Concepts in Support Vector Machines

Support Vector Machines (SVMs) are powerful machine learning models known for their ability to find optimal decision boundaries by maximizing the margin between different classes.

Like logistic regression, SVMs are a binary classification algorithm that uses a hyperplane to classify the points.

This pre-lecture quiz[1] will introduce you to the following ideas that will be used in the lecture:

- **Functional Margin:** The Functional margin of a point reflects the correctness of a data point's classification relative to the hyperplane. The functional margin of the set reflects if all the points are correctly classified.
- **Canonical Weights:** The term "canonical" refers to the standard form. Canonical weights represent a hyperplane in a normalized or standard form within the SVM framework.
- **Geometric Margin:** Similar to the functional margin, the geometric margin accounts for classification correctness. Its absolute value also quantifies the Euclidean distance from a point to the hyperplane. The geometric margin of a set evaluates classification correctness and how close the closest point is to the hyperplane.
- **Gradient Calculation:** The training of an SVM involves adjusting the hyperplane using gradients. Here you will perform the gradient calculation needed in the lecture.

The ideas above are the concepts we will use to describe our objective function and the optimizer.

To discuss these ideas, we will need many background concept such as the distinction between the positive and negative sides of a hyperplane, and the euclidean distance of a point to the hyperplane

[1] Assume all points (training examples) are not on the hyperplane. If needed, look up the definition of the vector normal to a hyperplane.

1 Functional Margin

In this section, we discuss the concept of functional margin, which is used in the SVM's objective function. We distinguish between:

- the functional margin of a *training example* with respect to the hyperplane,
- the functional margin of a *set of training examples* with respect to the hyperplane.

Before we define the functional margin for a *set of training examples*, we establish:

- the notion of the positive/negative side of a hyperplane,
- a new encoding for class labels,
- the functional margin of an individual training example (aka labeled point).

Positive/Negative Side of a Hyperplane:

Consider a hyperplane characterized by the equation $\mathbf{w}^T\mathbf{x}+w_0 = 0$, where $\mathbf{w}$ is the normal vector and w_0 is the bias/intercept term.

- A point $\mathbf{x}$ is on the **positive side** of the hyperplane if $\mathbf{w}^T\mathbf{x} + w_0 > 0$.
- Conversely, a point $\mathbf{x}$ is on the **negative side** if $\mathbf{w}^T\mathbf{x} + w_0 < 0$.

Multiple Choice Question

1. Which equation represents a point $\mathbf{x}$ is on the negative side of the hyperplane?

   a) $\mathbf{w}^T\mathbf{x} + w_0 = 0$

   b) $\mathbf{w}^T\mathbf{x} + w_0 > 0$

   c) $\mathbf{w}^T\mathbf{x} + w_0 < 0$

   d) $v = 0$

Encoding Class Labels and Intuitive Classification

When we're classifying with SVMs, we mark one class as +1 and the other as -1. This encoding simplifies the writing of the objective functions for SVMs.

Functional Margin of a Point

Think of functional margin as a way to observe if a training example is correctly classified. If a training example has a positive functional margin, it's like saying, "You're correctly classified" But if the functional margin is negative, it's like telling the training example , "Oops, you're incorrectly classified."

For a given training example $(\mathbf{x}^{(\mathbf{i})}, y^{(i)})$ where $\mathbf{x}^{(i)}$ is the input and $y^{(i)}$ is its label, the functional margin $\gamma_f^{(i)}$ of this training example with respect to the hyperplane is given by:

$$\gamma_f^{(i)} = y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + w_0)$$

Given the functional margin formula, consider the following scenarios:

* For a training example with label +1 (i.e., $y^{(i)} = +1$):

  - If $\mathbf{x}^{(i)}$ is on the **positive side** of the hyperplane, $\mathbf{w}^T\mathbf{x}^{(i)} + w_0$ is positive. Thus, the functional margin $\gamma_f^{(i)}$ is also positive, indicating *correct classification.*
  - If $\mathbf{x}^{(i)}$ is on the **negative side** of the hyperplane, $\gamma_f^{(i)}$ will be negative, indicating *incorrect classification.*

* For a training example with label -1 (i.e., $y^{(i)} = -1$):

  - If x_i is on the **negative side** of the hyperplane, multiplying by -1 turns the value positive. Hence, $\gamma_f^{(i)}$ is positive, indicating *correct classification.*
  - Conversely, if x_i is on the positive side, $\gamma_f^{(i)}$ will be negative, indicating *incorrect classification.*

Multiple Choice Questions

2. A positive functional margin of a training example indicates:

   a) Incorrect classification

   b) Correct classification

   c) The point is on the hyperplane

   d) The point is far from the hyperplane

3. If the functional margin for a training example with label +1 is negative, this means:

a) The example is correctly classified

b) The example is on the hyperplane

c) The example is incorrectly classified

d) None of these

4. When a training example labeled as -1 lies on the negative side of the hyperplane, its functional margin is:

a) Negative

b) Zero

c) Positive

d) Undefined

5. For a training example with label -1 and a positive functional margin, where does it lie with respect to the hyperplane?

a) On the positive side

b) On the negative side

c) On the hyperplane

d) Far from the hyperplane

Functional Margin of the Set of Training Examples

The functional margin of a set of training examples tells us if a hyperplane correctly classifies all the training examples.

