# **K-Mean Clustering and Ethical Use of Public Data**

Krishna Kumar Veeraputhiran

Grand Canyon University

DSC 540 – O500 – Machine Learning for Data Science

Dr. Aiman Darwiche

September 22,2021

# **K-Mean Clustering and Ethical Use of Public Data**

In this article, we will discuss about the implementation of K-Mean clustering algorithm on a publicly available dataset. We will discuss on the key questions that need to be asked and answered before the implementation of the algorithm. Since we are using data from a public database for our study, we will discuss on the ethical aspect associated with it.

**About the Data**

The data for this article is the Wine dataset obtained from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/wine>). This dataset is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivators. This dataset has 13 dimensional variables describing the wine and one wine class identifier (label) variable that indicates from which cultivator it is obtained from (we will not be using this label in our model). There are 178 wine samples available in this dataset.

**K-Means Clustering**

For our report, we need to identify patterns in the wine data without a set label. This process of identifying patterns in an unlabeled data and grouping them is called clustering. One of the most common unsupervised algorithm that can be used for clustering is the k-mean algorithm. The goal of k-means algorithms is to obtain a) K centroids representing the centre of the clusters b) Labels for the training data.

**Questions to Ask before Modeling**

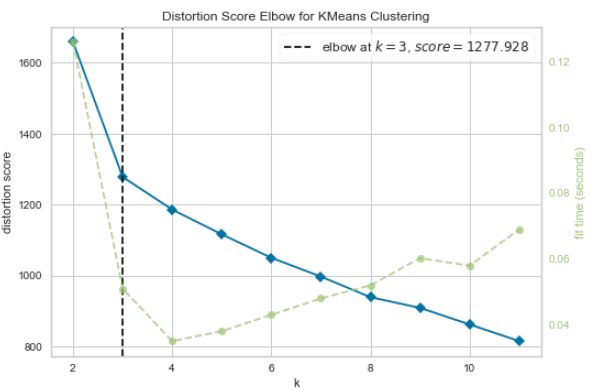
Before we build the k-mean model and fit our wine dataset, we need to ask some questions about the data. This helps us to understand the data in a better way and create a model that is efficient and predicts the outcome more precisely. Some of the key questions that can be asked while looking at the dataset are a) What is the purpose of grouping the data into clusters using the k-mean model? b) How many groups should we split the dataset into? c) What are all the dimensional variable should one use to classify the dataset into a given set of clusters? Answering above set of questions will help us in building a better cluster model.

**Purpose of Clustering**

K-Mean clustering is an Unsupervised learning algorithm that we will use on the Wine data set. As mentioned earlier, the wine dataset doesn’t have a model classifying what kind of wine each wine is. Other than the wine properties we do not have any additional information about the data. One of the strengths of k-mean clustering is to identify patterns among the data and use these commonly identified patterns in the data to classify them. In our case identifying the patterns helps us to classify the wines and this information can be used by the cultivators to grade their wine and fix a price for their wine based on these patterns. Also grouping similar group of wine in our case helps us to classify a new wine cultivated can be classified properly based on its properties in the future.

**Number of Clusters**

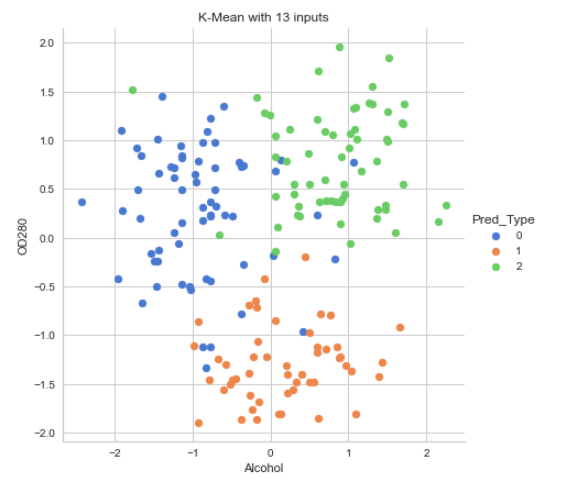
One key question that need to be answered before performing a k-mean clustering is how many clusters we need to define on the dataset. This is a crucial step before we perform the clustering algorithm. Defining wrong number of clusters can lead to less informatics clustering. Selecting fewer clusters can cause some disjoint groups of data to be forced to fit in one large cluster. Also creating too many clusters can lead to creating artificial boundaries within real data cluster. So it is crucial to identify the optimal number of clusters for our dataset before we perform our k-mean clustering. There are two ways to identify the optimal number of clusters. One is the Elbow method and the other is the Silhouette score method. In our program we will use the Elbow method to identify the optimal number of clusters. Elbow method calculates the Within cluster sum of squares (WCSS) for various k (cluster number). The sum of squares is greatest when k=1. As the k value increases, the distance decreases rapidly. For a certain k value the distance changes drastically creating an elbow shape when plotted in the graph. The k value corresponding to this is selected as an optimal cluster. Below is the elbow plot for our wine dataset.



