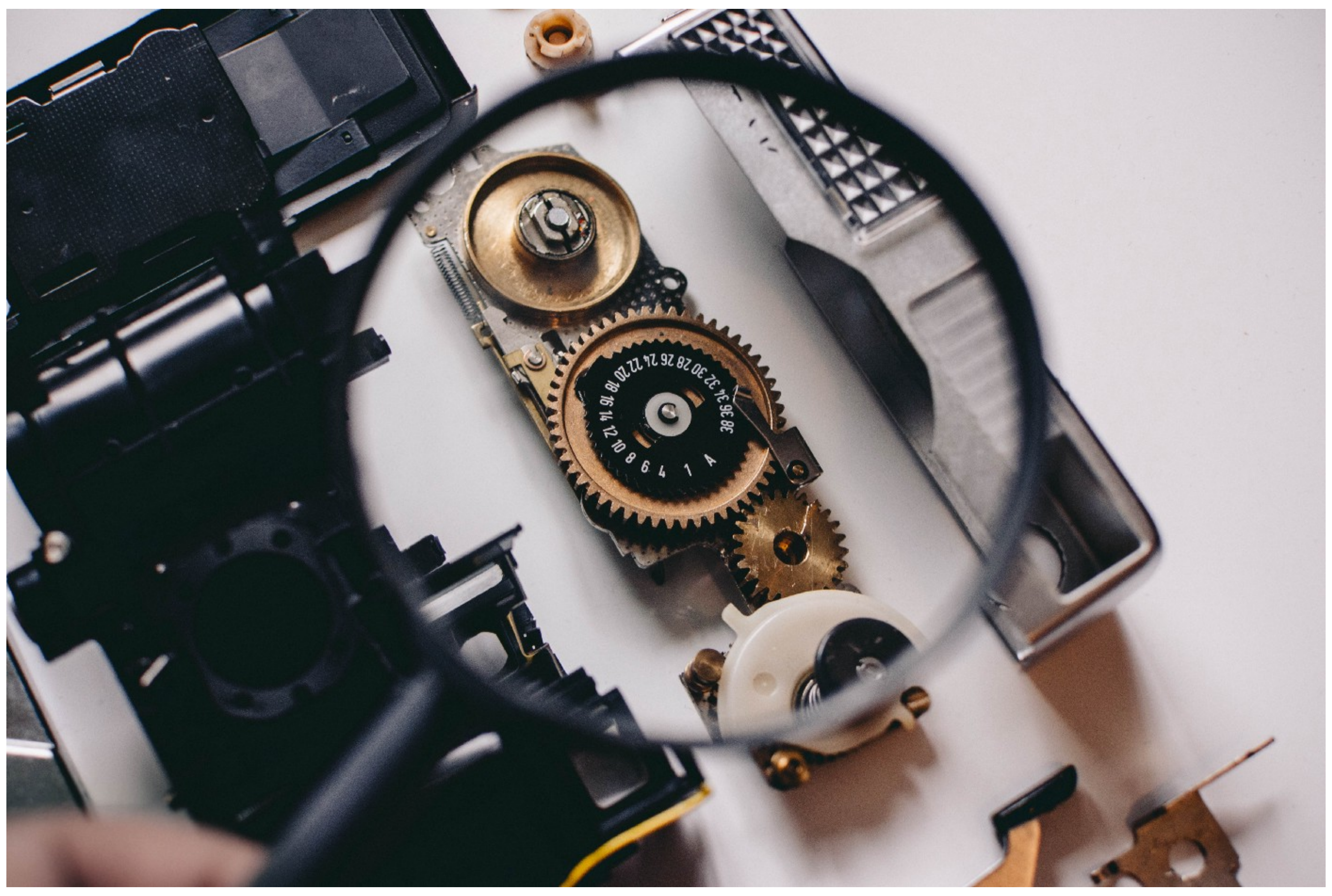
Published in Towards Data Science

Eugenio Zuccarelli
Jun 7, 2020 · 4 min read ★ · Listen



Interpretable Clustering

How to use CART to take the guesswork out of describing your clusters

Clustering algorithms such as K-Means, Agglomerative Clustering and DBSCAN are powerful unsupervised machine learning techniques. However, summarising the key characteristics of each cluster requires quite a qualitative approach, becoming a lengthy and non-rigorous process that requires domain expertise.

