An Introduction to K-means Clustering

Bárbara Bettencourt

November 2019

Abstract

A study into K-means clustering; an unsupervised machine learning problem where the goal is to have data points of similar characteristics put into different clusters. Image compression is one application of clustering where fixing a number of k colours gives an individual variation of a given picture. However, the lower the value of k the worse the image quality.

1 Introduction

Machine learning is a method of data analysis where the objective is to have computers operate without being explicitly programmed to. In other words, the goal is to have automated systems that learn from experience.
We can assign different classes to the problems in machine learning, such as:

- Supervised learning: in this type of learning we are given explicit data with correct labels. Regression and Classification are two supervised learning problems, where the Regression problem fits the given data and the Classification problem separates the data.
- Unsupervised learning: in this type of learning we are not given a training set with the correct labels. Association and Clustering are two unsupervised learning problems, the Association problem seeks to find patterns that describe large sections of the data, and the Clustering problem seeks to find permanent groupings in the data.
- Reinforced learning: in this type of learning we are not given a training set with correct labels, but the learner can obtain details about the output by communicating with the environment.

2 Theory

Now that we're familiar with all the types of machine learning, we can move on to clustering. Clustering is an unsupervised learning problem which seeks to divide the data into a number of groups such that each group has similar traits. There are multiple types of clustering algorithms such as:

- Centroid-Based clustering: "These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters" [2]. The K-means algorithm fits in this type of clustering.
- Density-Based clustering: this links regions of high density into clusters.
- Distribution-Based clustering: this assumes that the data is formed of distributions.
- Connectivity-Based clustering, also referred as Hierarchical clustering: this is based on the idea of objects being better associated with nearby objects rather than farther apart objects.

2.1 K-means Clustering

In this project we are going to be looking at the K-means algorithm in particular as it is the most popular algorithm when it comes to Centroid-based clustering. The objective of k-means is to group similar data points and to find patterns by fixing k number of clusters in a dataset. In order to demonstrate this we choose $k = 2$ clusters in a scenario with 7 points where these points are randomly assigned to a cluster. We can then find the cluster centroids; a centroid represents the center of a cluster.

Figure 1: Example of 7 points with 2 clusters.

Figure 2: Squares represent the centroids.

Next, we reassign each point to the closest centroid. In this case we reclassify two points. Finally we recompute the cluster centroids and iterate these steps until no further improvements are possible.

Figure 3: Reassign the points to closest centroid.

Figure 4: Recompute cluster centroids.

This algorithm has some notable advantages: it guarantees convergence, it is easily adaptable to examples and it scales to large data-sets. There are also some disadvantages to this algorithm such as the need to choose k manually, the lack of certainty in regards to which k to choose, and the fact that the clustering outliers can drag the centroids so they might get their own cluster rather than being ignored.
However, K-means clustering can be quite useful in real world problems, examples of this include: Detecting insurance fraud; by using past data on criminal claims we isolate new claims based on their closeness to clusters that designate criminal patterns. K-means is also helpful in identifying crime regions; this also uses past data such as the type of crime and the place where it was committed. The connection between this data can provide useful insight to law enforcement bodies about the areas that are more inclined to have crime.

2.2 Image Compression

Image compression is one implementation of clustering. It is"the type of data compression applied to digital images to reduce their cost of storage or transmission" [1]. Hence, by using K-means clustering we will join the same colour data points, in this case pixels, into separate clusters.
We can illustrate this by taking a random picture (Figure 5) and applying the K-means algorithm using k=16 and k=6 colours.

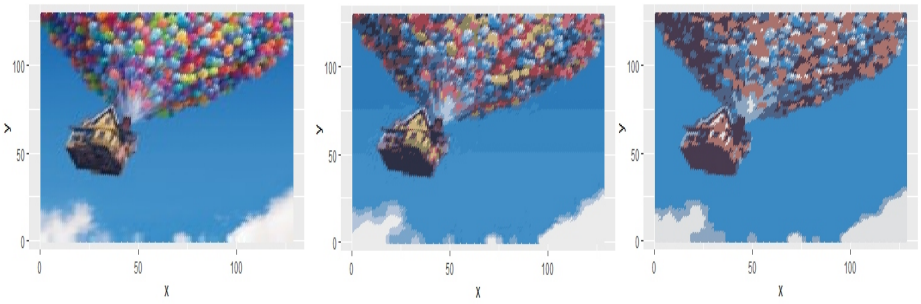

Figure 5: Original Photo. Figure 6: k=16 colours. Figure 7: k=6 colours.

There is a small variation from the original photo to the compressed photo when k=16 (Figure 6); we can see that a few colours have disappeared due to the fact that the original picture has more RGB (Red-Green-Blue) values. When comparing the original picture to the compressed photo when k=6 (Figure 7), there is a massive distinction as expected. Therefore when k is minimized, the compression becomes more dissipated. This means that the details of the picture may fade away. If we maximize k, it will lessen the dissipation of the picture.

3 Conclusion

The purpose of machine learning is to teach machines to learn from experience, and clustering is one of multiple problems, such as Classification and Association, that we encounter when studying machine learning.
The K-means algorithm is the leading procedure when faced with an unsupervised clustering problem. This procedure groups the given data points, and finds patterns by fixing k number of clusters. As discussed before, the k-means does have some flaws, nonetheless it is still an essential method when it comes to clustering. Image compression is a method using K-means to lessen the amount of storage the image takes by decreasing the number of colours. However, when compressing images it may not be ideal as the image quality will be poor when choosing a small k.

References

[1] Vibhor Agarwal. Image compression using k-means clustering, 2018 (Acessed November 16, 2019).

[2] Saurac Kaushik. An introduction to clustering and different methods of clustering, 2016 (Accessed November 15, 2019).

Appendices

A R-code

```
image.path<-paste(getwd(),"/up2.jpg",sep="")
img <- readJPEG(image.path)

imgDm <- dim(img)

imgRGB <- data.frame(
  x = rep(1:imgDm[2], each = imgDm[1]),
  y = rep(imgDm[1]:1, imgDm[2]),
  R = as.vector(img[,,1]),
  G = as.vector(img[,,2]),
  B = as.vector(img[,,3])
)
plot1<-ggplot(data = imgRGB, aes(x = x, y = y)) +
geom_point(colour = rgb(imgRGB[c("R", "G", "B")]))
+ labs(title = "Original Image")

k <- 16
kMeans <- kmeans(imgRGB[, c("R", "G", "B")], centers = k)
num.of.colours <- rgb(kMeans$centers[kMeans$cluster,])

plot2<-ggplot(data = imgRGB, aes(x = x, y = y))
+ geom_point(colour = num.of.colours)
+ labs(title = paste("k-Means Clustering of", k, "Colours"))

grid.arrange(plot1, plot2, nrow=2)
```