# Support Vector Machines (SVM): Lecture Notes

## 1. Intuition Behind SVM

Support Vector Machines (SVM) are supervised learning algorithms used for classification and regression. The core idea is to find the **best decision boundary** (hyperplane) that separates data points from different classes with the **maximum margin**.

### Key Concepts

- **Hyperplane:** In a 2D space, it is a line; in a 3D space, it is a plane; in higher dimensions, it is a hyperplane. The hyperplane separates the feature space into regions for different classes.
- **Margin:** The margin is the perpendicular distance between the hyperplane and the closest data points (called **support vectors**). SVM maximizes this margin.
- **Support Vectors:** These are the data points closest to the hyperplane. Only support vectors influence the position and orientation of the hyperplane.

## 2. Formulating the Optimization Problem

### Hyperplane Representation

The hyperplane is represented as:

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

where:

- $\mathbf{w}$ is the **weight vector**, which determines the orientation of the hyperplane.
- $b$ is the **bias term**, which shifts the hyperplane.
- $\mathbf{x}$ is a point in the feature space.

### Classification Condition

For correct classification:

- For $y_i = +1$: $\mathbf{w} \cdot \mathbf{x}_i + b \geq 1$,
- For $y_i = -1$: $\mathbf{w} \cdot \mathbf{x}_i + b \leq -1$.

These can be combined into a single constraint:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \forall i$$

### Margin Maximization

The margin width is given by:

$$\text{Margin Width} = \frac{2}{\|\mathbf{w}\|}$$

Maximizing the margin is equivalent to minimizing $\frac{1}{2}\|\mathbf{w}\|^2$. The optimization problem becomes:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

subject to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad \forall i$$

# 3. Solving Using Lagrange Multipliers

To handle the constraints, we use **Lagrange multipliers** $\alpha_i \geq 0$. The Lagrangian is:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1\right]$$

## Stationarity Conditions

1. **With respect to w:**

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$

Solve for $\mathbf{w}$:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

2. **With respect to $b$:**

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

This implies:

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

## Dual Problem

Substitute $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ into the Lagrangian to eliminate $\mathbf{w}$. The dual problem becomes:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

subject to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0 \quad \forall i$$

# 4. Computing w and $b$

## Weight Vector w

Once $\alpha_i$ is obtained from the dual problem, compute:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

## Bias $b$

To compute $b$, choose any support vector $\mathbf{x}_k$ (where $\alpha_k > 0$):

$$b = y_k - \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}_k)$$

# 5. Classifying New Data Points

For a new data point $\mathbf{x}$, the decision function is:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

Classify based on the sign of $f(\mathbf{x})$:

- If $f(\mathbf{x}) > 0$, predict $y = +1$,
- If $f(\mathbf{x}) < 0$, predict $y = -1$.

# 6. What is $\mathbf{w}$?

The vector $\mathbf{w}$ in SVM is a key parameter of the model that determines the orientation of the hyperplane. It has the following characteristics:

## Geometric Interpretation

- $\mathbf{w}$ is the **normal vector** to the hyperplane, meaning it is perpendicular to the decision boundary.
- The direction of $\mathbf{w}$ determines which side of the hyperplane corresponds to each class:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$$

  - If $f(\mathbf{x}) > 0$, $\mathbf{x}$ is classified as $+1$. - If $f(\mathbf{x}) < 0$, $\mathbf{x}$ is classified as $-1$.

- The magnitude $\|\mathbf{w}\|$ is inversely related to the margin width. A smaller $\|\mathbf{w}\|$ implies a larger margin:

$$\text{Margin Width} = \frac{2}{\|\mathbf{w}\|}$$

## Mathematical Derivation

- $\mathbf{w}$ is computed as a weighted sum of the support vectors:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

  where:

  - $\alpha_i > 0$: Lagrange multipliers for support vectors.
  - $y_i$: Class label ($+1$ or $-1$).
  - $\mathbf{x}_i$: Feature vector of a support vector.

- Non-support vectors ($\alpha_i = 0$) do not contribute to $\mathbf{w}$.

## Importance of $\mathbf{w}$

- $\mathbf{w}$ encodes the direction of the hyperplane and is essential for classification.
- It determines the influence of each feature on the classification decision.
- The SVM training process focuses on finding $\mathbf{w}$ that maximizes the margin while satisfying the constraints.

# 7. Summary

- SVM aims to find the hyperplane that maximizes the margin.
- The optimization problem is solved using Lagrange multipliers, leading to the dual problem.
- The final classifier is based on the weight vector $\mathbf{w}$ and bias $b$, with contributions primarily from the support vectors.