Given training examples $\{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$, the functional margin of the set of training examples with respect to the hyperplane $\mathbf{w}^T\mathbf{x} + w_0 = 0$ is the minimum of the individual functional margins:

$$\gamma_f = \min_{i=1}^{N} y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) = \min_{i=1}^{N} \gamma_f^{(i)}$$

Multiple Choice Questions

6. If the functional margin of a set of training examples is positive, it indicates:

   a) The hyperplane poorly classifies the training examples.

   b) At least one example is close to the hyperplane.

   c) The hyperplane correctly classifies all the training examples.

   d) The hyperplane passes through the origin.

7. The functional margin of a set of training examples is:

   a) The average functional margin of all training examples.

   b) The smallest functional margin across all training examples.

   c) The product of functional margins of all training examples.

   d) The largest functional margin across all training examples.

2 Canonical Weights for Functional Margin

The term ”canonical” in the context of a hyperplane’s equation refers to a standard or normalized form.

Definition: For an SVM hyperplane, the canonical form is achieved by adjusting the weights and bias such that the *functional margin for the set* is 1. This standardization simplifies the optimization problem.

Scaling the weights and bias of a hyperplane by any positive scalar c retains the decision boundary while altering the functional margin. A hyperplane is in canonical form when the smallest functional margin of the dataset is 1, defined as:

$$1 = \gamma_f = \min_{i=1}^{N} y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b)$$

This implies that for the closest training example $(\mathbf{x}^{(i)}, y^{(i)})$ to the hyperplane, the functional margin under canonical weights $\mathbf{w}$ and w_0 satisfies:

$$y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + w_0) = 1$$

Example: Assume a non-canonical hyperplane with weights $\mathbf{w}^T = [2, 2]$ and bias $w_0 = -2$ in a 2D space. Consider training examples $\{([1, -4]^T, -1), ([1, 1]^T, 1)\}$. The functional margin for this set is:

$$1 \cdot ((2 \cdot 1) + (2 \cdot 1) - 2) = 2$$

For correctly classified training examples, to convert this hyperplane into canonical form, we divide the weights and bias by the current functional margin for the set, which is 2. Thus, the canonical weights $\mathbf{w_c}$ and w_{0c} become:

$$\mathbf{w_c} = \frac{1}{2} \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad w_{0c} = \left(\frac{1}{2}\right)(-2) = -1$$

Now, the functional margin for the set with the canonical hyperplane is:

$$1 \cdot ((1 \cdot 1) + (1 \cdot 1) - 1) = 1$$

This confirms that the hyperplane equation $x_1 + x_2 - 1 = 0$ is now in canonical form.

Multiple Choice Questions

8. In the canonical form of a hyperplane, the functional margin with respect to the closest training example is:

a) 0

b) -1

c) 1

d) Undefined

9. For correctly classified training examples, adjusting the weights and bias to their canonical form:

a) Changes the position of the hyperplane.

b) Makes the functional margin negative.

c) Ensures the smallest functional margin is 1.

d) Reduces the geometric margin.

3 Signed Distance

Before defining the signed distance of a point to a hyperplane, we first define the hyperplane, the normal vector to the plane, and the *(unsigned) distance* of a training example to a plane.

Distance Formula

Once you have the coefficient vector $(w_1, w_2, \ldots, w_d)$ and a point $(x_1, x_2, \ldots, x_d)$ not on the plane, you can calculate the distance between the training example and the plane using the formula:

$$\frac{|w_0 + w_1x_1 + w_2x_2 + \cdots + w_dx_d|}{\sqrt{w_1^2 + w_2^2 + \cdots + w_d^2}}$$

When we're training an SVM, we're really trying to get this hyperplane to sit in just the right spot so that this distance is as big as possible to the closest training example.

Example:

Let's say we have a plane with equation $2x_1 - 3x_2 + x_3 - 7 = 0$, and we want to find the distance from training example $\mathbf{x}^T = (4, -2, 5)$ to this plane.

- Coefficients of the normal vector: $\mathbf{w}^T = (w_1, w_2, w_3) = (2, -3, 1)$
- Using the formula:

$$\frac{|2 \cdot 4 - 3 \cdot (-2) + 1 \cdot 5 - 7|}{\sqrt{2^2 + (-3)^2 + 1^2}} = \frac{8}{\sqrt{14}}$$

Multiple Choice Questions:

10 If a plane has the equation $2x_1 - 2x_2 + 4x_3 + x_4 - 6 = 0$, and a training example $\mathbf{x}^T = (-4, -2, -4, 1)$ is given, what is the distance from the training example to the plane?

a) 5

b) 3

c) 7

d) 2

Signed Distance Formula

The hyperplane in SVMs is not just a boundary between classes but also provides information on the relative position of training examples. The *signed distance* measures not only how far a point is from the hyperplane but also indicates which side of the hyperplane the training example is on.

