserved after the algorithm split the data into groups. For example, if two customers belong to two groups that are close to each other on the map in Figure 2(c), then the two customers have similar characteristics.

Although Figure 2 illustrates the nature of SOMs, it does not capture their main purpose because situations where there are only two features can be visualized without difficulty. SOMs are most useful when they reduce many dimensions to two dimensions for visualization purposes. For example, consider a map that takes the form of the U-shaped data with 3 characteristics (see Figure 3).

(a)



(b)



(c)

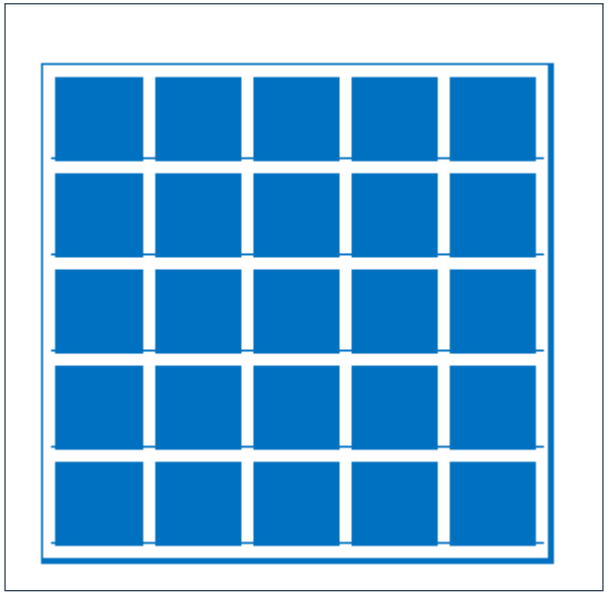


Figure 3. *Synthetic data with three characteristics X, Y, and Z. Part (a) data (green) and SOM with initial specification of 25 group averages (blue); part (b) data (green) with adjusted by ML algorithm group averages (blue); part (c) map of group positions: each square corresponds to a group.*

As it follows from Figure 3(a,b), neighbouring group centres remain neighbours even after adjustments. This feature of the SOM algorithm allows us to reduce the original three-dimensional data to the two-dimensional map shown in Figure 3(c).

SOM-based data analysis starts with "colouring" the map: assigning certain colours to groups that share a certain characteristic. For instance, if the data represents banks' balance sheet data, then one can identify groups of banks that went bankrupt (colour those in red) and the group of financially stable banks (colour those in green). The map can later be used to analyse the financial condition of a given bank by checking the proximity of the bank to either of the groups. An example of a coloured map is shown in Figure 4.

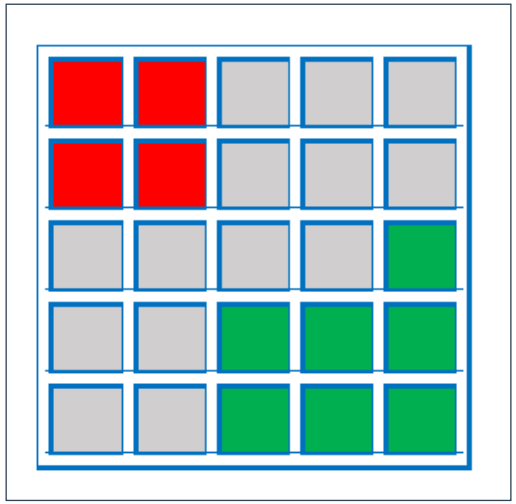


***Figure 4**. An example of a coloured map where SOM algorithm was applied to banks' balance sheet data. Red area represents a group of banks that went bankrupt; green area represents a group of financially stable banks; grey area represents banks with intermediate financial situations.*

The second category of UL algorithms is concerned with identifying the main factors that best describe the data. Alternatively, these algorithms can also be used in reducing the number of variables used in various analyses. For example, the Bank of Canada provides the information on zero-coupon yields for 120 different maturities (yield curve).[§] Since the yields of close maturities are highly correlated, it makes sense to seek ways to reduce this data to a smaller number of features. Principal Component Analysis (PCA), an UL technique, can be used toreduce the amount of the data.[¶] Figure 5 shows zero-coupon yields of maturities 6 and 7 years together with two Principal Components (PCs) identified by the UL algorithm.

