Week-10: Support Vector Machine

Sherry Thomas
21f3001449

Contents

Abstract

This week's curriculum entails a further examination of the perceptron algorithm, followed by a comprehensive exploration of support vector machines (SVM) and subsequently, an elaboration on the concept of soft-margin SVM.

Perceptron and Margin Maximization in Linear Separability

In this week, we explore the concept of perceptron learning in the context of linearly separable datasets with a specified margin, denoted as γ. The dataset is defined as $D = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$, where $\mathbf{x_i} \in \mathbb{R}^d$ represents data points and $y_i \in \{-1, 1\}$ indicates their respective classes.

Defining Key Parameters

We begin by introducing essential parameters:

1. **Weight Vector**: Let $\mathbf{w}^* \in \mathbb{R}^d$ be the weight vector such that $(\mathbf{w^{*T}x_i})y_i \geq \gamma$ for all i. This weight vector ensures the desired margin γ.
2. **Bounding Radius**: We define $R > 0 \in \mathbb{R}$ such that for all i, $||\mathbf{x_i}|| \leq R$. This represents a constraint on the norm of input data.

With these parameters established, we can formulate the upper bound on the number of mistakes made by the perceptron learning algorithm:

$$\#\text{mistakes} \leq \frac{R^2}{\gamma^2}$$

Observations and Objectives

Observations

We make several observations regarding the perceptron learning algorithm:

1. The "quality" of the solution depends on the margin γ.
2. The number of mistakes is influenced by the margin associated with $\mathbf{w}^*$.
3. The weight vector $\mathbf{w_{perc}}$ need not be identical to $\mathbf{w}^*$.

Goal: Margin Maximization

Given these observations, our overarching goal is to find the solution that maximizes the margin. It is crucial to note that a single dataset can have multiple linear classifiers with varying margins, as depicted in the diagram below:

To formalize our objective, we aim to maximize the margin γ:

$$\max_{\mathbf{w},\gamma} \gamma$$

Subject to the following constraints:

$$\begin{aligned} (\mathbf{w^T x_i})y_i &\geq \gamma \quad \text{for all } i \\ ||\mathbf{w}||^2 &= 1 \end{aligned}$$

Reformulating the Objective

To simplify our objective, we can express γ in terms of the width of $\mathbf{w}$:

$$\max_{\mathbf{w}} \text{width}(\mathbf{w})$$

Subject to the constraint:

$$(\mathbf{w^T x_i})y_i \geq 1 \quad \text{for all } i$$

Width Calculation

The width of the margin, denoted as width($\mathbf{w}$), can be calculated as the distance between two parallel margins. Consider two points $\mathbf{x}$ and $\mathbf{z}$ lying on opposite sides of the decision boundary such that $\mathbf{w^T x} = -1$ and $\mathbf{w^T z} = 1$ or vice versa.

Let $\mathbf{x_1}$ and $\mathbf{x_2}$ be two points on the margins and opposite sides of the decision boundary.

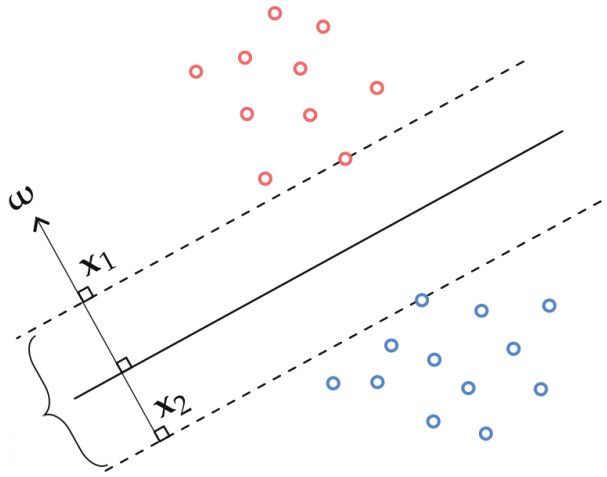

Figure 1: Margin Width

The width is given by:

$$\begin{aligned} \mathbf{x_1^T w} - \mathbf{x_2^T w} &= 2 \\ (\mathbf{x_1} - \mathbf{x_2})^T \mathbf{w} &= 2 \\ ||\mathbf{x_1} - \mathbf{x_2}||_2 ||\mathbf{w}||_2 \cos(\theta) &= 2 \\ \therefore ||\mathbf{x_1} - \mathbf{x_2}||_2 &= \frac{2}{||\mathbf{w}||_2} \end{aligned}$$

Hence, our objective can be restated as:

$$\max_{\mathbf{w}} \frac{2}{||\mathbf{w}||_2^2} \quad \text{subject to} \quad (\mathbf{w^T x_i}) y_i \geq 1 \quad \text{for all } i$$

Equivalently, we can frame it as a minimization problem:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 \quad \text{subject to} \quad (\mathbf{w^T x_i}) y_i \geq 1 \quad \text{for all } i$$

In summary, finding the separating hyperplane with the maximum margin is equivalent to finding the one with the smallest possible normal vector $\mathbf{w}$.