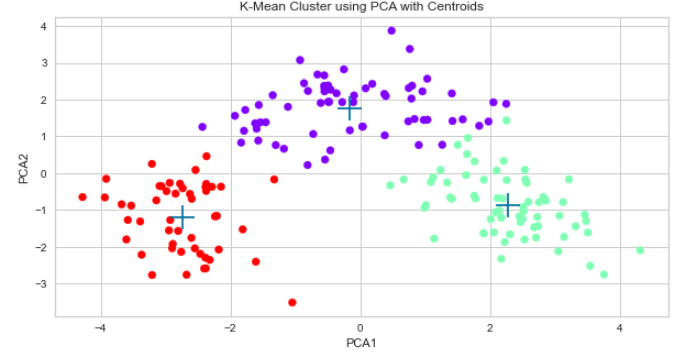
We have the distortion on the x-axis, which is the sum of squared distance from each point to its assigned centers and the y-axis gives the number of clusters. As we can see at k=3, there is a sharp change in the distortion score. So for our case we will select the total number of clusters to be 3.

**Selecting Dimensional Variables**

In our dataset there are 13 different dimensional input parameters that can be used for clustering. One problem for dataset with high dimension is that we run the risk of massively overfitting our model, which would lead to performance issue. Also too many dimension causes the data points in the wine dataset to appear equidistant from all the others. If the distance all appear equal and hence all the observation appear equally alike and no meaningful cluster can be formed. Below is how the wine dataset would be clustered with all 13 variables.



As we can see that the cluster boundaries overlap here with all 13 variables. Hence, we need to select nominal amount of input variables for our k-means algorithm. We have used Principal Component Analysis (PCA) to select the two attributes that has a higher impact on the variance. Below is the outcome of the k-mean cluster after using the two PCA’s for the algorithm.



**Ethical Use of Public Data**

For our research we have used publicly available Wine dataset from the UCI Machine learning repository. As per Cooper (2020), data that are available publicly are created and distributed by public organizations. Some of the common characteristics of Publicly-available data are provided below.

*Surrogates* – These dataset are not real world data but a closer representation of real world.

*False Precision* – These dataset are not precise enough with real world data. For example in our example the precision of the Alcohol percentage to the actual data would be different.

*Quality* – Quality of the data is depreciated and not understood by the users of this data.

*Metadata* – The Metadata of the public data is not clearly defined for the end users.

Some of the common ethical issues of these publicly-available data are bias, liability, censorship and privacy. Data are often shared globally and usage of such data is subject to diverse value system. End users who do not have the moral understanding of the data may fail to appreciate the value of the data and may also tend to push the ethical boundaries. As a data scientist we must be able to understand the data that is pulled from diverse systems and explore it by questioning the context of the data. We also need to identify any personal traits available in these dataset and wean away from using it for any analysis without proper consent of an individual or a group. This very well breaches the Privacy of the individual/group and hence lead to ethical violations.

# **References**

Lee, W.-M. (2019). *Python machine learning*. Wiley.

Yiu, T. (2019, July 20). *The curse of dimensionality*. Medium. Retrieved September 22, 2021, from https://towardsdatascience.com/the-curse-of-dimensionality-50dc6e49aa1e.

Yellowbrick. (n.d.). *Elbow method*. Elbow Method - Yellowbrick v1.3.post1 documentation. Retrieved September 22, 2021, from https://www.scikit-yb.org/en/latest/api/cluster/elbow.html.

Cooper, A. K. (2020). *Responsible design, implementation and use of information and communication technology: 19th Ifip Wg 6.11 conference on e-business, e-services, and e-society, i3e 2020, Skukuza, South Africa, April 6-8, 2020, proceedings*. Springer.