An often overlooked technique can be an ace up the sleeve in a data scientist's arsenal: using Decision Trees to quantitatively evaluate the characteristics of each cluster. Specifically, after having clustered the unlabelled data, we can assign to each sample the corresponding cluster as a label. We can then train a CART model using the label as target variable, and then inspecting the resulting decision tree to highlight the characteristics of the cluster.

Clustering the Data

After having processed the data accordingly, we can select the Clustering algorithm that we prefer. Usually, the options are:

- K-Means
- Agglomerative Clustering
- DBSCAN

K-Means

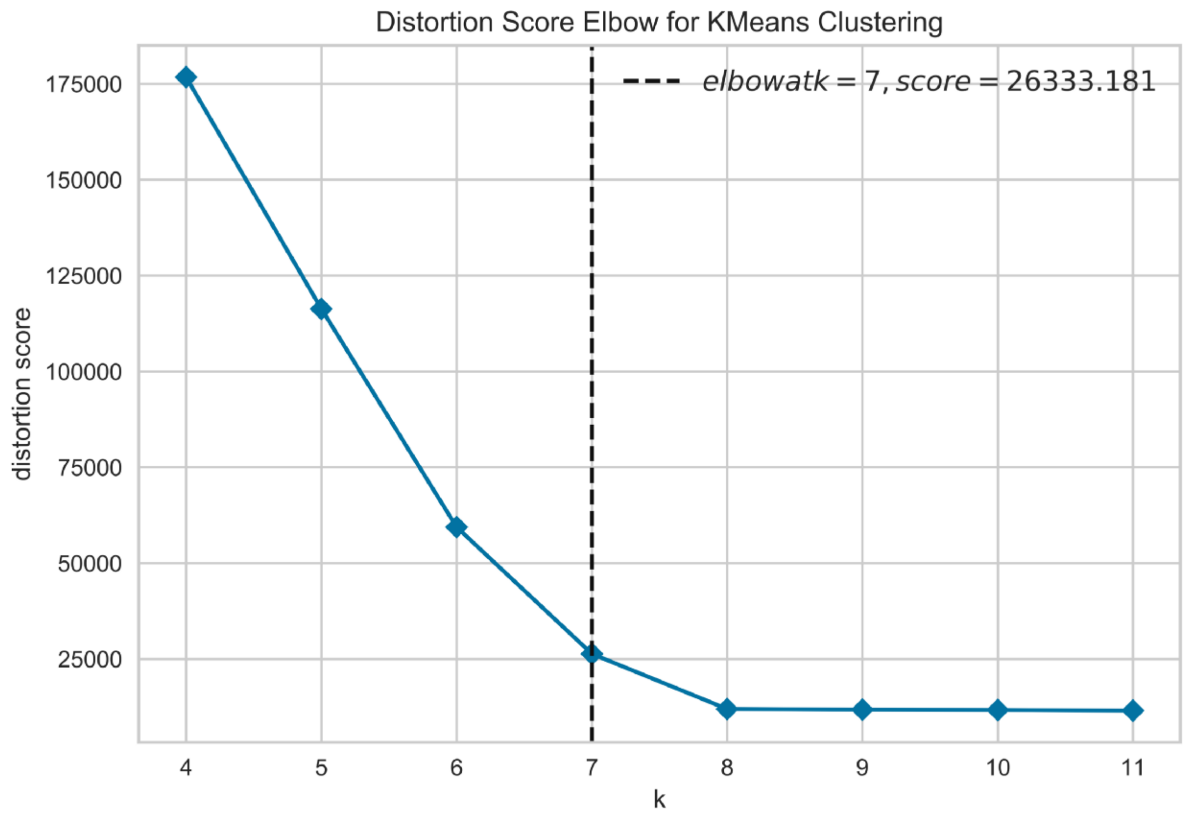
K-Means is the most common unsupervised learning technique, mostly due to its simplicity and effectiveness. It is a fast algorithm, that can run on millions of observations, but it does have its drawbacks. First of all, it does not perform well with data structures that are not spherical. It also does not perform well when the density of points is heterogeneous, meaning that in some areas of the distribution the density is higher than in others. Finally, the number of clusters K has to also be chosen, requiring a somewhat qualitative decision to be made.