Constrained Optimization and Dual Problem in Lagrange Multipliers

In the realm of constrained optimization, we encounter problems formulated as follows:

$$\min_{\mathbf{w}} f(\mathbf{w})$$
$$s.t.\ g(\mathbf{w}) \leq 0$$

To tackle such problems effectively, the method of **Lagrange Multipliers** comes into play. Lagrange multipliers are instrumental in solving constrained optimization problems by identifying optimal values of an objective function while adhering to a set of constraints. In this method, the constraints are seamlessly integrated into the objective function through the introduction of auxiliary variables known as Lagrange multipliers.

Lagrange Function

For the optimization problem described above, the Lagrange function, denoted as $\mathcal{L}(\mathbf{w}, \alpha)$, is defined as follows:

$$\mathcal{L}(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha^T g(\mathbf{w}) \quad \text{for all } \mathbf{w}$$

Here, α is a vector of Lagrange multipliers, constrained to be non-negative ($\alpha \geq \mathbf{0}$).

Maximizing the Lagrange function with respect to α leads us to the following formulation:

$$\begin{aligned}\max_{\alpha \geq \mathbf{0}} \mathcal{L}(\mathbf{w}, \alpha) &= \max_{\alpha \geq \mathbf{0}} \left(f(\mathbf{w}) + \alpha^T g(\mathbf{w})\right) \\ &= \begin{cases} \infty & \text{if } g(\mathbf{w}) > \mathbf{0} \\ f(\mathbf{w}) & \text{if } g(\mathbf{w}) \leq \mathbf{0} \end{cases}\end{aligned}$$

Since the Lagrange function is equivalent to $f(\mathbf{w})$ when $g(\mathbf{w}) \leq \mathbf{0}$, we can rewrite our original problem as follows:

$$\begin{aligned}\min_{\mathbf{w}} f(\mathbf{w}) &= \min_{\mathbf{w}} \left[\max_{\alpha \geq \mathbf{0}} \mathcal{L}(\mathbf{w}, \alpha)\right] \\ &= \min_{\mathbf{w}} \left[\max_{\alpha \geq \mathbf{0}} \left(f(\mathbf{w}) + \alpha^T g(\mathbf{w})\right)\right]\end{aligned}$$

In general, interchanging the positions of the min and max functions is not permissible unless all involved functions are convex. In our case, both f and g are convex functions, allowing us to express this as:

$$\min_{\mathbf{w}} \left[\max_{\alpha \geq \mathbf{0}} \left(f(\mathbf{w}) + \alpha^T g(\mathbf{w})\right)\right] \equiv \max_{\alpha \geq \mathbf{0}} \left[\min_{\mathbf{w}} \left(f(\mathbf{w}) + \alpha^T g(\mathbf{w})\right)\right]$$

Multiple Constraints

Now, extending our discussion to scenarios with m constraints, denoted as $g_i(\mathbf{w}) \leq \mathbf{0}$ for $i \in [1, m]$, we can represent the problem as follows:

$$\begin{aligned}\min_{\mathbf{w}} f(\mathbf{w}) &\equiv \min_{\mathbf{w}} \left[\max_{\alpha \geq \mathbf{0}} f(\mathbf{w}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{w})\right] \\ &\equiv \max_{\alpha \geq \mathbf{0}} \left[\min_{\mathbf{w}} f(\mathbf{w}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{w})\right]\end{aligned}$$

Formulating the Dual Problem

To formalize the dual problem, we start with our objective function:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 \quad \text{s.t. } (\mathbf{w}^T \mathbf{x}_i) y_i \geq 1 \quad \forall i$$

The constraints can be expressed as:

$$\begin{aligned}(\mathbf{w}^T \mathbf{x}_i) y_i &\geq 1 \quad \forall i \\ 1 - (\mathbf{w}^T \mathbf{x}_i) y_i &\leq 0 \quad \forall i\end{aligned}$$

Introducing Lagrange multipliers $\alpha \in \mathbb{R}^d$, we construct the Lagrange function:

$$\mathcal{L}(\mathbf{w}, \alpha) = \frac{1}{2}||\mathbf{w}||_2^2 + \sum_{i=1}^{n} \alpha_i (1 - (\mathbf{w}^T \mathbf{x}_i) y_i)$$

