Data Mining Machine Learning

Md. Abdullah-Al-Kafi
Lecturer
Daffodil International University

Problem Setup

We are given the following data for food classification based on Sweetness and Calories:

Food	Sweetness(x_1)	Calories(x_2)	Label(y)
Apple	7	52	1
Banana	9	89	1
Carrot	3	41	1
Donut	10	400	-1
Eggplant	2	25	1

The goal is to use Support Vector Machine (SVM) to separate the data into two classes: *Healthy* (label 1) and *Unhealthy* (label -1).

SVM Formulation

The objective of SVM is to find the optimal hyperplane that separates the two classes. The decision boundary is given by the equation:

$$w_1 x_1 + w_2 x_2 + b = 0$$

where: - x_1 is the Sweetness, - x_2 is the Calories, - w_1 and w_2 are the weights to be learned, - b is the bias to be learned.

We want to maximize the margin, subject to the constraint that the data points are correctly classified. This can be written as:

$$y_i(w_1 x_{i1} + w_2 x_{i2} + b) \geq 1 \quad \text{for all data points}$$

Support Vectors

Assume that the support vectors are Banana and Donut, which are the closest points to the decision boundary.

For Banana:

$$w_1(9) + w_2(89) + b = 1$$

For Donut:

$$w_1(10) + w_2(400) + b = -1$$

This results in the following system of equations:
1. $9w_1 + 89w_2 + b = 1$ 2. $10w_1 + 400w_2 + b = -1$

Solving the System of Equations

Subtract equation 1 from equation 2 to eliminate b:

$$(10w_1 + 400w_2 + b) - (9w_1 + 89w_2 + b) = -1 - 1$$

$$w_1 + 311w_2 = -2$$

Solving for w_1:

$$w_1 = -2 - 311w_2$$

Substitute this into equation 1:

$$9(-2 - 311w_2) + 89w_2 + b = 1$$

$$-18 - 2800w_2 + 89w_2 + b = 1$$

$$-18 - 2711w_2 + b = 1$$

$$b = 19 + 2711w_2$$

Choosing $w_2 = 0$

For simplicity, assume $w_2 = 0$. Then:

$$w_1 = -2$$
$$b = 19$$

Thus, the equation of the hyperplane is:

$$-2x_1 + 0x_2 + 19 = 0$$

or:

$$x_1 = 9.5$$

Conclusion

The decision boundary is approximately $x_1 = 9.5$, which means that foods with a sweetness value greater than 9.5 (such as Donut) are classified as *Unhealthy*, while foods with a sweetness value less than 9.5 (such as Apple, Banana, etc.) are classified as *Healthy*.

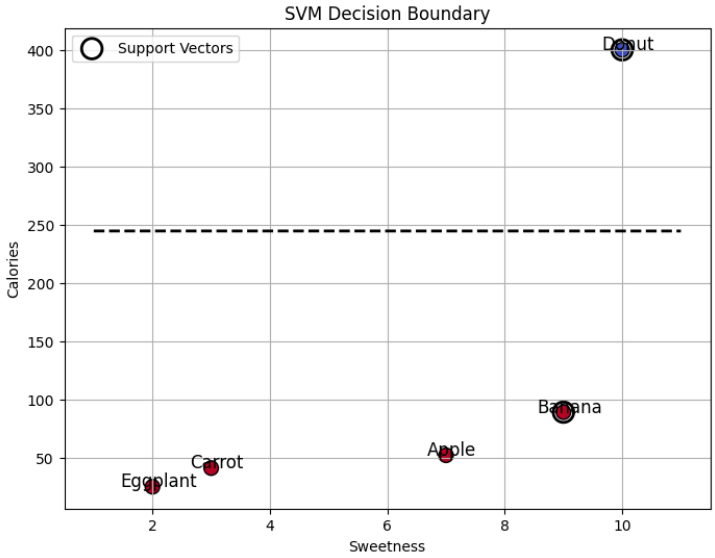

Figure 1: svm

Basic Concepts

1. What is a Support Vector Machine (SVM) and what problem does it solve?
2. How does SVM classify data into different classes?
3. What is a hyperplane in the context of SVM?
4. How does SVM find the optimal hyperplane for classification?
5. What is the "margin" in SVM, and why is it important?
6. What are support vectors, and how do they relate to the optimal hyperplane?

Mathematical Foundations

1. What is the objective function that SVM tries to optimize?
2. How is the margin maximized in the context of SVM?
3. Explain the role of the kernel function in SVM. Why is it used?
4. What is the difference between a soft margin and a hard margin in SVM?
5. What is the concept of slack variables in soft margin SVM?

Types of SVM

1. How does the Linear SVM differ from a Nonlinear SVM?
2. What is the role of the kernel trick in transforming data into a higher-dimensional space?
3. Name and explain different types of kernel functions commonly used in SVM.
4. How does the choice of kernel function affect the performance of the SVM model?

SVM Implementation

1. What are the advantages and disadvantages of using SVM for classification tasks?
2. How do you handle multiclass classification in SVM, given that SVM is inherently binary?
3. How is the performance of an SVM model evaluated, and what metrics are typically used?
4. What is the role of the regularization parameter (C) in SVM, and how does it affect the model's performance?
5. What are the computational complexities of training an SVM with a large dataset, and how can they be mitigated?