Classification of Wines using FTIR Spectrum Analysis

Winter in Data Science

Analytics Club and DAV Team, IIT Bombay

Mentors :
Samuel Raja Bethala & Vanshika Sindhu

Yash Mehta
22B1504
UID : 07

14 January, 2024

Contents

1 INTRODUCTION

Fourier Transform Infrared (FTIR) spectroscopy is a powerful analytical technique widely employed in various scientific disciplines to investigate the molecular composition of materials. This spectroscopic method utilizes the interaction of infrared radiation with matter, providing valuable insights into the vibrational modes of molecules. The key advantage of FTIR spectroscopy lies in its ability to rapidly collect a broad spectrum of information, allowing for the identification of functional groups and the characterization of chemical structures.

In FTIR spectroscopy, a sample is exposed to infrared light, leading to the absorption, transmission, or reflection of specific wavelengths corresponding to the vibrational frequencies of molecular bonds. The resulting spectrum, generated through Fourier transformation of the raw data, offers a detailed fingerprint of the sample's chemical composition. FTIR has found widespread applications in diverse fields, including chemistry, biology, pharmaceuticals, and materials science, making it an indispensable tool for qualitative and quantitative analysis.

This project deals with the usage of FTIR spectra of 37 wine samples to classify them into two classes using K-Means Clustering.

2 ABSTRACT

The dataset comprises FTIR spectra from 37 red wines, with each bottle contributing three spectra, resulting in a total of 111 spectra. This dataset was sourced from a GitHub repository [1]. The wines were sampled in triplicate, and the FTIR spectra, obtained through **Attenuated Total Reflectance** sampling in the mid-infrared range, were collected randomly over multiple days.

The assertion is that among the 37 red wines, 19 are Cabernet Sauvignon, while the remaining 18 are Shiraz. However, the specific assignment of each wine to these categories is unknown.

The objective of this project is to employ an unsupervised classification algorithm, specifically K-Means Clustering, to categorize these 37 wines into the two proposed classes. The coding for the same is done using Python programming language with the help of Numpy, Pandas, Matplotlib and Scikit-Learn packages. The Jupyter Notebook is uploaded on a Github Repository [2].

3 DATA ANALYSIS, PRE-PROCESSING AND DIMENSIONALITY REDUCTION

3.1 Data Analysis

The plots of spectra of all the samples were plotted. Three major peaks were observed. The following figure shows a plot of one of the samples, along with the three important peaks.

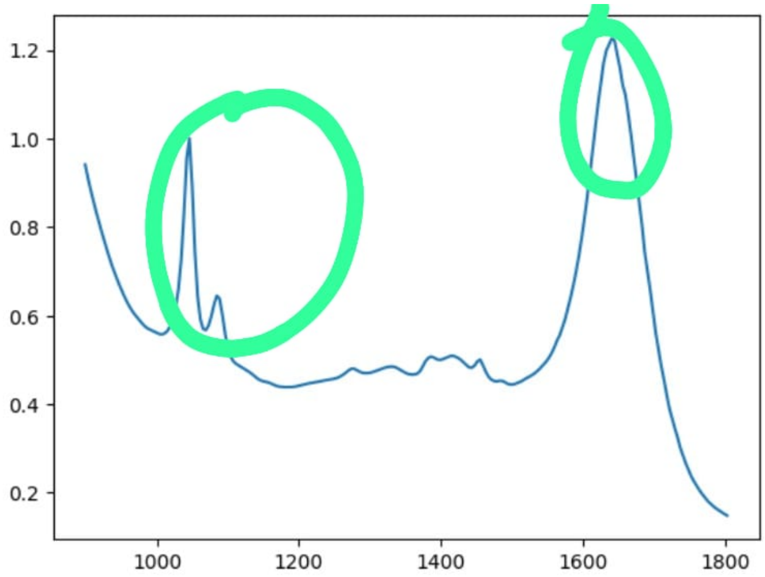


Figure 1: FTIR Spectrum of a wine sample along with three important peaks

3.2 Data Pre-Processing

The vector of each spectrum was re-scaled using **MinMaxScaler** function of **sklearn.preprocessing** package. This re-scales all the data points of the vector into a new value between 0 and 1 corresponding to the position of its original value with respect to the maximum and minimum values of that vector.