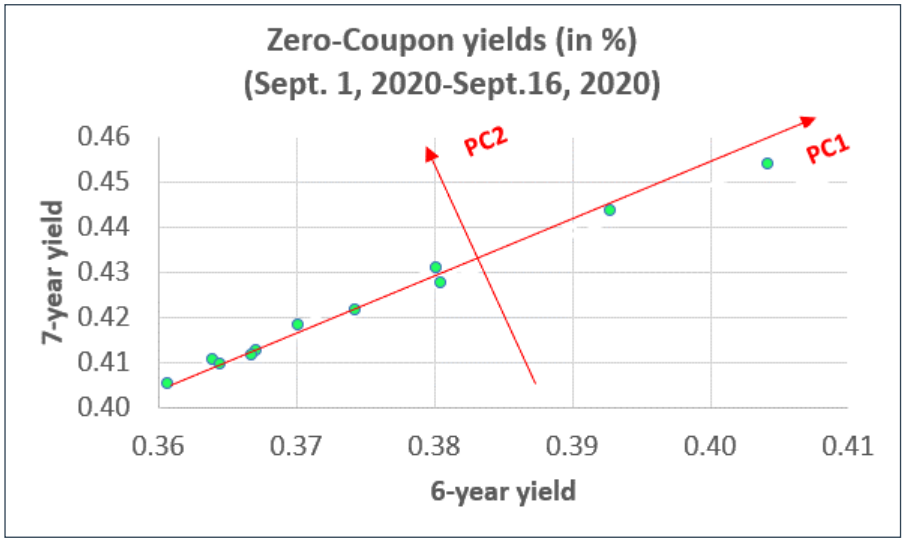


***Figure 5.** Zero-coupon yields (green dots) and Principal Components (red axes)*

Notice that PCA allows us to reduce the dimensionality of the data: instead of reporting the yields as a pair (*6-year yield, 7-year yield*) one might as well report them as a number, namely, the position on the PC1 axis because their variability

§ See https://www.bankofcanada.ca/rates/interest-rates/bond-yield-curves/

¶ A Neural Network implementation of PCA is known as Oja's learning rule (see Oja (1982)).

along PC2 is insignificant. In other words, two yields were condensed to one factor – PC1. This allows us to reduce the yields data by half and simplify further analysis/use of the data.

Finally, we note that as we reduce two features to one factor in Figure 5, we can often use PCA to identify a small number of factors that describe the behaviour of a large number of features. For instance, 120 zero-coupon yields (features) provided by the Bank of Canada are often reduced to only 3 factors (3 PCs).

APPLICATIONS IN FINANCE

In this section we describe some of the applications of UL in Finance.

Country Risk. UL algorithms can be used to understand the risk of countries for foreign investment. Hull (2020) uses four characteristics for each of 122 countries: the real GDP growth rate, a corruption index, a peace index, and a legal risk index. The clustering algorithm identifies three groups of countries that can be labeled as having high, moderate, and low risk. Such an analysis allows us to identify countries that are similar based on the risk characteristics used in ML algorithm.

Trading. It has been common in the financial investment industry to conduct backtests, which are simulations of how an investment portfolio would have performed under a particular historical scenario. However, the probability of making a false discovery (finding a false positive) increases as more and more tests are conducted on the same data. Lopez de Prado and Lewis (2018) employ an UL clustering algorithm to compute the probability that an investment strategy is a false positive, while controlling for selection bias under multiple testing.

Model Risk Management. A report by the Financial Stability Board (2017) gives an example where one global corporate and investment bank is using UL algorithms in model validation. Its equity derivatives business has used this type of machine learning to detect anomalous projections generated by its stress-testing models. Each night, these models produce over three million computations to inform regulatory, internal capital allocations and limit monitoring. A small fraction of these computations is extreme and knocked out of the normal distribution of results by a quirk of the computation cycle or faulty data inputs. UL algorithms help model validators in the ongoing monitoring of internal and regulatory stress-testing models, as they can help determine whether those models are performing within acceptable tolerances or drifting from their original purpose. They can also provide additional input to operational risk models, such as the vulnerability of organizations to cyber-attacks.

Fraud Detection. Like the previous application in Model Risk Management, UL algorithms can be used to detect credit card fraud, cyber fraud, wire fraud, and insurance fraud. The idea is that UL algorithms can be successful in identifying outliers (frauds in this example) in the data.

Financial Situation of Companies. The basic objective is to analyze solvency of a given company. Various financial ratios can be used as characteristics of each firm. The advantage of a graphical representation is that a simple preliminary explanation can be given for the classification of a company into the bankrupt or solvent group. Using SOM one can see a firms' financial situation in a particular year. We can also observe the evolution of a company, by introducing the financial information from various years. If it clusters with the failed firms, then it must be treated with care, on the grounds that its financial structure is not different from that of other companies that have failed in the past. If it clusters with non-failed firms, that concern disappears.

