Module 3

3.3 Unsupervised learning

By Rafael Alcalá Fernández
Senior professor at the University of Granada, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI).

1. WHAT IS UNSUPERVISED LEARNING?

As seen in the previous capsule, when implementing supervised learning, we normally have access to characteristics (input variables) $X_1, X_2, \ldots, X_p$ for a set of $n$ observations (instances), and a response $Y$ (output variable) associated with the same $n$ instances. Thus, the goal is to predict $Y$ using $X_1, X_2, \ldots, X_p$. However, an associated response is not always available, or even if it is available, we might be interested in discovering other types of associations. In these cases, a series of techniques can be used that are referred to as ‘unsupervised learning’. The term unsupervised refers to the fact that this learning is not based on the existence of a previously known response. That is, the learning is unmonitored, with the correct answer being provided after each failed attempt, much like a teacher might correct a student (see figure 1).

Unsupervised learning comprises a set of statistical tools intended for an environment in which we only have (or use) one set of variables $X_1, X_2, \ldots, X_p$ rather than a set of $n$ instances. In this case, we are not interested in prediction, because we do not have an associated output variable $Y$. Rather, the goal is to discover interesting things about the various measurements $X_1, X_2, \ldots, X_p$ such as

- Whether there is an informative way to represent the data.
- If we can recognize subgroups between the instances.
- The possibility of finding relationships of interest between the variables themselves.

Unsupervised learning refers to a diverse set of techniques used to help us answer questions like the ones above. Among the best known are clustering and association rules, although techniques such as principal component analysis (a technique we saw in Section 2.3 of Capsule 2 in Module 2) are also part of unsupervised learning. In this massive open online course we will examine the first two questions as examples.
2. THE DIFFICULTY OF UNSUPERVISED VERSUS SUPERVISED LEARNING

Supervised learning is an area with well-defined objectives. For example, if you are asked to predict a binary output for a data set, an extremely extensive set of very well-developed tools will be at your disposal. These include logistic regression, linear discriminant analysis, decision trees, and support vector machines, among others. In addition, you will already have a clear understanding of how to assess the quality of the results obtained (through cross-validation, validation with an independent data set, and direct error estimation metrics, etc.).

In contrast, **unsupervised learning is often much more challenging**. Its application tends to be more subjective and there is no single clear objective (such as prediction of an output variable) for the analysis. Therefore, unsupervised learning is often performed as part of the exploratory data analysis. Furthermore, it can be very difficult to evaluate the results obtained from these methods because there is no universally accepted mechanism for cross-validating or for simple validation of the results on an independent data set.

The reason for this difference is simple. If we fit a predictive model using a supervised learning technique, then it is possible to verify our work by seeing how well our model predicts the output variable $Y$ in the instances not used to fit the model. However, in unsupervised learning, there is no direct way to verify our work because we do not know the true answer: the problem is unsupervised.
3. THE IMPORTANCE AND APPLICABILITY OF UNSUPERVISED LEARNING

Techniques for unsupervised learning are becoming increasingly important in a multitude of areas. For example, an online shopping website could attempt to identify groups of shoppers with similar browsing and purchasing histories, as well as items of particular interest to distinct groups of shoppers. Likewise, retailers may also want to look for causal relationships between the purchase of some items with the purchase of others. This would then allow individual buyers to be preferentially shown items they would most likely be interested in based on the purchase histories of similar buyers. Thus, a search engine could choose which search results to show a particular individual based on the click history of other individuals with similar search patterns. Or it could show them items that they are more likely to purchase after having previously made certain other purchases, all based on machine-learned causal relationships.

Similarly, moving into the area of bioinformatics, a cancer researcher could evaluate the gene expression levels in 100 breast cancer patients. They could then look for subgroups among breast cancer samples or among genes to gain a better understanding of the disease. Similarly, we could look for causal relationships between the activation of some genes and others (which could perhaps then be considered markers of interest). Perhaps we could even identify a relationship between the activation of certain groups of genes and the development of a specific type of breast cancer. These statistical learning tasks, and many more, can be performed through unsupervised learning techniques.

