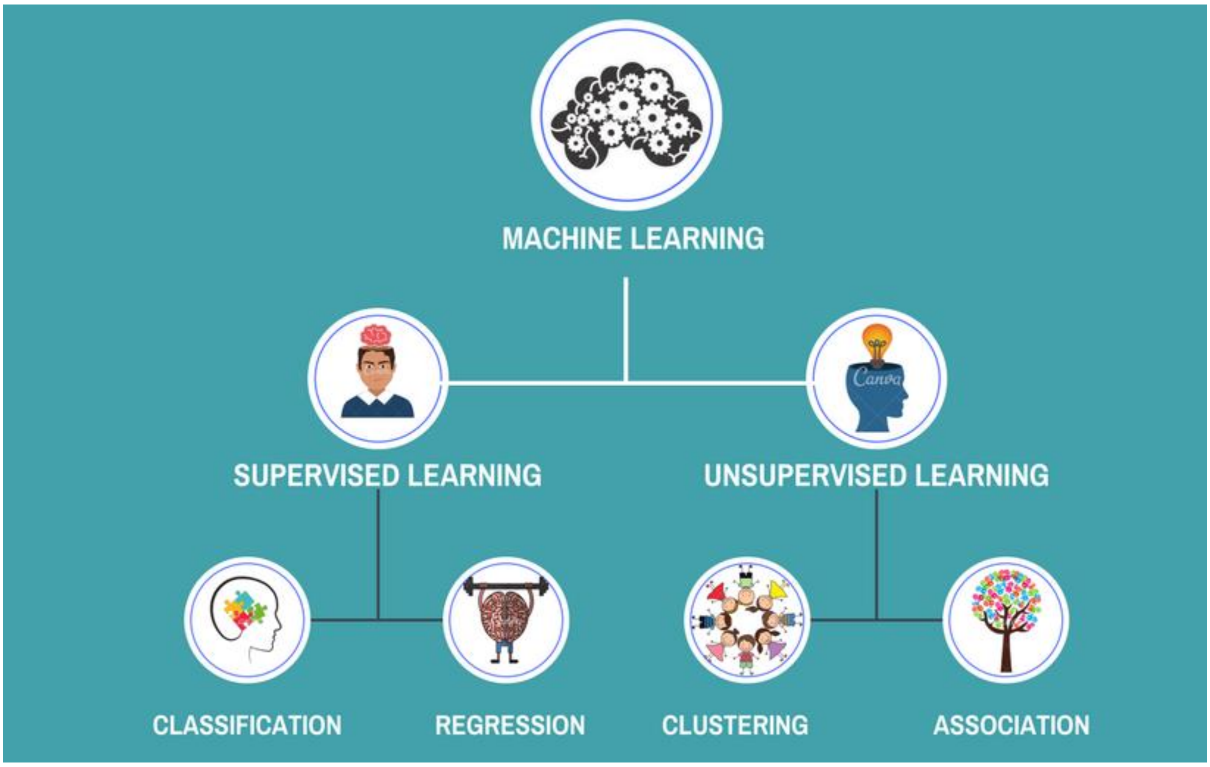
Blog on Clustering (Part-2)

By Deepak Kaura

In previous part you learned and read about clustering , Now this part will revolve around Unsupervised vs supervised learning and who is better.



Supervised and unsupervised learning represent the two key methods in which the machines (algorithms) can automatically learn and improve from experience.

This process of learning starts with some kind of observations or data (such as examples or instructions) with the purpose to seek for patterns. We use those patterns to make better decisions or forecast based on the examples/ instruction that we provide.

The goal is to let the computers (machines) learn automatically without people assistance and adjust actions suitably.

What is Supervised Learning?

Supervised Learning when a computer uses given labels as examples to take and sort series of data and thus to predict future events.

In supervised learning people teach or train the machine using labeled data. Labeled data means it is already tagged with the right answer.

In other words, the machine algorithm starts from the analysis of a well-known training dataset (also called input data) and then model a function to make predictions about future outcomes.

That's why it is called supervised – because there is a teacher or supervisor.

Example:

Let's give an example to make things clearer:

Suppose you have a bunch of different kinds of flowers. First, you need to train the machine on how to classify all different flowers: You can train it like this:

- If there are thorns and the head has color Red then it will be labeled as Rose.
- If there aren't thorns and the head has color White then it will be labeled as Daisy.

Now, let's say that after training the data, there is a new separate flower (say Rose) from the bunch and you need to ask the machine to identify it.

