DELHI TECHNOLOGICAL UNIVERSITY

DEPARTMENT OF APPLIED PHYSICS

EP-208 COMPUTATIONAL METHODS

Image Compression using K-Means

Created by

Sakshi Arora
2K19/EP/084

Pranay Khosla
2K19/EP/073

Under the Guidance of

Dr. Ajeet Kumar
Department of Applied Physics

May 22, 2021

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Dr. Ajeet Kumar sir, for providing guidance in the subject of Computational Methods.

We would also like to thank the Applied Physics department and the Delhi Technological University for giving us an opportunity to present this project.

Working on this project was a great learning experience for us as we worked on the python programming language and also learned about unsupervised machine learning algorithms and explored their scope in the future.

CONTENTS

ABSTRACT

In this project, we learn about K-Means, which is an unsupervised clustering algorithm. We look at the algorithm and develop a working python code for it.

We have implemented the K-means clustering algorithm and applied it to compress an image (by reducing the number of colors). We have used the Python programming language for this purpose.

We have also looked at some applications of unsupervised learning and real life cases of the K-Means clustering algorithm.

Files used in the Project

compression.py - Python script to perform image compression

choose_k_random_centroids.py - To initialize K random centroids

k_means.py - Main script to run the K-Means algorithm

find_closest_centroids.py - To find closest centroids to each point

dist_squared.py - To compute squared distance between two points

compute_centroids.py - To update the centroids

WHAT IS LEARNING?

Machine Learning

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies existing in today's world. As it is evident from its name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

In 1997, Tom Mitchell gave a well-posed mathematical and relational definition of Machine Learning, which is that "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

Supervised Learning

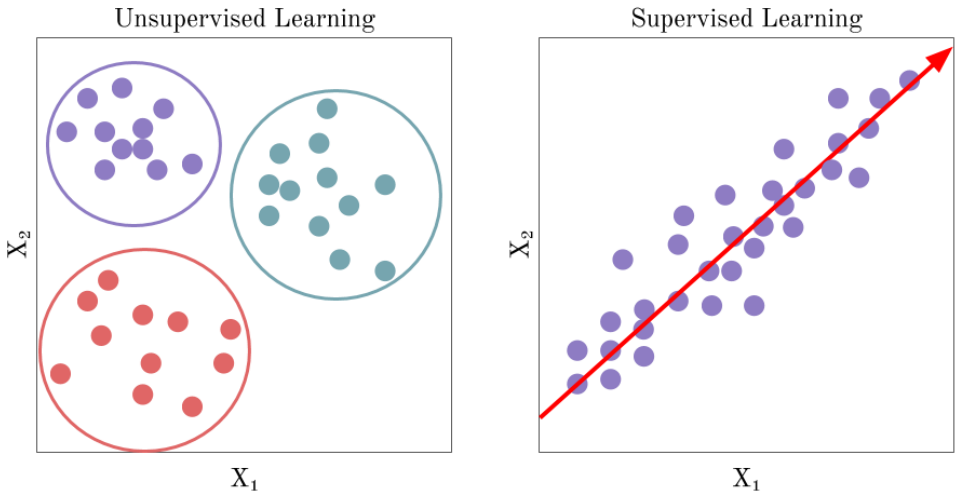
Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Formally, supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

UNSUPERVISED LEARNING

Unsupervised learning (UL) is a type of algorithm that learns patterns from untagged data. The hope is that, through mimicry, the machine is forced to build a compact internal representation of its world and then generate imaginative content. In contrast to supervised learning where data is tagged by a human, e.g. as "car" or "fish" etc, UL exhibits self-organization that captures patterns as neuronal predilections or probability densities. Two broad methods in UL are Neural Networks and Probabilistic Methods.



