#4 Unsupervised Learning

Once we address supervised learning problems, we will understand how to solve ML problems with unsupervised learning. Many times, in machine learning we don't have labels related to our X. And in fact, much of the ML is about this, unsupervised learning.

This week we will see:

- **Clustering (PCA, k-means)**
- **Dimension Reduction**

We enter a series of algorithms not as intuitive as the previous ones but which are one of the most important parts together with the analysis and exploration of data.

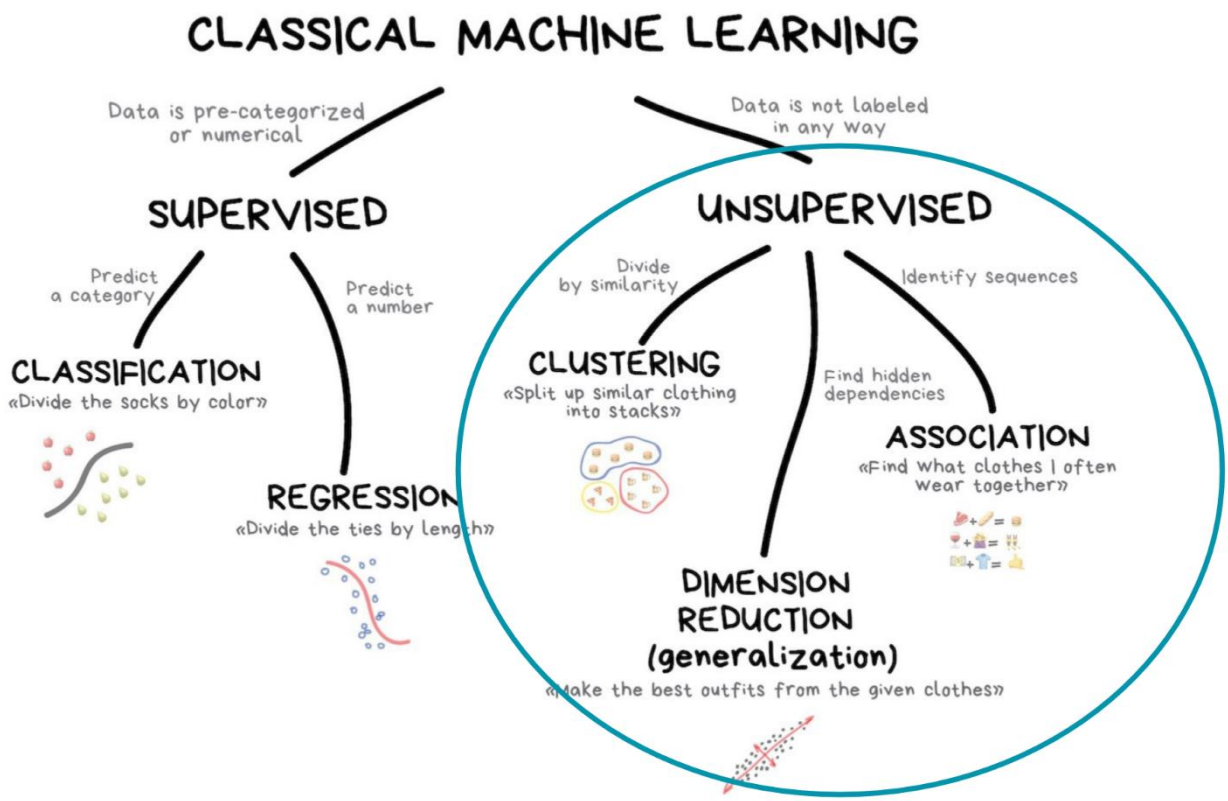
As a guide:

1. We have left 3 sets of videos to have a global unsupervised learning, what is a PCA, dimensionality reduction, etc ...
2. In addition, I have attached an introductory notebook for you to run and study during this week.
3. If you have the Hands-On or the Hundred Page ML Book, I recommend studying this topic there.
4. If you want to practice, we leave you an exercise attached too.
5. *If someone wants to go into this topic in depth, I attach a zip with some notebooks taken from MLcourse.ai if you want to know how to use PCA to compress files, MNIST, preprocessed, etc ...*

Let's start!

Unsupervised Learning:

Welcome! We are going to delve into the topic of unsupervised learning!



Unsupervised Learning is the process of identifying patterns in a dataset. Identifying patterns is often an early step in understanding data. Unsupervised learning methods are a set of techniques designed to explore and find "hidden structure" rather than predict outcomes.

Unsupervised learning does not require labeled data, therefore works for broader range of data. In fact, most data in the world is unlabeled. However, since there are no labels / correct answers there is not always a clear feedback to validate that the results are correct.

There are three main techniques in the domain of unsupervised learning:

- **Dimensionality Reduction** - Some datasets have too many features causing problems with over-fitting, slow model fitting time and issues with metric interpretability (look up the Curse of Dimensionality!). For this reason, we look for methods to reduce the number of features used to train the model while maintaining most of the variance/signal in the data.

- **Clustering** - Clustering is relatively self-explanatory. These are methods which divide the dataset into subgroups based on similar characteristics. These sub-groups can be then be used in further supervised learning algorithms or act as an intuitive way to understand the natural subsets in your dataset. Clustering is sometimes referred to as data segmentation or data partitioning.

Clustering:

PCA:

1- Standardize the data
2- Build covariance matrix
3- Calculate the Eigenvectors and Eigenvalues
4- Compute Principal components
5- Reduce the data dimensions

- PCA step by step Irises
- PCA explained notebook

https://youtu.be/FgakZw6K1QQ

https://youtu.be/hJZHcmJBk1o

Kmeans:

- Kmeans explained notebook

https://youtu.be/yR7k19YBqiw

Dimensionality Reduction:

- Interactive introduction to dimensionality reduction: A comprehensive introduction to three dimensionality reduction methods, PCA, LDA and t-SNE from kaggle. Interactive examples with code are provided so that you can see the impact of these methods on the features.

Further reading:

- https://medium.com/machine-learning-for-humans/unsupervised-learning-f45587588294
- https://blogs.oracle.com/meena/anomaly-detection

RESOURCES

unsupervised.zip

intro_unsupervised_learning.zip

assignment07_unsupervised_learning.ipynb