Since your machine has already learned things, it needs to use that knowledge.

The machine will classify the flower regarding the presence (or absence of thorns) and color and would label the flower name like Rose.

This is how machines learn from training data (the bunch of flowers in our case) and then use the knowledge to label data.

That is why the process is widely known as machine learning.

Nowadays, supervised machine learning is the more common method that has applications in a wide variety of industries where data mining is used.

Types of supervised learning algorithms:

Supervised learning techniques can be grouped into 2 types:

- Regression – we have regression problem when the output variables are continuous (to know what they mean see this post "Refer link 1" *). The output variable is a real value, such as "euros" or "height".

- Classification – machine learning classification algorithms("Refer link 2" *) are at the heart of a vast number of data mining problems and tasks. Classification can be used only for simple data such as nominal data, categorical data, and some numerical variables (see these posts "Refer link 3" * and "Refer link 4" *). There are 2 types of classification: binary and multi- classification. Binary classification is when there are only two possible outcomes (such as Yes/ No). For example, binary problem is to predict if a student will pass or fail. There are only two possible answers. Multi-classification is when there are more possible results such as "red", "green", "yellow",and "white".

Here is a list of common supervised machine learning algorithms:

- Decision Trees
- K Nearest Neighbors
- Linear SVC (Support vector Classifier)
- Logistic Regression
- Naive Bayes
- Neural Networks
- Linear Regression
- Support Vector Regression (SVR)
- Regression Trees (e.g. Random Forest)
- Gradient boosting
- Fisher linear discriminant.

Advantages and Disadvantages of Supervised Learning

Advantages:

- It allows you to be very specific about the definition of the labels. In other words, you can train the algorithm to distinguish different classes where you can set an ideal decision boundary.
- You are able to determine the number of classes you want to have.
- The input data is very well known and is labeled.
- The results produced by the supervised method are more accurate and reliable in comparison to the results produced by the unsupervised techniques of machine learning. This is mainly because the input data in the supervised algorithm is well known and labeled. This is a key difference between supervised and unsupervised learning.
- The answers in the analysis and the output of your algorithm are likely to be known due to that all the classes used are known.

Disadvantages:

- Supervised learning can be a complex method in comparison with the unsupervised method. The key reason is that you have to understand very well and label the inputs in supervised learning.
- It doesn' take place in real time while the unsupervised learning is about the real time. This is also a major difference between supervised and unsupervised learning. Supervised machine learning uses of-line analysis.
- It is needed a lot of computation time for training.
- If you have a dynamic big and growing data, you are not sure of the labels to predefine the rules. This can be a real challenge.

Moreover, the pros and or cons of supervised machine learning highly depend on what exactly supervised learning algorithm you use.

Some examples of supervised learning applications include:

- In finance and banking for credit card fraud detection (fraud, not fraud). In fact, supervised learning provides some of the greatest anomaly detection algorithms.
- Email spam detection (spam, not spam).
- In marketing area – a range of text mining algorithms are used for text sentiment analysis (happy, not happy).
- In medicine, for predicting patient risk (such as high-risk patient, low-risk patient) or for predicting the probability of congestive heart failure.

What is Unsupervised Learning?

Unsupervised Learning works things out without using predefined labels. The unsupervised machine learning algorithms act without human guidance.

The task of the machine is to sort ungrouped information according to some similarities and differences without any previous training of data.

The machine is expected to find the hidden patterns and structure in unlabeled data by their own. That's why it is called unsupervised – there is no supervisor to teach the machine what is right and what is wrong.

Unsupervised learning has two categories of algorithms:

- **Clustering** - Clustering is the assignment of a set of objects into subsets (also called clusters) so that objects in the same cluster have similar characteristics in some sense. The goal of clustering is to segregate groups with similar characteristics and then assign them into clusters. A good example here is when you want to group customers by their purchasing behavior. Analytics Vidhya has a great introduction to clustering that will help you to understand better the whole idea.

- **Association** - Here we have association rules that aim to find associations amongst data objects within large databases. Association is about discovering some interesting relationships between variables in large databases. For example, people that buy a new house also tend to buy new furniture.

List of key unsupervised machine learning algorithms and techniques:

- K-means clustering
- K-NN (k nearest neighbors)
- Dimensionality Reduction
- Neural networks / Deep Learning
- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis
- Distribution models
- Hierarchical clustering
- Mixture models

Advantages and Disadvantages of Unsupervised Learning

Again here, the pros and or cons of unsupervised machine learning depend on what exactly unsupervised learning algorithms you need to use.

