Artificial Intelligence and Machine Learning Fundamentals

Activity 12: k-means Clustering of Sales Data

This section will detect product sales that perform similarly in nature to recognize trends in product sales.
We will be using the Sales Transactions Weekly Dataset from this URL:
https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly
Perform clustering on the dataset using the k-means Algorithm. Make sure you prepare your data for clustering based on what you have learned in the previous lessons. Use the default settings for the k-means algorithm.

1. Load the dataset using pandas.
```
import pandas
pandas.read_csv('Sales_Transactions_Dataset_Weekly.csv')
```
2. If you examine the data in the CSV file, you can realize that the first column contains product id strings. These values just add noise to the clustering process. Also notice that for weeks 0 to 51, there is a W-prefixed label and a Normalized label. Using the normalized label makes more sense, so we can drop the regular weekly labels from the data set.
```
import numpy as np
drop_columns = ['Product_Code']
for w in range(0, 52):
drop_columns.append('W' + str(w))
features = data_frame.drop(dropColumns, 1)
```
3. Our data points are normalized except for the min and max
```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaled_features = scaler.fit_transform(features)
```
4. Create a k-means clustering model and fit the data points into 8 clusters.
```
from sklearn.cluster import KMeans
k_means_model = KMeans()
k_means_model.fit(scaled_features)
```
5. The labels belonging to each data point can be retrieved using the labels_property. These labels determine the clustering of the rows of the original data frame.
```
k_means_model.labels_
```
6. Retrieve the center points and the labels from the clustering algorithm:
```
k_means_model.cluster_centers_
```

The output will be as follows:
array([5, 5, 4, 5, 5, 3, 4, 5, 5, 5, 5, 5, 4, 5, 0, 0, 0, 0, 0, 4, 4,
4,
4, 0, 0, 5, 0, 0, 5, 0, 4, 4, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0,
0, 0, 0, 0, 0, 5, 0, 0, 5, 0, 0, 0, 0, 0, 4, 0, 0, 5, 0, 0, 5,
0,
...
1, 7, 3, 2, 6, 7, 6, 2, 2, 6, 2, 7, 2, 7, 2, 6, 1, 3, 2, 2, 6,

6,
7, 7, 7, 1, 1, 2, 1, 2, 7, 7, 6, 2, 7, 6, 6, 6, 1, 6, 1, 6, 7,
7,
1, 1, 3, 5, 3, 3, 3, 5, 7, 2, 2, 2, 3, 2, 2, 7, 7, 3, 3, 3, 3,
2,
2, 6, 3, 3, 5, 3, 2, 2, 6, 7, 5, 2, 2, 2, 6, 2, 7, 6, 1])

How are these labels beneficial?

Suppose that in the original data frame, the product names are given. You can easily recognize that similar types of products sell similarly. There are also products that fluctuate a lot, and products that are seasonal in nature. For instance, if some products promoted fat loss and getting into shape, they tend to sell during the first half of the year, before the beach season.