Note Taking: Support Vector Machine(September 17th)

Minxuan Jin

GT ID: 904135121

1 Support Vector Machines

Several points need to be classified into two classes. Thus, a decision boundary which gives zero training error needs to be found.

The goal of SVM is to find a hyperplane such that the minimum distance from the data points to the hyperplane (i.e., the margin) is maximized.

We define the decision boundary as

$$w^T x + b = 0,$$

which represents a hyperplane. **Dash lines** are parallel to the decision boundary and they hit the data points.

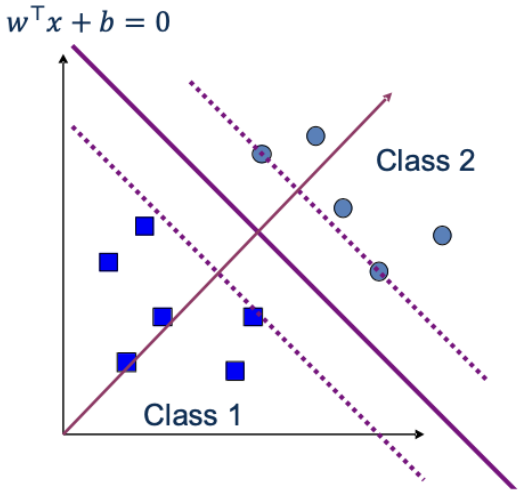

Figure 1: Linear Classifier

Select two points from each dash line respectively,

$$w^T x + b = c \tag{1}$$

$$w^T x + b = -c \tag{2}$$

Assume y = $w^T x + b$. If $y > c$, then this node will be classified as Class 2. If $y < c$, this node will be classified as Class 1. If $-c <= y <= c$ then this node will be in **unnormalized margin**.

The **margin** is defined as $\gamma = \frac{2c}{\|w\|}$. (margin: middle of the two dash lines)

From a geometric perspective, $w^T x$ represents the projection of a point x onto the normal vector w of the hyperplane.

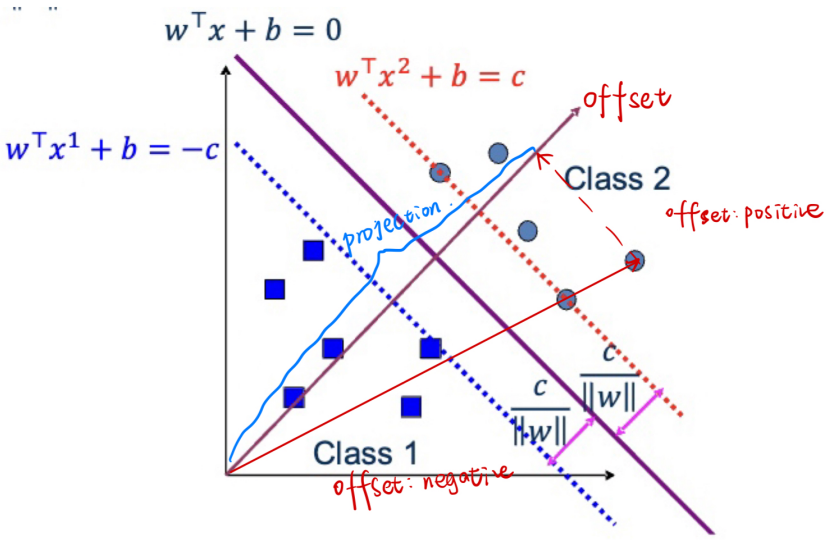

Figure 2: Offset

For all x in Class 2, $y = 1, w^T x + b \geq c$.

For all x in Class 1, $y = -1, w^T x + b \leq c$.

Thus, Formal definition of the linear equation can be defined as

$$(w^T x + b)y \geq c.$$

The margin between two classes need to be as large as possible. The **standard form** shows as:

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

$$s.t. 1 - y^i(w^T x_i + b) \leq 0,$$

The margin is defined as $\frac{2}{\|w\|}$.

2 Lagrangian Duality

The problem is formulated as:

$$\min_{w} f(w)$$

Subject to:

$$\begin{aligned} g_i(w) &\leq 0, \quad i = 1, 2, \ldots, k \\ h_i(w) &= 0, \quad i = 1, 2, \ldots, l \end{aligned}$$

We define the **Lagrangian Function**:

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w).$$

The Lagrangian function transforms the original "constrained optimization problem" into an "unconstrained optimization problem," where $\alpha_i g_i(w)$ and $\beta_i h_i(w)$ are constrainer.

If there exists some saddle points of L, then the saddle point follow **KKT** conditions:

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \\ g_i(w) \le 0, \quad h_i(w) = 0 \\ \alpha_i \ge 0 \\ \alpha_i g_i(w) = 0 \end{cases}$$

For any $\alpha_i \ge 0$ and β_i,

$$\min_{\substack{w \text{ feasible} \\ g_i(w) \le 0 \\ h_i(w) = 0}} f(w) \ge \inf_w L(w, \alpha, \beta)$$

Why?

For every possible $\tilde{w}$ in the primal problem, it satisfies

$$\forall i: \quad g_i(\tilde{w}) \le 0 \quad \text{and} \quad h_i(\tilde{w}) = 0.$$

Put these restrictions back to Lagrangian Function, consider that $\alpha_i \ge 0$:

$$L(\tilde{w}, \alpha, \beta) = f(\tilde{w}) + \sum_i \underbrace{\alpha_i}_{\ge 0} \underbrace{g_i(\tilde{w})}_{\le 0} + \sum_i \beta_i \underbrace{h_i(\tilde{w})}_{=0}.$$

Because $\sum_i \alpha_i g_i(\tilde{w})$ is non-positive, and $\sum_i \beta_i h_i(\tilde{w})$ equals to zero, we can prove that

$$L(\tilde{w}, \alpha, \beta) \le f(\tilde{w}).$$

$$\inf_w L(w, \alpha, \beta) \le L(\tilde{w}, \alpha, \beta).$$

Thus, we can get

$$\inf_w L(w, \alpha, \beta) \le \min_{w \text{ is feasible}} f(w).$$

$$\min_{\substack{w \text{ feasible} \\ g_i(w) \le 0 \\ h_i(w) = 0}} f(w) \ge \max_{\substack{\alpha, \beta \text{ feasible} \\ \alpha_i \ge 0}} \inf_w L(w, \alpha, \beta)$$

Why?