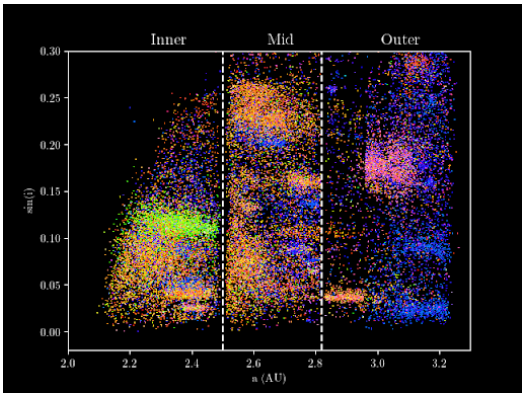
Source: towardsdatascience.com

Applications of Unsupervised Learning

- **Market Segmentation -** It is desirable to group a database of customers into different market segments to so you can sell to them separately or serve different segments better.
- **Astronomical Data Analysis -** Unsupervised learning has made it easier to understand galaxy formation, supernovae, mergers and other astronomical events.
- **Social Network Analysis -** Social networks, such as Facebook, Twitter, and LinkedIn, have greatly facilitated communication between web users around the world. The analysis of social networks helps summarizing the interests and opinions of users (nodes), discovering patterns from the interactions (links) between users, and mining the events that take place in online platforms.

Source: businessyield.com

Source: researchtoaction.org



Source: astroml.org

K-Means Clustering

Clustering

Clustering is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance.

Source: developer.squareup.com

K-Means Algorithm

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping clusters where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

Let us consider the following variables which shall be used in our algorithm:

- The number of clusters K
- Training set $x^{(1)}, x^{(2)}, ..., x^{(m)}$, where $x^{(i)} \epsilon \mathbb{R}^n$
- Set of centroids $\mu_1, \mu_2, ..., \mu_K$, where $\mu_k \epsilon \mathbb{R}^n$

The k-means algorithm works as follows:

- Specify the number of clusters.
- Initialize the centroids by randomly selecting K data points.
- Iterate over the following steps until the algorithm converges and there is no change to the centroids, i.e assignment of data points to clusters is not changing:
  - Compute the sum of the squared distance between data points and all centroids.
  - Assign each data point to the closest cluster (centroid).

$$c^{(i)} = min_k \left\| x^{(i)} - \mu_k \right\|^2$$

  - Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

$$\mu_k = \frac{1}{n}[x^{(k_1)} + x^{(k_2)} + ... + x^{(k_n)}] \epsilon \mathbb{R}^n$$

where each of $x^{(k_1)}, \ldots, x^{(k_n)}$ are the training examples assigned to group μ_k.

Optimization Objective

The cost function is defined as:

$$J(c^{(i)}, ..., c^{(m)}, \mu_1, ..., \mu_K) = \frac{1}{m} \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$$

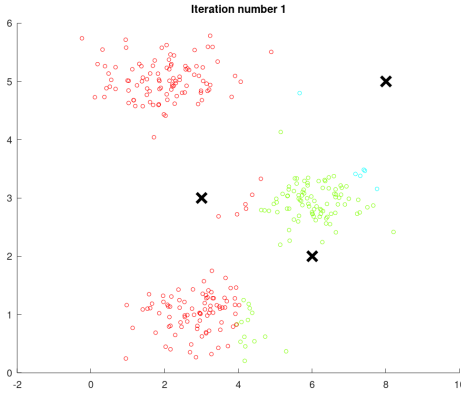
Our optimization objective is to minimize all our parameters using the above cost function:

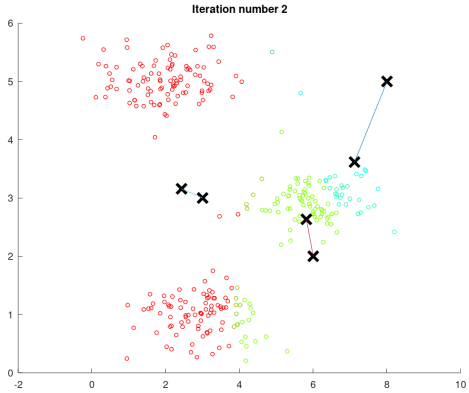
$$min_{c,\mu} J(c, \mu)$$

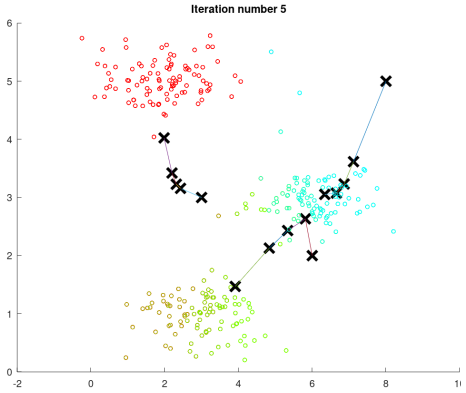
that is, we are finding all the values in sets c, representing all our clusters, and μ, representing all our centroids, that will minimize the average of the distances of every training example to its corresponding cluster centroid.

