Practical Business Python

Lecture 13: Intro to Machine Learning in Python (2)

Iegor Vyshnevskyi

Woosong University

December 07, 2023

Agenda

1. Unsupervised Machine Learning
2. Unsupervised Machine Learning in Practice
3. In-class Assignment
4. Final Exam instructions

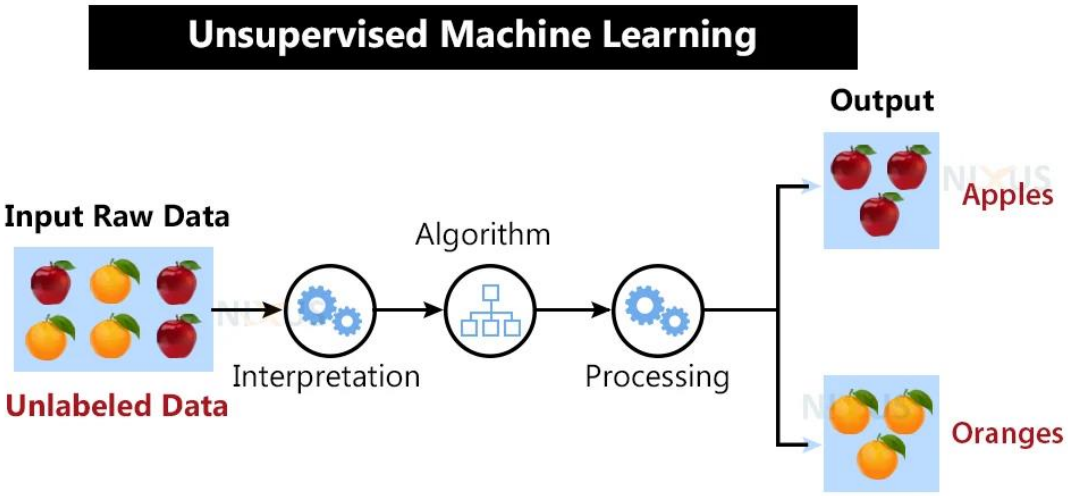
Acknowledgment: Used a number of open sources and materials from the web.

1. Unsupervised Machine Learning

What is ***Unsupervised ML?***

Unsupervised learning, also known as *unsupervised machine learning*, uses machine learning algorithms to analyze and cluster unlabeled datasets.

These algorithms discover hidden patterns or data groupings without the need for human intervention.



Why Unsupervised ML?

Here, are prime reasons for using Unsupervised Learning in Machine Learning:

- Unsupervised machine learning finds all kind of unknown patterns in data and it helps identify anomalies and outliers.
- Unsupervised machine learning helps identify the data structure and reduce data daimonions for better visualization.
- Unsupervised methods help you to find features which can be useful for categorization.
- It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.
- It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention.

*What are **Unsupervised ML techniques?** (cont.)*

Unsupervised ML generally differentiates between

- **Clustering**, where the goal is to find homogeneous subgroups within the data; the grouping is based on distance between observations.

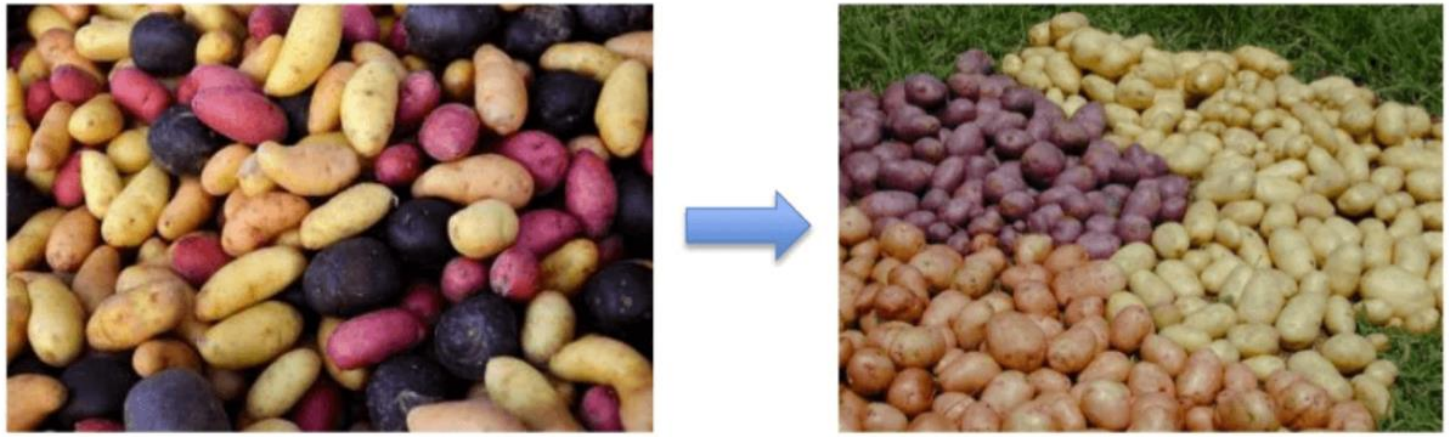
- **Dimensionality reduction**, where the goal is to identify patterns in the features of the data. Dimensionality reduction is often used to facilitate visualisation of the data, as well as a pre-processing method before supervised learning (*can be under both supervised and unsupervised ML*).