However, before scaling, a new dataframe was formed by excluding the data for wavenumbers less than 1000 and greater than 1750 from the original dataframe. This was done to restrict our attention to the three peaks.

Further, after scaling, a threshold was imposed on the dataframe to reduce those values which are less than that threshold, to zero. This was done to further restrict the data to the peaks.

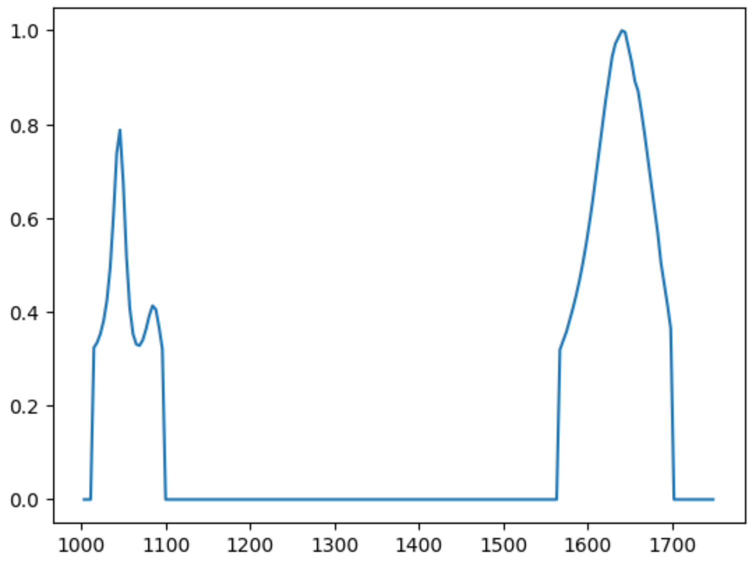


Figure 2: FTIR spectrum of a wine sample for a threshold of about 0.32

3.3 Dimensionality Reduction

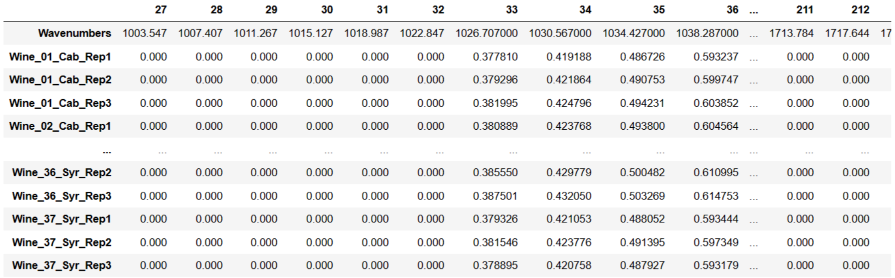
Principal Component Analysis was used to reduce the number of features (i.e., the number of wavenumbers) from about 195 to 20-30. PCA was carried out using the PCA function of sklearn.decomposition package. Above 99% of variance was captured in the new reduced dimensions. The dataset with reduced dimensions was stored in a new dataframe.

4 DATASET OF AVERAGE SPECTRA

Another dataset was made which had the average value of 3 intensities for each wavenumber for each wine. As a result, the new data set had 37 rows with each row containing the average spectrum of a wine. All the analyis, pre-processing and dimensionality reduction was done for this dataset as well. The purpose of this dataset was to compare the results of K-Means Clustering on both the data-sets to find the optimal value of the threshold. At this optimal value, the results of clustering on both the data-sets must match and the classification must be such that one of the classes must contain 19 wines and the other must contain the remaining 18.