4. CLUSTERING

Clustering refers to a set of unsupervised learning techniques aimed at identifying groups in the data. A group (or cluster) is a set of objects (instances) that resemble each other. Thus, the objective of a clustering algorithm is to group the available objects in such a way that objects within the same group are similar to each other, and different from objects in other groups.

Considering that we have already seen classification in the previous capsule, another possible definition is that clustering consists of performing a classification on a set of data without previously knowing its classes; it is not only that we do not have the information about the class to which the instance may belong, but that we do not even know if there are any classes or how many may exist. Therefore, we try to group by similarity in groups that, a priori, are unknown, but are sufficiently different from other groups to be able to state that each one represents a cluster. In other words, they represent an entity or a possible class.

It is almost intuitive to think that a clustering problem is, a priori, more difficult to solve than a classification problem. This is because, in the case of classification, at least the data set classes are known. In general, this statement is true: if the classes are known, it will be easier for the data to be well classified (or grouped by classes). The information discovery provided by clustering techniques is often extremely useful when very clear clusters are obtained. This is the case even when only one cluster is clearly identified, if this happens to be the group in which we are most interested.
Numerous functions are available to measure the similarity or ‘distance’ between objects. The use of one or the other depends on the types of variables and the problem itself. In addition to the proposed definition of similarity (distance), a multitude of clustering algorithms can identify different clusters, even for the same problem. Here we will present some of the best-known distance measures along with the methodology implemented with some of the most widely used algorithms (presented in depth in Module 6). Figure 2 shows the evolution (convergence) of the \( k \)-means algorithm in an example with \( K = 3 \) clusters, which were initially generated centered on three randomly chosen instances. Before watching the animation, look at the data distribution and imagine how the clusters should look at the end.

Figure 2. Convergence created by implementing the \( k \)-means algorithm. Move your mouse over the image until the link appears; click it to view the animation in a browser. Source: available in the public domain (Wikipedia).

5. ASSOCIATION RULES

Association rules are used to identify and represent dependencies between the elements or values in a data set in which, likewise, we do not know the class to which these dependencies belong. These rules are defined as expressions of the type:

\[
A \rightarrow C,
\]

where \( A \) and \( C \) are sets of elements that verify \( A \cap C = \emptyset \); \( A \) is known as the ‘antecedent’ of the rule and \( C \) as the ‘consequent’ of the rule. These rules represent the fact that when the elements of \( A \) appear in an instance of the data set, there is a high probability that the elements of \( C \) will also appear in that same instance.
Association rules were initially used to detect associations among products purchased by consumers in supermarkets. In this particular case example,

- The products in the supermarket are the elements among which we want to find associations, called *items*.
- A set of $k$ products is called an *itemset*. More precisely, it is called $k$-itemset, where $k$ is the number of items comprising it.
- Each of the supermarket sales is called a *transaction*.

By analyzing these data sets we can extract rules such as:

```
“Diapers → Wipes”,
```

indicating that every time diapers are purchased, wipes are also usually purchased. These types of rules can help the supermarket owner to make decisions about what offers to make and how to distribute the products in the supermarket to increase sales, improve the quality of service provided, and increase customer satisfaction. While the rule shown seems obvious, others such as “Diapers → Beers” may also appear that might not always be so obvious. Thus, there is a duality between the appearance of rules that only represent a way to confirm that the algorithm is working correctly (information already known) and those that represent the real discoveries (the rules we are really looking for).