Unsupervised ML presents specific challenges and benefits:

- there is no single goal in UML
- there is generally much more unlabelled data available than labelled data.

Clustering

- Clustering is an important concept when it comes to unsupervised learning.
- It mainly deals with finding a structure or pattern in a collection of uncategorized data.
- Unsupervised Learning Clustering algorithms will process your data and find natural clusters(groups) if they exist in the data.
- You can also modify how many clusters your algorithms should identify.
- It allows you to adjust the granularity of these groups.



sample

Cluster/group

Clustering Types

Following are the clustering types of Unsupervised ML:

- K-mean Clustering: The k-means clustering algorithms aims at partitioning n observations into a fixed number of k clusters. The algorithm will find homogeneous clusters.

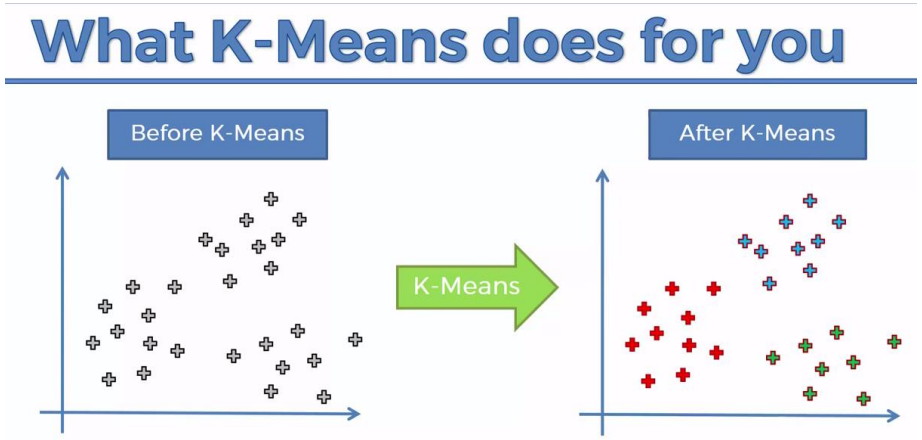
- Hierarchical Clustering: Hierarchical clustering is an algorithm which builds a hierarchy of clusters. It begins with all the data which is assigned to a cluster of their own. Here, two close cluster are going to be in the same cluster. This algorithm ends when there is only one cluster left.

2. Unsupervised Machine Learning in Practice

What we do today

K-means clustering

1. It starts with K as the input which is how many clusters you want to find. Place K centroids in random locations in your space.
2. Now, using the euclidean distance between data points and centroids, assign each data point to the cluster which is close to it.
3. Recalculate the cluster centers as a mean of data points assigned to it.
4. Repeat 2 and 3 until no further changes occur.



Example from: https://towardsdatascience.com/clustering-unsupervised-learning-788b215b074b

Example Dataset

Edgar Anderson's Iris Data

- From the *iris* manual page:
  - This famous (Fisher's or Anderson's) *iris* data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.
  - K-means Clustering is used with *unlabeled data*, but in this case, we have a labeled dataset so we have to use the iris data without the Species column. In this way, algorithm will cluster the data and we will be able to compare the predicted results with the original results, getting the accuracy of the model.

3. In-class Assignment