The signed distance r from a point (x_1, x_2, x_3) to a hyperplane given by the formula:

$$r = \frac{w_0 + w_1x_1 + w_2x_2 + w_3x_3}{\sqrt{w_1^2 + w_2^2 + w_3^2}}$$

Here, the sign of r indicates the training example's position relative to the hyperplane. A positive r implies that the training example is on the same side as the normal vector's direction, whereas a negative r indicates it is on the opposite side.

Multiple Choice Questions

11. If a plane has the equation $2x_1-2x_2+4x_3+x_4-6 = 0$, and a point $\mathbf{x}^T = (-4, -2, -4, 1)$ is given, what is the signed distance from the training example to the plane?

a) 5

b) 3

c) 7

d) 2

e) -5

f) -3

g) -7

h) -2

12. If the signed distance from a data point to the decision boundary (hyperplane) is negative, what does this indicate about the training example?

a) The training example lies on the boundary.

b) The training example lies on the same side of the boundary as the direction of the normal vector.

c) The training example is classified with high confidence.

d) The training example lies on the opposite side of the boundary from the direction of the normal vector.

4 Geometric Margin

Like the functional margin, the geometric margin for *individual points* and a *set* are defined.

The geometric margin of a *point* (e.g. training examples) is a measure of distance to the decision boundary and classification correctness.

For a hyperplane defined by $\mathbf{w}^T\mathbf{x} + w_0 = 0$, the geometric margin $\gamma_g^{(i)}$ of a data point $(\mathbf{x}^{(i)}, y^{(i)})$ is:

$$\gamma_g^{(i)} = y^{(i)} \left(\frac{\mathbf{w}^T\mathbf{x}^{(i)} + w_0}{\|\mathbf{w}\|} \right)$$

Here, $\|\mathbf{w}\|$ is the Euclidean norm of the weight vector $\mathbf{w}$. The geometric margin for a *set* (e.g. training set) is the smallest of these individual margins:

$$\gamma_g = \min_i \gamma_g^{(i)}$$

Example: Consider a 2D dataset with a hyperplane defined by the weights $\mathbf{w}^T = [1, 2]$ and bias $w_0 = -3$. Let's compute the geometric margin for a correctly classified point $(\mathbf{x}, y) = ([3, 3]^T, 1)$.

First, calculate the functional margin without normalization:

$$\gamma_f = (1 \cdot ((1 \cdot 3) + (2 \cdot 3) - 3)) = 9 - 3 = 6$$

To find the geometric margin, we normalize by the magnitude of $\mathbf{w}$:

$$\|\mathbf{w}\| = \sqrt{1^2 + 2^2} = \sqrt{5}$$

$$\gamma_g = \frac{\gamma_f}{\|\mathbf{w}\|} = \frac{6}{\sqrt{5}}$$

Now, if $\mathbf{w}$ is a unit vector, say $\mathbf{w}^T = \left[\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right]$, the geometric margin is directly the functional margin since the weight vector is already normalized:

$$\gamma_g = (1 \cdot \left(\frac{1}{\sqrt{5}} \cdot 3 + \frac{2}{\sqrt{5}} \cdot 3 - \frac{3}{\sqrt{5}} \right)) = \frac{6}{\sqrt{5}}$$

In both cases, with and without $\mathbf{w}$ being a unit vector, the geometric margin γ_g remains the same.

Multiple Choice Questions

13 Consider a hyperplane in 3D space defined by the equation $\mathbf{w}^T\mathbf{x} + w_0 = 0$ with $\mathbf{w}^T = [1, -2, 2]$ and $w_0 = 3$. For training example $(\mathbf{x}, y) = (([x_1, x_2, x_3]^T, y) = ([4, 1, -2]^T, -1)$, what is the geometric margin $\gamma_g^{(i)}$ for this example?

a) $-1 \times \frac{-4}{3}$

b) $-1 \times \frac{2}{3}$

c) $-1 \times \frac{1}{3}$

d) $-1 \times \frac{11}{3}$

14 Given the following geometric margins for individual points in the dataset (aka training examples):

- $\gamma_g^{(1)} = 2$
- $\gamma_g^{(2)} = 11$
- $\gamma_g^{(3)} = 0.5$
- $\gamma_g^{(4)} = 0.8$

What is the geometric margin γ_g for the dataset?

a) $\gamma_g = 2$

b) $\gamma_g = 11$

c) $\gamma_g = 0.5$

d) $\gamma_g = 0.8$

e) None of these

5 Gradient Computations in SVM

The following is the gradient computation that will be needed in the lecture.

Multiple Choice Question

15. Given the expression $f(y^{(i)}, \mathbf{w}, \mathbf{x}^{(i)}, w_0) = 1 - y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + w_0)$ compute the gradient of f with respect to the coefficient/weight vector $\mathbf{w}$.

a) $-y^{(i)}\mathbf{x}^{(i)}$

b) $y^{(i)}\mathbf{x}^{(i)}$

c) $-y^{(i)}$

d) $y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + w_0)$

e) $1 - y^{(i)}$