MLT Week-10

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 Margins

Let dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be linearly separable with γ-margin where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$.

Let $w^* \in \mathbb{R}^d$ be the weight vector s.t. $(w^{*T} x_i) y_i \geq \gamma \ \forall i$.

Let some $R > 0 \in \mathbb{R}$, s.t. $\forall i \ ||x_i|| \leq R$.

Therefore, the number of mistakes made by the algorithm is given by,

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

Observations

Let w_{perc} be any weight vector which can linearly separate the dataset.

Therefore, we observe the following:

1. "Quality" of the solution depends on the margin.
2. Number of mistakes depend on w^*'s margin.
3. w_{perc} need not necessarily be w^*.

Hence, our **goal** should be to **find the solution that maximizes the margin**.

Margin Maximization

From the previous analysis, it is clear that a single dataset could have multiple linear classifiers with varying margins. The following diagram illustrates this phenomenon,

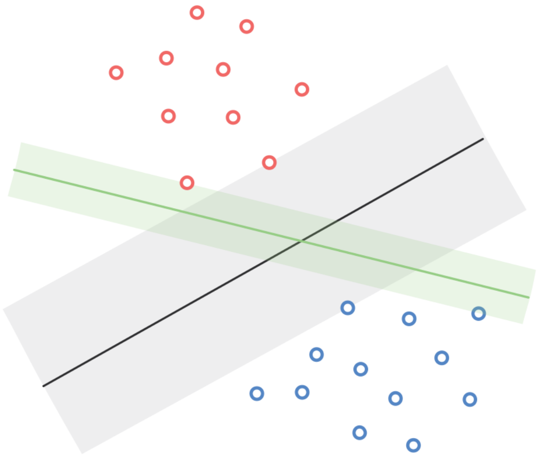

Figure 1: Multiple Classifiers

Therefore, for getting the best classifier, our goal can be written as,

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

$$s.t. (w^T x_i) y_i \geq \gamma \quad \forall i$$
$$||w||^2 = 1$$

The boundary of the margin is given by,

$$\{x : (w^T x_i) y_i = \gamma\}$$
$$\{x : (\frac{w^T}{\gamma} x_i) y_i = 1\}$$

From the above equation, we can see that γ depends on the width of w. Therefore, we reformulate our goal as,

$$\max_w \text{width}(w)$$

$$s.t. (w^T x_i) y_i \geq 1 \quad \forall i$$

Let the width be the distance between the two parallel margins, and let x and z be two points who are on the two lines exactly opposite to each other s.t. $w^T x = -1$ and $w^T z = 1$ or vice versa.

Let x_1 and x_2 be two points which lie on opposite side of the decision boundary as well as on the margins.

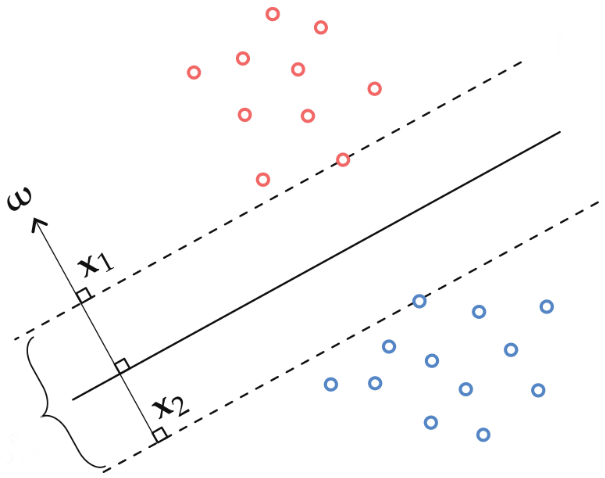

Figure 2: Margin Width

Therefore, the width is given by,

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

Therefore, our objective function can be written as,

$$\max_w \frac{2}{||w||_2^2} \quad s.t. (w^T x_i) y_i \geq 1 \quad \forall i$$

Equivalently,

$$\min_w \frac{1}{2} ||w||_2^2 \quad s.t. (w^T x_i) y_i \geq 1 \quad \forall i$$

Therefore finding the separating hyperplane with maximum margin is equivalent to finding the one with the smallest possible normal vector w.

Constrained Optimization

Let a constrained optimization problem be formulated as follows,

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

We can solve this problem using **Lagrange Multipliers**.

Lagrange multipliers are used in constrained optimization problems to find the optimal values of the objective function subject to a set of constraints. In Lagrange multipliers method, the constraints are incorporated into the objective function by introducing additional variables called Lagrange multipliers.

The Lagrange function $\mathcal{L}(x, \alpha)$, for our above function, is defined as follows:

$$\mathcal{L}(w, \alpha) = f(w) + \alpha g(w) \quad \forall w$$

where $\alpha \geq 0$.