Now, considering the duality principle, we arrive at:

$$\min_{\mathbf{w}} \max_{\alpha \geq \mathbf{0}} \left[\frac{1}{2}||\mathbf{w}||_2^2 + \sum_{i=1}^{n} \alpha_i (1 - (\mathbf{w}^T \mathbf{x}_i) y_i)\right] \equiv \max_{\alpha \geq \mathbf{0}} \min_{\mathbf{w}} \left[\frac{1}{2}||\mathbf{w}||_2^2 + \sum_{i=1}^{n} \alpha_i (1 - (\mathbf{w}^T \mathbf{x}_i) y_i)\right]$$

Solving for the inner function of the dual problem, we obtain:

$$\mathbf{w}_\alpha^* - \sum_{i=1}^{n} \alpha_i \mathbf{x}_i y_i = 0$$

This leads to a vectorized form:

$$\mathbf{w}_\alpha^* = \mathbf{X}\mathbf{Y}\alpha \quad \dots [1]$$

Here, $\mathbf{X} \in \mathbb{R}^{d \times n}$ represents the dataset, $\mathbf{Y} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with label values on its diagonals, and $\alpha \in \mathbb{R}^n$.

Rewriting the outer dual function, we get,

$$\begin{aligned}
&\max_{\alpha \geq \mathbf{0}} \left[\frac{1}{2} ||\mathbf{w}||_2^2 + \sum_{i=1}^{n} \alpha_i (1 - (\mathbf{w}^T \mathbf{x}_i) y_i) \right] \\
&= \max_{\alpha \geq \mathbf{0}} \left[\frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{i=1}^{n} \alpha_i - \mathbf{w}^T \mathbf{X}\mathbf{Y}\alpha \right] \\
&= \max_{\alpha \geq \mathbf{0}} \left[\frac{1}{2} (\mathbf{X}\mathbf{Y}\alpha)^T (\mathbf{X}\mathbf{Y}\alpha) + \sum_{i=1}^{n} \alpha_i - (\mathbf{X}\mathbf{Y}\alpha)^T (\mathbf{X}\mathbf{Y}\alpha) \right] \quad \dots \text{from } [1] \\
&= \max_{\alpha \geq \mathbf{0}} \left[\sum_{i=1}^{n} \alpha_i - \frac{1}{2} (\mathbf{X}\mathbf{Y}\alpha)^T (\mathbf{X}\mathbf{Y}\alpha) \right]
\end{aligned}$$

Observations

1. **Dual and Primal Variables:** The duality of the SVM problem manifests in the dimensions of its variables. The dual problem is characterized by $\alpha \geq \mathbf{0}$, residing in $\mathbb{R}^n$, while the primal problem deals with $\mathbf{w}$ in $\mathbb{R}^d$.
2. **Simplicity of Dual Problem:** Solving the dual problem is often considered "easier" compared to the primal problem.
3. **Kernelization Potential:** The dual problem's dependence on $\mathbf{X}^T\mathbf{X}$ opens the door for kernelization techniques.

Let's now examine the equation:

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

Observations on the Equation

1. **Interpreting $\mathbf{w}^*$:** The optimal $\mathbf{w}^*$ emerges as a linear combination of data points, where each data point's significance is determined by the corresponding Lagrange multiplier α_i.
2. **Varying Importance:** Consequently, some data points exert greater influence on $\mathbf{w}^*$ than others.

Support Vector Machine (SVM)

Duality Revisited

Revisiting the Lagrangian function:

$$\min_{\mathbf{w}} \left[\max_{\alpha \geq \mathbf{0}} f(\mathbf{w}) + \alpha^T g(\mathbf{w}) \right] \equiv \max_{\alpha \geq \mathbf{0}} \left[\min_{\mathbf{w}} f(\mathbf{w}) + \alpha^T g(\mathbf{w}) \right]$$

Here, the primal function is on the left-hand side, and the dual function is on the right. The solutions for the primal and dual functions are denoted as $\mathbf{w}^*$ and α^*, respectively. When these solutions are substituted into the equation, we obtain:

$$\max_{\alpha \geq \mathbf{0}} f(\mathbf{w}^*) + \alpha^T g(\mathbf{w}^*) = \min_{\mathbf{w}} f(\mathbf{w}) + \alpha^{*T} g(\mathbf{w})$$

Given that $g(\mathbf{w}^*) \leq \mathbf{0}$, the left-hand side simplifies to $f(\mathbf{w}^*)$:

$$f(\mathbf{w}^*) = \min_{\mathbf{w}} f(\mathbf{w}) + \alpha^{*T} g(\mathbf{w})$$