While for the first two issues there is not a remedy other than choosing another algorithm, the number of clusters K can be chosen in a relatively rigorous manner. We can iteratively run the K-Means algorithm with an increasing number of K until a somewhat large number (e.g. 50). Then, we can plot a performance measure for each value of K on a chart. By inspecting the chart, we can find the "elbow", meaning where the performance peaks, for then yielding marginal returns at higher values of K. The value of K we can choose is exactly when the performance gain is marginal with respect to the increase in K. Out of

Eugenio Zuccarelli
287 Followers
Data Scientist for Fortune 100 | Fulbright Scholar | MIT '20, Harvard, Imperial College | Follow on Twitter at @JayZuccarelli and at eugeniozuccarelli.com
Follow

More from Medium
Ka... in Towards ...
Calculating Machine Learning Model...
Maz... in Python ...
Data Science Project | Clustering Mixe...
Angell... in MLe...
Data sampling methods for

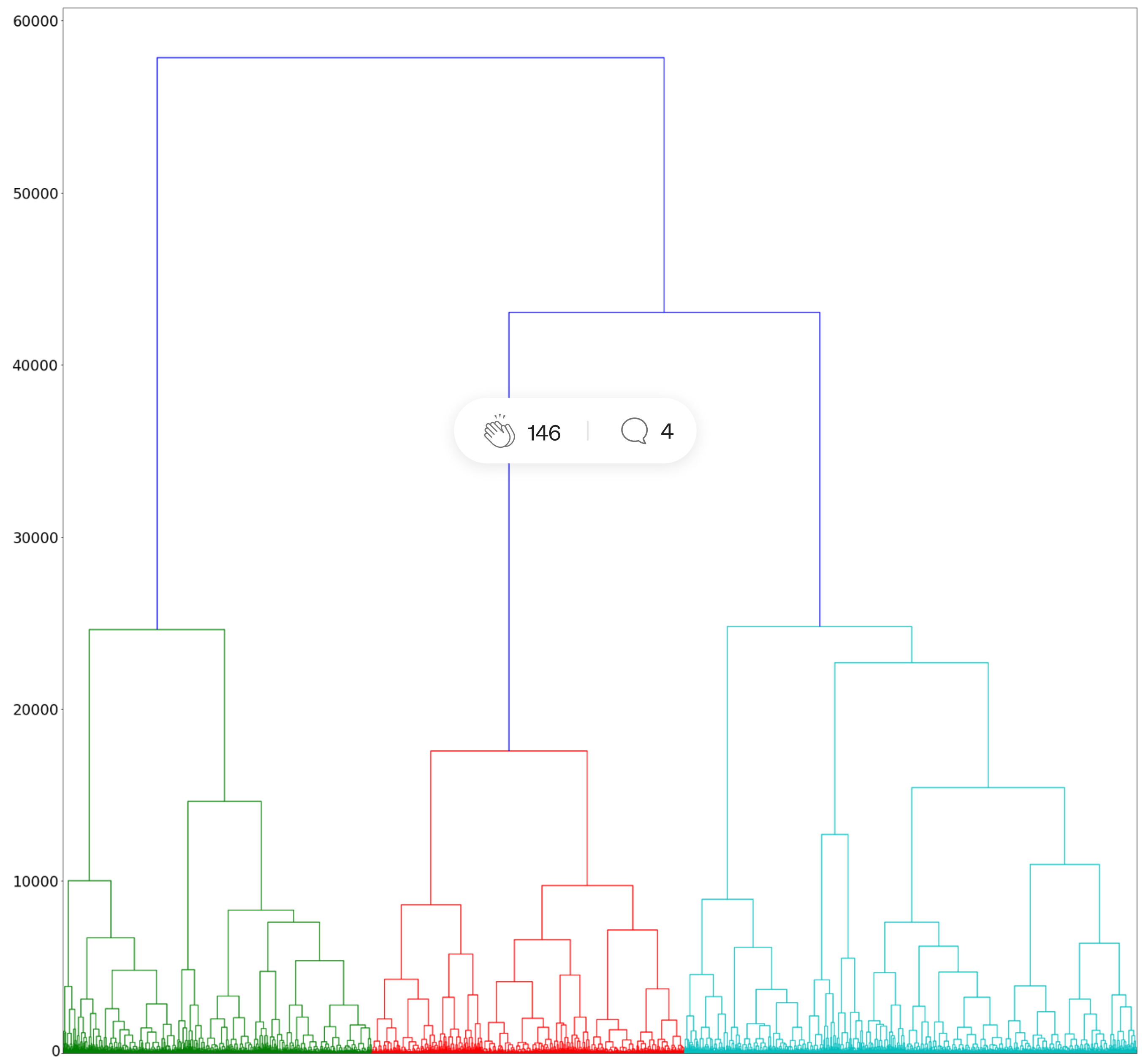
the many performance measures, the most used are usually, Silhouette Coefficient, Distortion and Gap Statistic.