Association rules are commonly evaluated using the classical measures of ‘**support**’ (the frequency with which the rule is satisfied in the data set) and ‘**confidence**’ (indicating how many data set transactions into which the antecedent fits, the consequence of the rule also appears). In general, algorithms try to obtain rules with high confidence by requiring them to comply with minimum support levels. However, there are now many quality measures that can be applied to select and classify rules according to their potential interest to the user (e.g., *Lift*, *leverage*, *conviction*, etc.). In Module 6, we present the different quality measure alternatives and the methodology used for the most popular algorithms, also explaining how to run them in Python.
6. DATA CHARACTERISTICS THAT INFLUENCE LEARNING

As in the case of supervised learning, it is important to discuss whether it is possible to determine the optimal amount of data required for correct learning, as well as the relationship between the number of instances and the number of variables that represent them. Of note, in this case we also have the same problems as in the case of supervised learning. Indeed, these issues are even aggravated for the same reasons indicated in section 2.1 of this capsule, because unsupervised learning usually presents more difficulties, as explained below.

- **Problems related to the number of instances**: We must have a minimum number of instances (data) which, as mentioned above, usually also depends on the number of variables. We can be guided by the values indicated for supervised learning in section 2.1 in the previous capsule. However, we already know that there is no exact answer to this question.

- **Difficulties regarding the input variables**: In this case, the instances are all input variables so we are not looking for a dependency relationship with respect to an output variable. However, there will also be confounding variables that may cause spurious groupings. In the case of association rules, this will mean that they do not represent true causality, or in the case of clustering, they may make us think that there are more groups than there really are. Similarly, dimensionality can be a problem for obtaining good results and so pre-processing is often applied to try to reduce this problem.

- **Hindrances in relation to data bias**: Apart from the problem of confounding variables (usually caused by bias in the data), unsupervised learning is also severely affected by imbalances in the distribution of any of the variables. It will be very difficult to group or associate instances if we do not have a representative sample of all the possible cases or situations in our data.

Finally, because we need to calculate the similarity to form the clusters, the **heterogeneity of the variables** can also represent a serious problem. If the attributes are numerical and are directly comparable, the calculation of similarity or distance is very simple. However, when the examples contain complex and heterogeneous attributes, things become more complicated. The Euclidean distance is usually used, but when we have categorical variables such as sex, this cannot be directly calculated. Numerical values are usually assigned (for example, male = 0, female = 1), but this can only be done if there are only two values or if there is an ordinal relationship between the categorical values, for example, “prepubescent” = 1, “pubescent” = 2, and “adult” = 3.

Scaling factors must also be considered. If distances are being calculated, the data should be normalized to the same range. In this case, a normalization of the values to the range [0,1] is usually applied. Because of the above, it is very important to pay attention to the exploratory analysis of the data and to perform data set preprocessing or preparation to facilitate the learning tasks.
7. ASSESSMENT OF UNSUPERVISED LEARNING RESULTS

As in the case of supervised learning, an important issue to discuss is whether the results obtained are of sufficient quality. In this case, it makes no sense to apply cross-validation, or any kind of validation based on knowledge of any kind of output values. The question to answer in this case would be: is the information we have discovered reliable? In our case, from a bioinformatics point of view, it would be equivalent to asking: is the information we have discovered supported by the data and does it have a biological foundation?

Whenever clustering or association rule learning is applied to a data set, we will find clusters and rules, respectively. Particularly in the case of association rules, we can even obtain hundreds or thousands of rules for some problems. However, we really want to know if the clusters or rules that have been found represent true subgroups in the data, or if they are simply the result of noise clustering or associations. For example, if we could have an independent set of data, would that data also show the same set of clusters, or would it still provide the same support for the rules?

This is a difficult question to answer. For example, there are several techniques for assigning a p-value to a cluster to assess whether there is more evidence for that cluster than would be expected due to chance. Similarly, as noted above, there are also quality metrics for association rules. However, there is no consensus on a single best approach to this issue because, in any case, the true assessment is usually linked to a check/validation performed by experts in the relevant area. In bioinformatics, in the vast majority of cases, the findings usually guide the laboratory research, and it is there that the information discovered is truly checked to see whether or not it is reliable.
BIBLIOGRAPHIC REFERENCES

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer, 2013 (Capítulos 2 y 10).

ADDITIONAL REFERENCES