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.

Some examples of unsupervised learning applications are:

- In marketing segmentation, when a company wants to segment its customers to better adjust products and offerings.
- Social network analysis.
- Image Segmentation.

- Anomaly detection and etc.

Conclusion

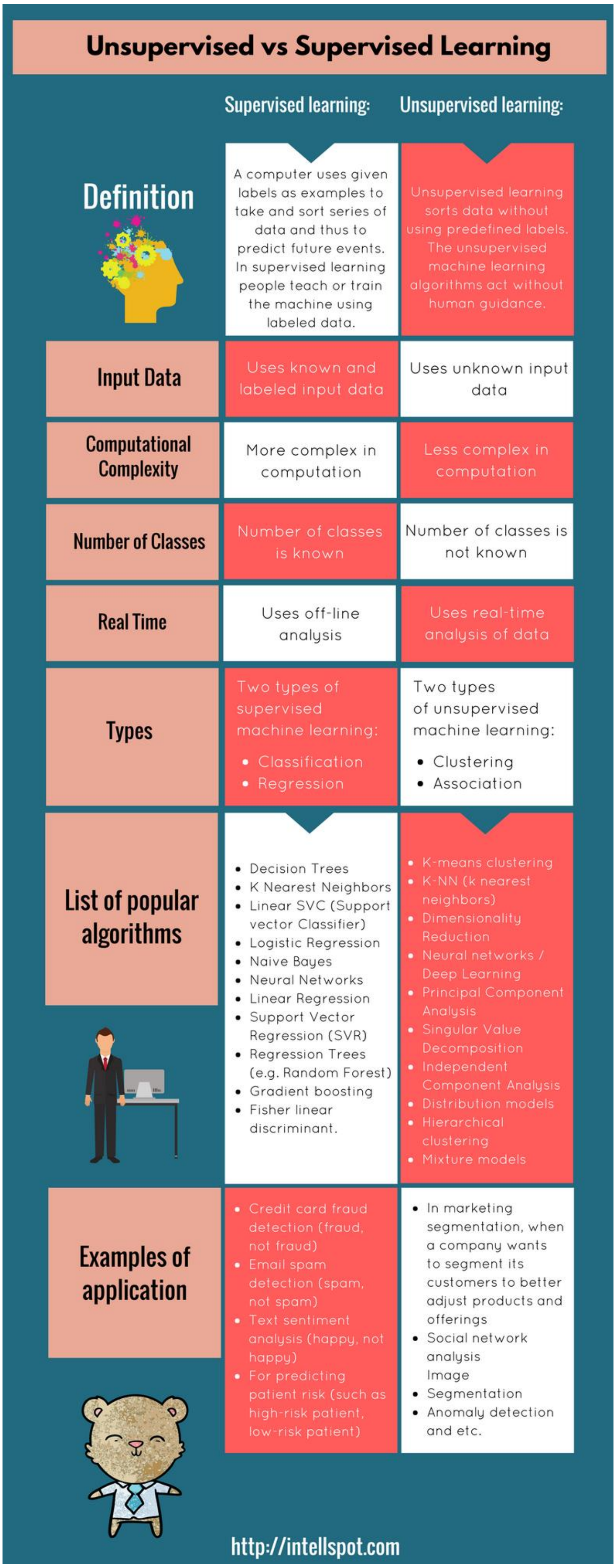
which is better supervised or unsupervised learning?

Despite we outlined the benefits and the disadvantages of supervised and unsupervised learning, it is not much accurate to say that one of those methods have more advantages than the other.

With this in mind, it’s not right to say that unsupervised and supervised methods are alternatives to each other.

The basic tasks and problems you can resolve with supervised and unsupervised methods are different. When to use the one or the other methods, depends on your needs and the problems you have to solve.

They are not only one of the hottest data science topics but also has a crucial role in data driven decision making. And as you already know, data driven decisions lead to more successful business results.