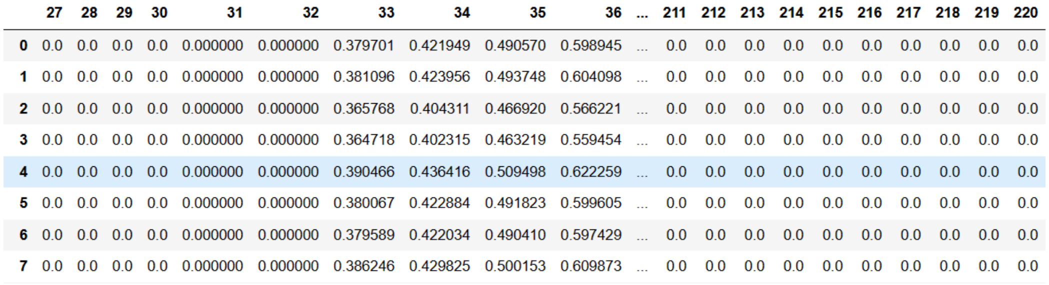
Upon the application of K-Means Clustering, the resulting classes will be denoted as ’0’ and ’1’. Given that the initial dataset included all three spectra of each wine in sequential order, it is imperative that, when classifying a particular wine, all three consecutive observations pertaining to that wine are assigned to the same class. In cases where two out of the three spectra share a common label and the third one differs, the majority’s classification will be adopted for the entire set of three observations corresponding to that particular wine.

For example,
0 0 0 1 1 1 means that first wine belongs to class 0 and second belongs to class 1.
1 1 0 0 1 0 means that first wine belongs to class 1 and second wine belongs to class 0.



	27	28	29	30	31	32	33	34	35	36	...	211	212	
Wavenumbers	1003.547	1007.407	1011.267	1015.127	1018.987	1022.847	1026.707000	1030.567000	1034.427000	1038.287000	...	1713.784	1717.644	17
Wine_01_Cab_Rep1	0.000	0.000	0.000	0.000	0.000	0.000	0.377810	0.419188	0.486726	0.593237	...	0.000	0.000	
Wine_01_Cab_Rep2	0.000	0.000	0.000	0.000	0.000	0.000	0.379296	0.421864	0.490753	0.599747	...	0.000	0.000	
Wine_01_Cab_Rep3	0.000	0.000	0.000	0.000	0.000	0.000	0.381995	0.424796	0.494231	0.603852	...	0.000	0.000	
Wine_02_Cab_Rep1	0.000	0.000	0.000	0.000	0.000	0.000	0.380889	0.423768	0.493800	0.604564	...	0.000	0.000	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Wine_36_Syr_Rep2	0.000	0.000	0.000	0.000	0.000	0.000	0.385550	0.429779	0.500482	0.610995	...	0.000	0.000	
Wine_36_Syr_Rep3	0.000	0.000	0.000	0.000	0.000	0.000	0.387501	0.432050	0.503269	0.614753	...	0.000	0.000	
Wine_37_Syr_Rep1	0.000	0.000	0.000	0.000	0.000	0.000	0.379326	0.421053	0.488052	0.593444	...	0.000	0.000	
Wine_37_Syr_Rep2	0.000	0.000	0.000	0.000	0.000	0.000	0.381546	0.423776	0.491395	0.597349	...	0.000	0.000	
Wine_37_Syr_Rep3	0.000	0.000	0.000	0.000	0.000	0.000	0.378895	0.420758	0.487927	0.593179	...	0.000	0.000	

