Data Mining Machine Learning

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

Dataset

Food	Sweetness	Calories
Apple	7	52
Banana	9	89
Carrot	3	41
Donut	10	400
Eggplant	2	25

Step 1: Calculate the Pairwise Euclidean Distances

The Euclidean distance between two points (x_1, y_1) and (x_2, y_2) is given by the formula:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

where (x_1, y_1) and (x_2, y_2) are the coordinates $(\text{Sweetness}, \text{Calories})$ for each food item.

1. Distance between Apple and Banana

$$d(\text{Apple, Banana}) = \sqrt{(9-7)^2 + (89-52)^2} = \sqrt{2^2 + 37^2} = \sqrt{4 + 1369} = \sqrt{1373} \approx 37.05$$

2. Distance between Apple and Carrot

$$d(\text{Apple, Carrot}) = \sqrt{(3-7)^2 + (41-52)^2} = \sqrt{(-4)^2 + (-11)^2} = \sqrt{16 + 121} = \sqrt{137} \approx 11.70$$

3. Distance between Apple and Donut

$$d(\text{Apple, Donut}) = \sqrt{(10-7)^2 + (400-52)^2} = \sqrt{3^2 + 348^2} = \sqrt{9 + 121104} = \sqrt{121113} \approx 348.02$$

4. Distance between Apple and Eggplant

$$d(\text{Apple, Eggplant}) = \sqrt{(2-7)^2 + (25-52)^2} = \sqrt{(-5)^2 + (-27)^2} = \sqrt{25 + 729} = \sqrt{754} \approx 27.46$$

5. Distance between Banana and Carrot

$$d(\text{Banana, Carrot}) = \sqrt{(3-9)^2 + (41-89)^2} = \sqrt{(-6)^2 + (-48)^2} = \sqrt{36 + 2304} = \sqrt{2340} \approx 48.37$$

6. Distance between Banana and Donut

$$d(\text{Banana, Donut}) = \sqrt{(10-9)^2 + (400-89)^2} = \sqrt{1^2 + 311^2} = \sqrt{1 + 96721} = \sqrt{96722} \approx 311.02$$

7. Distance between Banana and Eggplant

$$d(\text{Banana, Eggplant}) = \sqrt{(2-9)^2 + (25-89)^2} = \sqrt{(-7)^2 + (-64)^2} = \sqrt{49 + 4096} = \sqrt{4145} \approx 64.38$$

8. Distance between Carrot and Donut

$$d(\text{Carrot, Donut}) = \sqrt{(10-3)^2 + (400-41)^2} = \sqrt{7^2 + 359^2} = \sqrt{49 + 128881} = \sqrt{128930} \approx 359.02$$

9. Distance between Carrot and Eggplant

$$d(\text{Carrot, Eggplant}) = \sqrt{(2-3)^2 + (25-41)^2} = \sqrt{(-1)^2 + (-16)^2} = \sqrt{1 + 256} = \sqrt{257} \approx 16.03$$

10. Distance between Donut and Eggplant

$$d(\text{Donut, Eggplant}) = \sqrt{(2-10)^2 + (25-400)^2} = \sqrt{(-8)^2 + (-375)^2} = \sqrt{64 + 140625} = \sqrt{140689} \approx 375.09$$

Step 2: Summary of Pairwise Distances

Pair	Distance
Apple, Banana	37.05
Apple, Carrot	11.70
Apple, Donut	348.02
Apple, Eggplant	27.46
Banana, Carrot	48.37
Banana, Donut	311.02
Banana, Eggplant	64.38
Carrot, Donut	359.02
Carrot, Eggplant	16.03
Donut, Eggplant	375.09

Step 3: Applying DBSCAN with $\varepsilon = 40$ and MinPts $= 2$

Using the parameters $\varepsilon = 40$ and MinPts $= 2$:

- Apple and Carrot form a cluster because $d(\text{Apple, Carrot}) = 11.70 < \varepsilon$.
- Apple and Eggplant also satisfy ε (distance 27.46), so they may be in the same cluster.
- Carrot and Eggplant have a distance of $16.03 < \varepsilon$, so they could also be in the same cluster.
- Banana and Apple are within ε, making Banana potentially part of the same cluster.
- Donut is distant from all other items, making it an outlier.

Final Clustering Summary

Thus, **Apple, Carrot, Eggplant, and possibly Banana** could form a cluster, while **Donut** is likely an outlier.

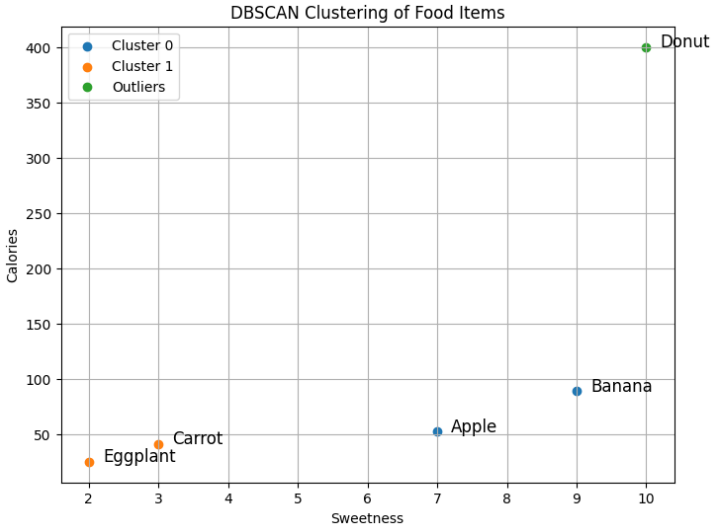

Figure 1: DBScan

1 Questions

1. What is DBSCAN, and how does it differ from other clustering algorithms like K-Means?
2. What are the primary parameters in DBSCAN, and how do they affect the clustering process?
3. How does DBSCAN handle outliers, and what is the significance of points labeled as 'outliers'?
4. How does DBSCAN form clusters based on proximity, and why is the concept of density crucial to this process?
5. What role does the ε (eps) parameter play in DBSCAN, and how does its value influence the number of clusters formed?
6. How does changing the MinPts parameter in DBSCAN affect the clustering outcome?
7. What happens if ε is too small in DBSCAN? What is the effect on outliers and cluster formation?
8. What is the effect of increasing ε on the formation of clusters and the identification of outliers?
9. Why is it important to standardize data before applying DBSCAN, and how does standardization impact the clustering result?
10. What would happen if the MinPts parameter is set to a value greater than the number of data points in a cluster?
11. How does DBSCAN categorize food items (like Apple, Banana, etc.) based on their sweetness and calorie content?
12. What insight does DBSCAN provide regarding the grouping of food items with similar sweetness and calorie values?
13. How might DBSCAN help in identifying food items that are outliers based on their sweetness and calorie values?
14. How do the distribution and range of values for sweetness and calories influence DBSCAN's ability to form meaningful clusters?
15. How can DBSCAN be used to detect hidden patterns in a dataset when clusters do not follow a specific shape (e.g., non-linearly separable clusters)?
16. What are the potential challenges when applying DBSCAN to large datasets, and how can they be addressed?

17. How can you visualize the results of DBSCAN clustering, and what can you infer from the visual representation?
18. What methods can be used to choose the optimal value for ε and MinPts in DBSCAN?
19. What are the practical applications of DBSCAN in clustering food items, and how can it be used for product recommendations or inventory management in the food industry?
20. What are the limitations of DBSCAN, and how can it be improved or combined with other clustering algorithms for better performance on complex datasets?