Financial Regulation. The U.S. Securities and Exchange Commission collects structured and unstructured data for investment advisers. UL algorithms are used to identify outlier reporting behaviours – including both topic modelling and tonality analysis (topic modelling lets the data define the themes of each filing and tonality analysis gauges the negativity of a filing by counting terms with a negative connotation). Once the output of this first stage is obtained, it

is then combined with past examination outcomes and fed into a second-stage machine learning algorithm to predict the presence of idiosyncratic risks for each investment advisor.

Identifying Complex Relationships in Stock markets. Goldman Sachs uses UL algorithms in its research to disentangle very high-frequency effects in global markets and understand complex, non-linear relationships between stocks.** Traditional techniques may fail to perform many tasks at the scale and speed required for electronic execution. On the other hand, UL algorithms have many computationally efficient tools that are particularly suitable for finding complex relationships in large amounts of data.

Early Warning Models for Financial Crises. In Lv et al. (2020) simple UL clustering algorithms combined with other ML tools were shown to be highly effective in predicting financial risks for companies.†† The model was tested on Chinese companies and showed results superior to traditional models typically used by firms.

RISKS AND CHALLENGES

UL algorithms are not free from drawbacks. Following, are some of the risks that should be considered before UL algorithms are used by financial institutions.

Given that UL is not "supervised" there is no direct measure of successful learning for the algorithm. In other words, it is hard to assess the validity of conclusions drawn from the output of UL algorithms. As a result, data scientists usually develop heuristic arguments to evaluate the performance of the algorithm. This could be a source of potential risk when assessing the quality of the results.

One of the most important factors in the success of UL is feature selection. Features that seem irrelevant in isolation may be important in combination with other features. On the other hand, some features that seem important can be disregarded if they are highly correlated with other features used in the analysis. Some experimentation may be necessary to identify the best set of features to be used by an UL algorithm.

It is incorrect to think that having more features does not hurt as, at worst, those features provide no new information. This problem is known as the "curse of dimensionality". As discussed earlier, UL clustering algorithms use distances among observations to determine groups. In this respect, the way one measures the distances among observations becomes critical. The most common distance measure is the Euclidean measure: the distance between two points is the length of the line joining the two points. However, if this measure is used, a large number of features makes it difficult for the UL algorithm to distinguish between observations that are close and that are far apart. One solution could be to change the distance measure. Another solution could be to use various techniques to reduce the number of features (see the discussion of PCA in section "Unsupervised Learning (UL)" on page 3).

Results of UL algorithm might have ambiguous interpretation. In the country risk example provided in the previous section, an analyst might assume that the chosen features are related to investment risk, but there is no guarantee that this is the case. For instance, the characteristics in that example were not related to losses incurred on investments in different countries, as could be the case in Supervised Learning. Thus, it is critical to have a particularly good understanding of how one should interpret/use the results obtained from UL algorithms.

** Prising open the black box of AI. Risk.Net, February 26, 2020

†† Lv, D., Wu, C., Dong, L.: A k-means++-improved radial basis function neural network model for corporate financial crisis early warning: an empirical model validation for Chinese listed companies. Risk.Net, August 27, 2020.

When SOMs are used to visualize data, it is known that observations close to each other on a map have similar characteristics. However, it is essential to understand that if two observations are indicated as being far apart on the map, it does not imply that they have quite different characteristics. It could be that data are structured in such a way that group centres, far apart on the map, actually represent groups with similar characteristics.

UL algorithms may reinforce biases. Given that the learning is unsupervised, there exists a risk to entrench biases that are present in the data that feeds the algorithm. For instance, UL algorithms may showcase fewer female applicants as weak only because the data is insufficient, having fewer female applications, not because females are actually weak applicants.

In summary, to successfully apply an UL algorithm, it is wise to use other exploratory data techniques to confirm the findings of the algorithm. In applications where UL and Supervised Learning are used in conjunction, to complement one another, one should also consider the risks stemming from Supervised Learning.

REFERENCES:

1. Artificial intelligence and machine learning in financial services. Market developments and financial stability implications. Report by Financial Stability Board (2017)
2. Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing. Global Quantitative & Derivatives Strategy, report by J.P. Morgan (2017)
3. Detection of false investment strategies using unsupervised learning methods by Lopez de Prado, M., Lewis, M.. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3167017
4. Machine Learning in Business: An Introduction to the World of Data Science by Hull, J.C.. Second Edition (2020)
5. Simplified neuron model as a principal component analyzer by Oja, E.. Journal of Mathematical Biology 15(3), 267–273 (1982)
6. The rise of the data scientist. Report by Refinitiv (2020)