Therefore, maximizing the Lagrange function w.r.t. α,

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

As the Lagrange function is equal to $f(w)$ where $g(w) \leq 0$, we can rewrite our original function as,

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

In general, we cannot swap the min and max functions unless all the functions involved are convex functions. Hence, as both f and g are convex functions in our example, we can rewrite them as follows,

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

Let's rewrite the above constrained optimization problem with m constraints $g_i(w) \leq 0$ where $i \in [1, m]$. This can written as,

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

Formulating the Dual Problem

Our Objective function is as follows,

$$\min_w \frac{1}{2}||w||_2^2 \quad s.t.(w^T x_i)y_i \geq 1 \quad \forall i$$

The constraints can be written as follows,

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

Let $\alpha \in \mathbb{R}^d$ be the Lagrange multipliers, and let our Lagrange function be written as,

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

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

Solving for the inner function of the dual problem, we get,

$$w_\alpha^* - \sum_{i=1}^{n} \alpha_i x_i y_i = 0$$

$$\therefore w_\alpha^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

Rewriting the above equation in vectorized form, we get,

$$w_\alpha^* = XY\alpha \quad ...[1]$$

where $X \in \mathbb{R}^{d\times n}$, $Y \in \mathbb{R}^{n\times n}$, and $\alpha \in \mathbb{R}^n$. X is the dataset, Y is the label diagonal matrix, where the diagonals are the labels. Rewriting the outer dual function, we get,

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

Observations:

1. As the dual problem solves for $\alpha \geq 0$, its variable dimension is in $\mathbb{R}^n_+$, while as the primal problem solves for w, its variable dimension is in $\mathbb{R}^d$.
2. Solving the dual problem is "easier".
3. As the dual problem depends on $X^T X$, it can be kernalized.

Some observations regarding the following equation,

$$w^*_\alpha = \sum_{i=1}^{n} \alpha_i x_i y_i$$

1. The optimal w^* is the linear combination of the datapoints where the importance of each datapoint is given by α_i for the i^{th} point.
2. Hence, there are points that are more important that others.

Support Vector Machine

Revisiting the Lagrangian function,

$$\min_w \left[\max_{\alpha \geq 0} f(w) + \alpha g(w)\right] \equiv \max_{\alpha \geq 0} \left[\min_w f(w) + \alpha g(w)\right]$$

The primal function is represented on the left-hand side of the equation, while the right-hand side represents the dual function. w^* and α^* are the solutions derived for the primal and dual functions, respectively. When these solutions are inserted into the equation, we obtain,

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

But as $g(w^*) \leq 0$, the left hand side equates to $f(w^*)$.

$$f(w^*) = \min_w f(w) + \alpha^* g(w)$$

Substituting w^* for w in the right-hand side of the equation would result in a new right-hand side that is greater than or equal to the current one.

$$f(w^*) \leq f(w^*) + \alpha^* g(w^*)$$

$$\therefore \alpha^* g(w^*) \geq 0 \quad ... [1]$$

But, according to our constraints, $\alpha^* \geq 0$ and $g(w^*) \leq 0$.

$$\therefore \alpha^* g(w^*) \leq 0 \quad ... [2]$$

From [1] and [2], we get,

$$\alpha^* g(w^*) = 0$$

Rewriting the equation for multiple constraints, we get,

$$\alpha^*_i g(w^*_i) = 0 \quad \forall i$$

Therefore, if one of the two is greater than zero, the other equals zero. We know that $g(w^*) = 1 - (w^T x_i) y_i$.

$$\alpha^*_i (1 - (w^T x_i) y_i) = 0 \quad \forall i$$

As the importance of the i^{th} datapoint is given by α_i, if $\alpha_i > 0$, we get,

$$(w^T x_i)y_i = 1$$

which means that the i^{th} datapoint lies on the **"Supporting"** hyperplane and contributes to w^*.

Therefore, the datapoints whose $\alpha_i > 0$ are known as **Support Vectors** and this algorithm is known as **Support Vector Machine**.

Support Vector Machines (SVMs) are a type of supervised learning algorithm used for classification and regression analysis. SVMs aim to find the optimal hyperplane that separates data points from different classes with the maximum margin. In the case of non-linearly separable data, SVMs use kernel functions to transform the data into a higher-dimensional space, where a linear decision boundary can be used to separate the data.

Insight: w^* is a sparse linear combination of the data points.

Hard-Margin SVM Algorithm

This algorithm only works if the dataset is linearly separable with a $\gamma > 0$.

1. Calculate $Q = X^T X$ directly or using a kernel as per the dataset.
2. Use the gradient of the dual formula ($\alpha^T 1 - \frac{1}{2}\alpha^T Y^T Q Y \alpha$), in the gradient descent algorithm to find a satisfactory α. Let the intial α be a zero vector $\in \mathbb{R}^n_+$.
3. To predict:
    - For non-kernelized SVM: label(x_{test}) = sign$(w^T x_{test})$ = sign $\left(\sum_{i=1}^n \alpha_i y_i (x_i^T x_{test})\right)$
    - For kernelized SVM: label(x_{test}) = sign$(w^T \phi(x_{test}))$ = sign $\left(\sum_{i=1}^n \alpha_i y_i k(x_i^T x_{test})\right)$

Soft-Margin SVM

Soft-Margin SVM is an extension of the standard SVM algorithm that allows for some misclassifications in the training data. This is useful when the data is not linearly separable, as it allows the SVM to still find a decision boundary that separates the classes as best as possible while allowing for some errors. The degree to which misclassifications are allowed is controlled by a regularization parameter(C), which is used to balance the trade-off between maximizing the margin and minimizing the number of misclassifications.

The primal formulation for this is given by,

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

where C is the hyperparameter that is used to balance the trade-off between maximizing the margin and minimizing the number of misclassifications, and ϵ_i is the additional value required to satisy the constraints.

Credits

Professor Arun Rajkumar: The content as well as the notations are from his slides and lecture.