Agglomerative Clustering

Agglomerative Clustering is a type of Hierarchical Clustering that iteratively groups samples together, starting from one cluster per sample to a single cluster overall. It is one of the preferred methods when performing clustering because it allows to quickly select the optimal value of K. Indeed, we can plot the dendrogram, a tree structure showing on the x-axis the samples groupings, while on the y-axis the information gain from additional groupings.

Similarly to the Elbow Analysis, we can simply select the number of clusters where the information gained from an ulterior split is marginally increasing, and therefore not worth the additional complexity.



146
4

DBSCAN

Finally, more for reference, DBSCAN is another technique widely used for Clustering analyses. It works by inspecting the density of the nearby samples, assigning samples to the cluster if the density is sufficiently high. However, since Agglomerative Clustering is already a successful technique, DBSCAN is less common in practice.

Interpreting the Clusters

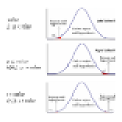
Now that we have clustered the unlabelled data, we can extract the cluster information and assign it to each sample as a label.

Using Agglomerative Clustering, for instance, we can first select the number of clusters K we chose…

imbalanced data

Renu Khandelwal

Hypothesis Testing: An Intuitive…



```
cut = scipy.cluster.hierarchy.cut_tree(Z, n_clusters=K)
```

...and then extract the labels to assign them to the data:

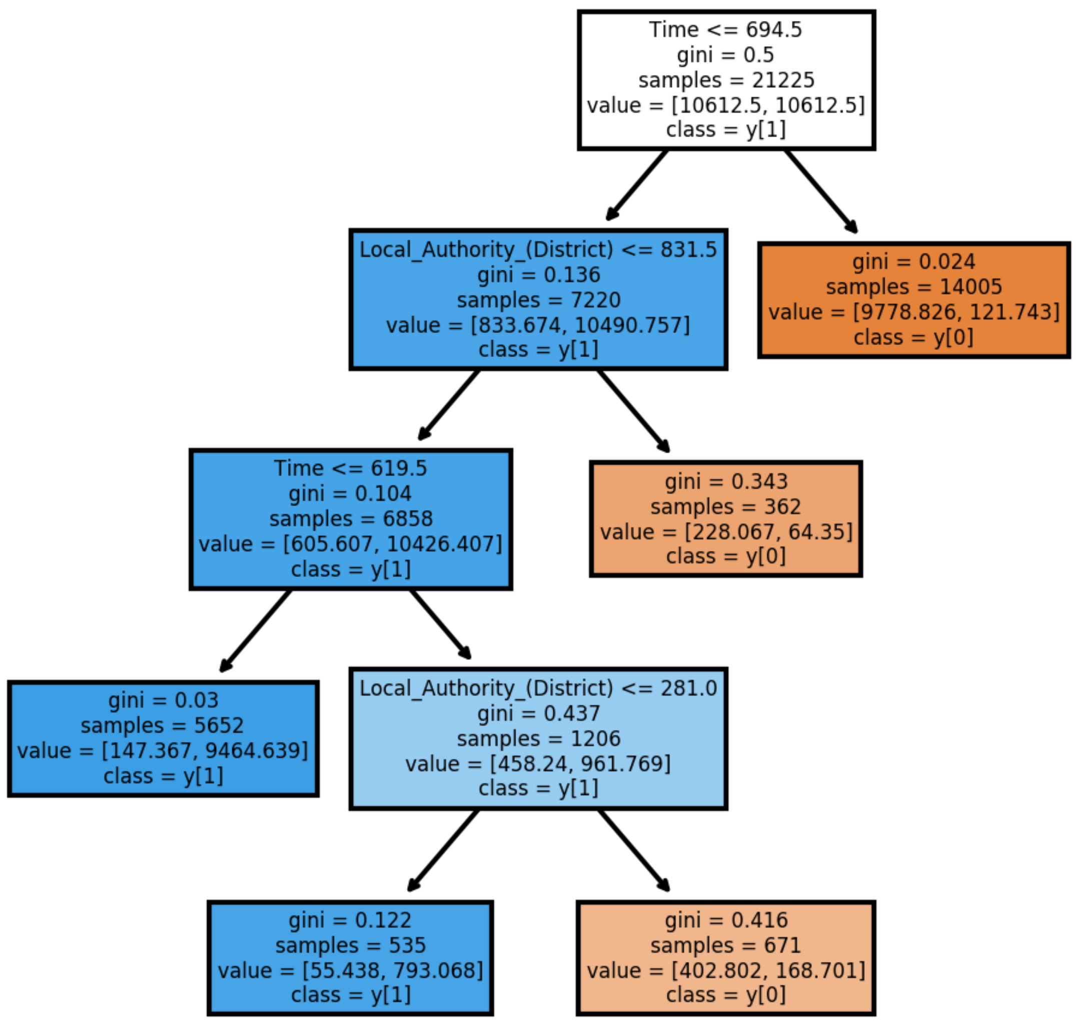
```
labels = list([i[0] for i in cut])
labeled_data = pd.DataFrame(data, columns=data_columns)
labeled_data['label'] = labels
```

Finally, we can proceed to train a Decision Tree using the labelled dataset as training. To avoid having one massive tree, I would recommend structuring this as a binary classification problem where y is equal to 1 if a point is in the selected cluster and 0 otherwise. This would be repeated for each cluster, giving you different simpler decision trees which are far easier to understand, and later present. Specifically, we can plot the Decision Tree, using:

```
fig, axes = mp.subplots(nrows = 1,
                        ncols = 1,
                        figsize = (4,4),
                        dpi=300)

sklearn.tree.plot_tree(model,
                       feature_names = X.columns,
                       filled = True,
                       class_names=True);
```

By inspecting the Decision Tree, we can highlight the characteristics of the corresponding cluster. For instance, we might be seeing that the observations get assigned to such cluster only under some conditions.



Summary

We can shed light on Clustering, by combining unsupervised and supervised learning techniques. Specifically, we can:

- First, cluster the unlabelled data with K-Means, Agglomerative Clustering or DBSCAN
- Then, we can choose the number of clusters K to use
- We assign the label to each sample, making it a supervised learning task
- We train a Decision Tree model
- Finally, we inspect the Decision Tree's output to quantitatively highlight the characteristics of the cluster

References

A recent interesting work has been completed by some colleagues at MIT, using Optimization to make Clustering interpretable. You can read more about it here.

To read more articles like this, follow me on Twitter, LinkedIn or my Website.