Substituting $\mathbf{w}^*$ for $\mathbf{w}$ in the right-hand side yields a new right-hand side that is greater than or equal to the current one:

$$f(\mathbf{w}^*) \leq f(\mathbf{w}^*) + \alpha^{*T} g(\mathbf{w}^*)$$

Hence, we can infer:

$$\alpha^{*T} g(\mathbf{w}^*) \geq \mathbf{0}$$

However, considering the constraints, where $\alpha^* \geq \mathbf{0}$ and $g(\mathbf{w}^*) \leq \mathbf{0}$, we arrive at:

$$\alpha^{*T} g(\mathbf{w}^*) \leq \mathbf{0}$$

From this, we deduce:

$$\alpha^{*T} g(\mathbf{w}^*) = \mathbf{0}$$

Extending this equation for multiple constraints, we obtain:

$$\alpha_i^* g(w_i^*) = 0 \quad \forall i$$

Hence, if one of the two values is greater than zero, the other must be zero. Considering $g(\mathbf{w}^*) = 1 - (\mathbf{w}^T \mathbf{x}_i) y_i$, we can express this as:

$$\alpha_i^* (1 - (\mathbf{w}^T \mathbf{x}_i) y_i) = 0 \quad \forall i$$

Importantly, when $\alpha_i > \mathbf{0}$, we deduce:

$$(\mathbf{w}^T \mathbf{x}_i) y_i = 1$$

This implies that the i^{th} data point resides on the **"Supporting"** hyperplane and significantly contributes to the determination of $\mathbf{w}^*$.

Consequently, data points with $\alpha_i > \mathbf{0}$ earn the title of **Support Vectors**, and this algorithm is known as the **Support Vector Machine (SVM)**.

Definition of Support Vector Machines (SVMs)

Support Vector Machines (SVMs) stand as a category of supervised learning algorithms designed for classification and regression analysis. SVMs aim to identify the optimal hyperplane that maximizes the margin between data points from different classes. In scenarios where data is not linearly separable, SVMs employ kernel functions to map the data into a higher-dimensional space where a linear decision boundary can effectively segregate the classes.

Insight: $\mathbf{w}^*$ represents a sparse linear combination of data points.

Hard-Margin SVM Algorithm

The Hard-Margin SVM algorithm is applicable only when the dataset is linearly separable with a margin parameter $\gamma > 0$. Its key steps include:

1. **Direct or Kernelized Calculation of Q:** Compute the matrix $Q = \mathbf{X}^T\mathbf{X}$ directly or using a kernel, based on the dataset.
2. **Gradient Descent:** Employ the gradient of the dual formula, $\alpha^T\mathbf{1} - \frac{1}{2}\alpha^T\mathbf{Y}^TQ\mathbf{Y}\alpha$, in a gradient descent algorithm to iteratively find a satisfactory set of Lagrange multipliers α. Initialize α as a zero vector in $\mathbb{R}^n_+$.
3. **Prediction:** For prediction, follow these formulas:
   - For non-kernelized SVM:

$$\textbf{label}(\mathbf{x}_{\text{test}}) = \text{sign}(\mathbf{w}^T\mathbf{x}_{\text{test}}) = \text{sign}\left(\sum_{i=1}^{n}\alpha_i y_i(\mathbf{x}_i^T\mathbf{x}_{\text{test}})\right)$$

   - For kernelized SVM:

$$\textbf{label}(\mathbf{x}_{\text{test}}) = \text{sign}(\mathbf{w}^T\phi(\mathbf{x}_{\text{test}})) = \text{sign}\left(\sum_{i=1}^{n}\alpha_i y_i k(\mathbf{x}_i^T\mathbf{x}_{\text{test}})\right)$$

Soft-Margin SVM

Soft-Margin SVM extends the standard SVM algorithm to accommodate some misclassifications in the training data. This extension is particularly useful when dealing with non-linearly separable data. It introduces a regularization parameter (C) to control the balance between maximizing the margin and allowing for misclassifications.

The primal formulation for this extension can be expressed as:

$$\min_{\mathbf{w},\epsilon}\frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n}\epsilon_i \quad s.t. \quad (\mathbf{w}^T\mathbf{x}_i)y_i + \epsilon_i \geq 1; \quad \epsilon_i \geq 0 \quad \forall i$$

Here, C serves as a hyperparameter governing the trade-off between maximizing the margin and minimizing misclassifications. The additional variable ϵ_i represents the adjustment needed to satisfy the constraints.

Acknowledgments

Professor Arun Rajkumar: The content, including the concepts and notations presented in this document, has been sourced from his slides and lectures. His expertise and educational materials have greatly contributed to the development of this document.