From the first inequality, we know that for every pair (α, β) satisfying $\alpha_i \ge 0$, we obtain a lower bound $g(\alpha, \beta) = \inf_w L(w, \alpha, \beta)$.

This process of "finding the greatest lower bound" is itself a new optimization problem:

$$\max_{\alpha_i \ge 0, \beta} g(\alpha, \beta) \quad \text{which is equivalent to} \quad \max_{\alpha_i \ge 0, \beta} \left(\inf_w L(w, \alpha, \beta) \right)$$

Let the optimal value of the primal problem be p^* and the optimal value of the dual problem be d^*:

$$p^* = \min_{w \text{ is feasible}} f(w)$$

$$d^* = \max_{\alpha_i \ge 0, \beta} g(\alpha, \beta)$$

For any valid (α, β), we have $p^* \ge g(\alpha, \beta)$, which implies that p^* is an upper bound for the set $\{g(\alpha, \beta) \,|\, \alpha_i \ge 0, \forall i\}$.

Therefore, p^* must be greater than or equal to the maximum element in this set, which is d^*.

$$p^* \ge d^*.$$

3 SVM Dual Problem

The Lagrangian function is

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

Dual objective is $g(\alpha) := \inf_{w,b} L(w, b, \alpha)$.

To find optimal w and b, we can take Lagrangian Function's derivative: $\begin{cases} \frac{\partial L}{\partial w} = w^* - \sum_{i=1}^{n} \alpha_i y^i x^i = 0 \\ \frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y^i = 0 \end{cases}$

Put above two condition back to Lagrangian function, we can get equation:

$$\max_{\alpha} g(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j ((x^i)^T x^j)$$

$$s.t. \alpha_i \geq 0, \forall i = 1, 2, ..., m$$

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

concave, global maximum can be found.

4 Inference and Support Vectors

From KKT condition: $\alpha_i g_i(w) = 0$. For data points with $(1 - y^i(w^T x^i + b)) < 0$, then $\alpha_i = 0$. For data points with $(1 - y^i(w^T x^i + b)) = 0$, then $\alpha_i \geq 0$.

Data points whose α_i' s are non-zero → **Support Vectors**. Those points support the margins.

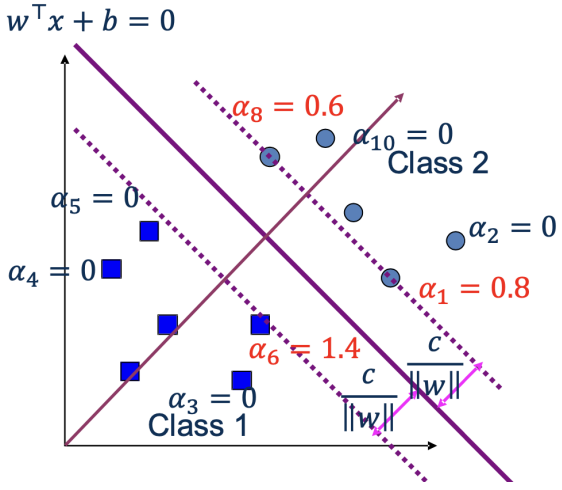

Figure 3: Support Vector

Since we've already had $w = \sum_{i=1}^{n} \alpha_i y^i x^i$, pick any data with $\alpha_i > 0$, compute b by

$$1 - y^i(w^T x^i + b) = 0.$$

Since w and b are computed, for any new data point z, we can compute the value of $w^T z + b$. If the result is positive, classify z as Class 1. If the result is negative, classify z as Class 2.

This model is highly efficient because its final decision function is determined only by a small number of support vectors rather than all the training data. This makes SVMs **sparse** in representation and memory-efficient.

5 Kernel SVM

All the previous discussion is based on the assumption that the data is linearly separable. But what if the data cannot be separated by a straight line?

The key idea is that instead of drawing complex curves in the original space to separate the data, we map the data into a **higher-dimensional space** where it becomes separable by a plane. This is the intuition behind the kernel method.

We define Kernel SVM as

$$\max_{\alpha} g(\alpha) := \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j).$$

$K(x_i, x_j)$ measure the similarity between x_i and x_j.

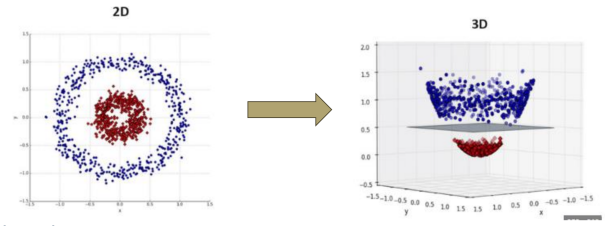

Figure 4: Projection to higher-dimensional place

$K(x_i, x_j)$ is equivalent to first mapping x_i and x_j into a higher-dimensional space through a mapping function $\phi(\cdot)$, and then taking their inner product, i.e.,

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle.$$

The most remarkable part is that we do not need to know the explicit form of ϕ, nor perform any actual mapping. We only need to compute the value of $K(x_i, x_j)$, which greatly simplifies the computation.