Figure 3: Figure of Original Dataset



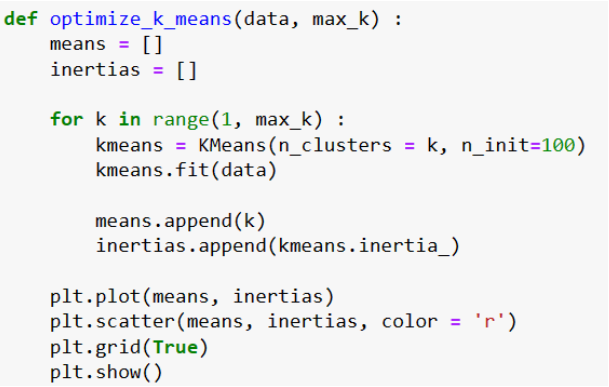
	27	28	29	30	31	32	33	34	35	36	...	211	212	213	214	215	216	217	218	219	220
0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.379701	0.421949	0.490570	0.598945	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.000000	0.000000	0.381096	0.423956	0.493748	0.604098	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.000000	0.000000	0.365768	0.404311	0.466920	0.566221	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.000000	0.000000	0.364718	0.402315	0.463219	0.559454	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.000000	0.000000	0.390466	0.436416	0.509498	0.622259	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.000000	0.000000	0.380067	0.422884	0.491823	0.599605	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.000000	0.000000	0.379589	0.422034	0.490410	0.597429	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.000000	0.000000	0.386246	0.429825	0.500153	0.609873	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 4: Figure of Dataset with Averaged Spectra

5 K-MEANS CLUSTERING

K-Means Clustering is an unsupervised classification ML algorithm. Elbow method was applied on both the datasets to find the optimal number of 'k' (which is the number of clusters). To apply the elbow method, a new function 'optimize_k_means' was defined, which plots the graph of SSE for different k-values. The k-value (known as the optimal k-value) where the elbow occurs in the graph is selected.

The optimal k-value is expected to come out to be 2 for both the datasets. After this, the K-Means Clustering algorithm is applied to both the datasets for k=2 for different values of the threshold.



```
def optimize_k_means(data, max_k) :
    means = []
    inertias = []

    for k in range(1, max_k) :
        kmeans = KMeans(n_clusters = k, n_init=100)
        kmeans.fit(data)

        means.append(k)
        inertias.append(kmeans.inertia_)

    plt.plot(means, inertias)
    plt.scatter(means, inertias, color = 'r')
    plt.grid(True)
    plt.show()
```

Figure 5: 'optimize_k_means' function

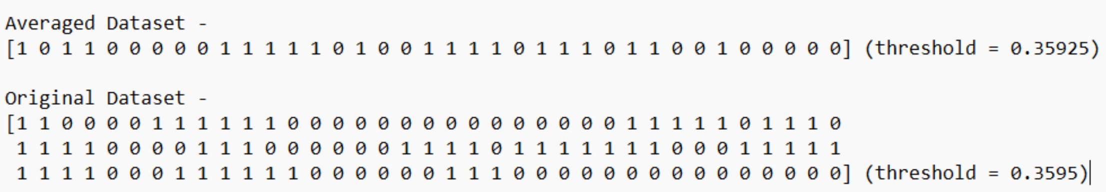
5.1 Impact of Threshold on the Optimal K-Value and Predicted Clusters

The main purpose of this project was to classify the 37 wines into two classes (one of 19 and other of 18). For that, two types of datasets were used (one being the original dataset and the other being the dataset of average spectra). The threshold was applied on both the datasets to control the amount of data from both the datasets being used for clustering. The main aim was to find such values of threshold for both the datasets that the optimal k-value for both the datasets came out to be 2 and the clustered wines from both the datasets matched at the end.

To achieve this, the K-Means Clustering algorithm was tested for a variety of thresholds for both the datasets. It was but obvious that the optimal thresholds for both the datasets would not differ much.