RUNNING K-MEANS ON EXAMPLE DATASET

The following images were obtained on running the algorithm on an example dataset having three clusters. At each step, the three clusters are represented by three different colors The algorithm was run upto 10 iterations.







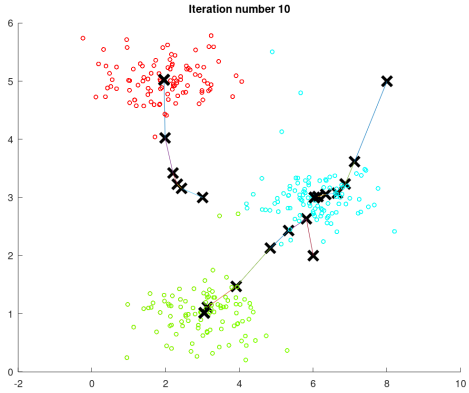


IMAGE COMPRESSION USING K-MEANS

In a straightforward representation of an image, each pixel is represented as three 8-bit unsigned integers (ranging from 0 to 255) that specify the red, green and blue intensity values. This encoding is often referred to as the RGB encoding. Our image contains thousands of colors, and we will reduce the number of colors to K, where K is input by the user.

By making this reduction, it is possible to represent (compress) the photo in an efficient way. Specifically, we only need to store the RGB values of the K selected colors, and for each pixel in the image we now need to only store the index of the color at that location. Reducing the colors in an image also imparts an animated look and can be used to make stylish images.

We will use the K-means algorithm to select the K colors that will be used to represent the compressed image. Concretely, we will treat every pixel in the original image as a data example and use the K-means algorithm to find the K colors that best group (cluster) the pixels in the 3-dimensional RGB space. Once we have computed the cluster centroids on the image, we will then use the K colors to replace the pixels in the original image with these K centroids.

We have run the K-Means Algorithm on 3 different images and compressed them by reducing the number of colors to 16 (4 bit), 8 (3 bit) and 4 (2 bit) for each image.

Original Image

Compressed, with 16 colors

Compressed, with 8 colors

Compressed, with 4 colors

Original Image

Compressed, with 16 colors

Compressed, with 8 colors

Compressed, with 4 colors

Original Image

Compressed, with 16 colors

Compressed, with 8 colors

Compressed, with 4 colors

APPLICATIONS OF K-MEANS

- **Identifying Fake News -** Fake news is being created and spread at a rapid rate due to technology innovations such as social media.

  The algorithm takes in the content of the fake news, the corpus, examines the words used and clusters them. Certain words are found more commonly in sensationalized, click-bait articles. When you see a high percentage of specific terms in an article, it gives a higher probability of the material being fake news.

- **Marketing and Sales -** Personalization and targeting in marketing is big business.

  This is achieved by looking at specific characteristics of a person and sharing campaigns with them that have been successful with other similar people.

- **Fantasy Cricket and Sports -** The challenge at the start of the season is that there is very little, if any, data available to help you identify the winning players for forming a team.

  When there is little performance data available to train the model on, unsupervised learning is used. You can find similar players using some of their characteristics. This has been done using K-Means clustering. Thus, you can get a better team more quickly at the start of the year, giving you an advantage.

- **Classifying network traffic -** For large websites, it is important to know where the traffic is coming from to be able to block harmful traffic and double down on areas driving growth.

  K-means clustering is used to group together characteristics of the traffic sources. By having precise information on traffic sources, you are able to grow your site and plan capacity effectively.

CONCLUSION

In this project, we have looked at what are supervised and unsupervised learning algorithms. We discussed some real world problems where unsupervised learning is applied.

We explore the K-Means clustering algorithm in detail and develop a python program for the same. We implement the algorithm and use it to compress an image (by reducing the number of colors). We go on to discuss some exciting real world applications of K-Means clustering.

REFERENCES

- Andrew Ng Lecture Notes, CS229 Stanford
- K-Means Clustering Algorithm, Javatpoint
- Use Cases for the K-Means Algorithm, DZone AI