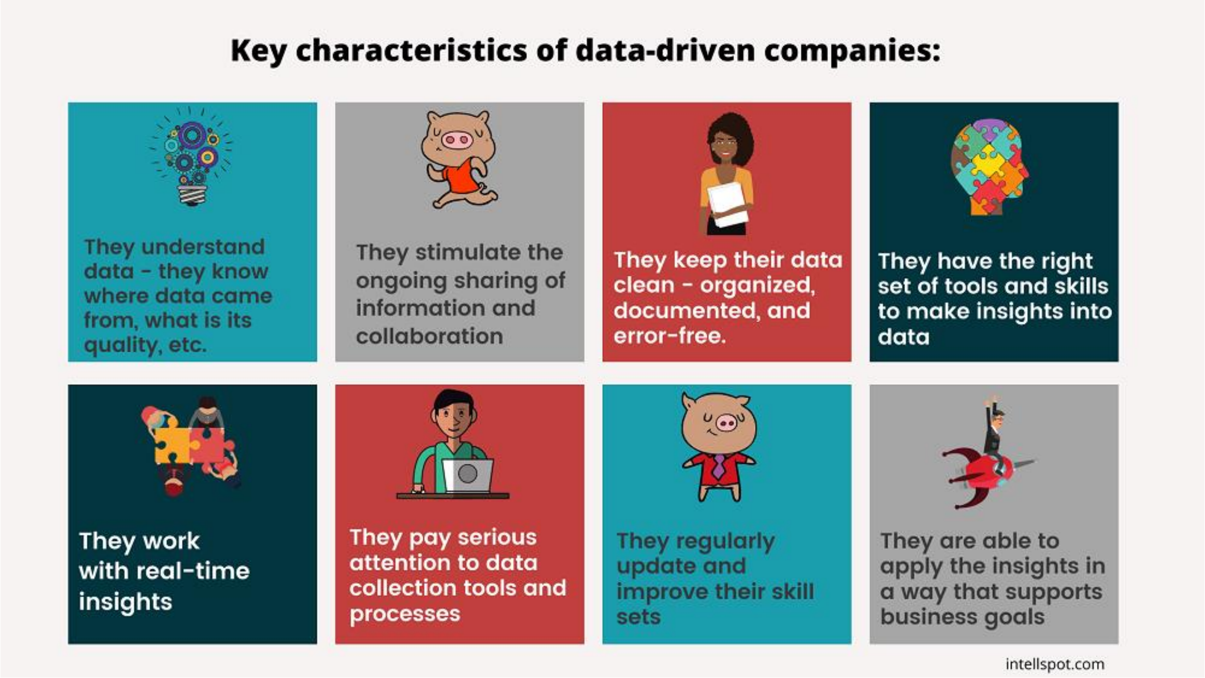


Unsupervised vs Supervised Learning

	Supervised learning:	Unsupervised learning:
Definition	A computer uses given labels as examples to take and sort series of data and thus to predict future events. In supervised learning people teach or train the machine using labeled data.	Unsupervised learning sorts data without using predefined labels. The unsupervised machine learning algorithms act without human guidance.
Input Data	Uses known and labeled input data	Uses unknown input data
Computational Complexity	More complex in computation	Less complex in computation
Number of Classes	Number of classes is known	Number of classes is not known
Real Time	Uses off-line analysis	Uses real-time analysis of data
Types	Two types of supervised machine learning: • Classification • Regression	Two types of unsupervised machine learning: • Clustering • Association
List of popular algorithms	• Decision Trees • K Nearest Neighbors • Linear SVC (Support vector Classifier) • Logistic Regression • Naive Bayes • Neural Networks • Linear Regression • Support Vector Regression (SVR) • Regression Trees (e.g. Random Forest) • Gradient boosting • Fisher linear discriminant.	• K-means clustering • K-NN (k nearest neighbors) • Dimensionality Reduction • Neural networks / Deep Learning • Principal Component Analysis • Singular Value Decomposition • Independent Component Analysis • Distribution models • Hierarchical clustering • Mixture models
Examples of application	• Credit card fraud detection (fraud, not fraud) • Email spam detection (spam, not spam) • Text sentiment analysis (happy, not happy) • For predicting patient risk (such as high-risk patient, low-risk patient)	• In marketing segmentation, when a company wants to segment its customers to better adjust products and offerings • Social network analysis Image • Segmentation • Anomaly detection and etc.

http://intellspot.com

data driven decisions



Key characteristics of data-driven companies:

They understand data - they know where data came from, what is its quality, etc.

They stimulate the ongoing sharing of information and collaboration

They keep their data clean - organized, documented, and error-free.

They have the right set of tools and skills to make insights into data

They work with real-time insights

They pay serious attention to data collection tools and processes

They regularly update and improve their skill sets

They are able to apply the insights in a way that supports business goals

intellspot.com

Thank You

!!!!!!!!!!