5.2 The Optimal Threshold and the Resulting Clusters

After a series of trials and errors, it was observed that at threshold = 0.3595 for the original dataset and 0.35925 for the dataset of average spectra, the resulting predicted clusters were -



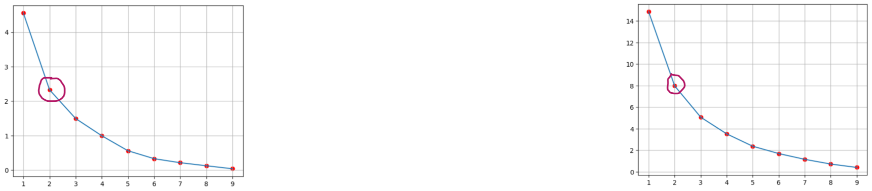
```
Averaged Dataset -
[1 0 1 1 0 0 0 0 0 1 1 1 1 1 0 1 0 0 1 1 1 1 0 1 1 1 0 1 1 0 0 1 0 0 0 0 0] (threshold = 0.35925)

Original Dataset -
[1 1 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 1 1 0
 1 1 1 1 0 0 0 0 1 1 1 0 0 0 0 0 0 1 1 1 1 0 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1
 1 1 1 1 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0] (threshold = 0.3595)
```

Figure 6: Optimal values of threshold

It can be seen from the figure 6 by using the rule described in section 4 that for the optimal values of threshold, the wines classified into 2 clusters almost match for both the datasets. Also, after interchanging the optimal values of thresholds for both the datasets, the classes of wines are interchanged.

It also observed that the optimal value of k for these thresholds is 2.

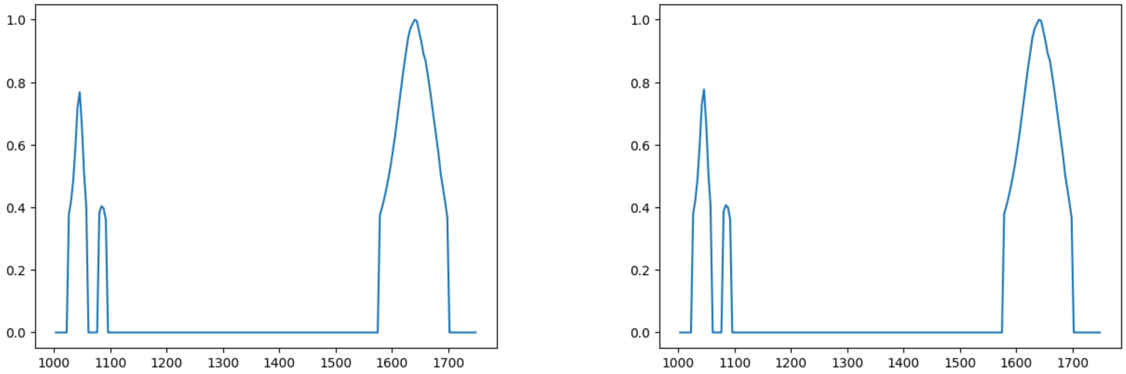
Let us now see the results pictorially.



(a) Elbow for Averaged Dataset

(b) Elbow for Original Dataset

Figure 7: Optimal K-Values



(a) A Spectrum from Original Dataset after applying Optimal Threshold

(b) A Spectrum from Wine Sample from Averaged Dataset after applying Optimal Threshold

Figure 8: Optimal Thresholds applied on Graphs

6 CONCLUSION AND REFERENCES

6.1 Conclusion

The 37 wines were successfully classified into two classes. For this, two datasets were used - the original dataset and the data set with the averaged spectra. This was done to compare the results of K-Means Clustering for both the datasets and to find the optimal threshold for both the datasets at which all the 37 wines are classified into the required two classes with one class having 18 wines and the other having 19 wines. The threshold was applied to restrict the input data to the three significant peaks of the wine samples.

The optimal threshold value come out to be 0.3595 for the original dataset and 0.35925 for the averaged dataset, for which the results of clustering on both the datasets match.

The classes of 37 wines are as follows -
1 0 1 1 0 0 0 0 0 1 1 1 1 1 0 1 0 0 1 1 1 1 0 1 1 1 0 1 1 0 0 1 0 0 0 0 0
where 1 denotes the class with 19 wines
and 0 denotes the class with 18 wines.

6.2 References

[1] Github Repository for Dataset - https://github.com/QIBChemometrics/Wine_Cabernet_Shiraz_FTIR/tree/main