

# Support Vector Machines

Hossein Falaki  
Computer Science Department,  
University of California, Los Angeles  
falaki@cs.ucla.edu

## 1 Introduction

This note introduces Support Vector Machines as an effective supervised learning technique. We will present the geometric interpretation of the SVM primal and dual problems. We will derive the dual problem in a more general setting where points are not separable using KKT theory. Finally we will introduce the use of the Kernel trick in SVM.

## 2 Support Vector Machines

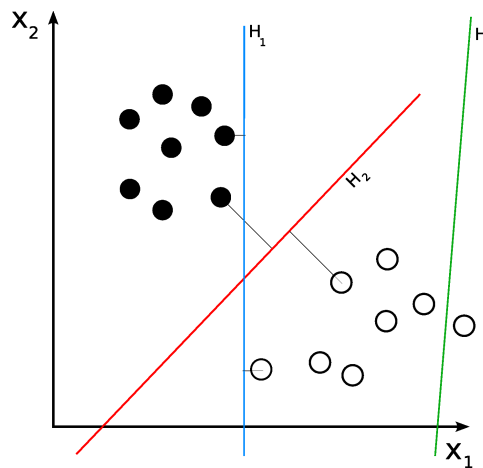


Figure 1: A separating hyperplane with no margin can be misleading.

As it can be seen in Figure 2 a separating hyperplane can fully separate the training data. But this zero training error does not mean it offers a good generalization. For example  $H_1$  in Figure 2 will not offer a good testing error.

An intuitive solution to this problem is separating data points with a hyperplane that has a margin as in Figure 2. If the margin is large enough

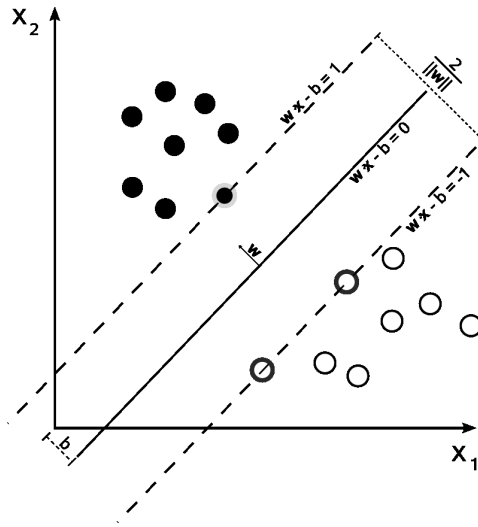


Figure 2: Separating training data points with the maximum possible margin.

we will be insured against testing error. This idea is the basis of *Support Vector Machines*:

$$\text{SVM classifier: } \text{sgn}(\langle x, w \rangle + b)$$

Where  $w$  is the weight and  $b$  is the bias of this SVM.

We can formulate the SVM problem as following:

$$\min ||w||^2$$

Subject to:

$$\text{goal: } y_i \text{sgn}(\langle x, w \rangle + b) \geq 1$$

In other words:

$$\begin{cases} y_i = + : & \langle x_i, w \rangle + b \geq 1 \\ y_i = - : & \langle x_i, w \rangle + b \leq -1 \end{cases}$$

We only care about the sign, therefore the scale does not matter, Therefore we can write:

$$\begin{cases} y_i = + : & \langle x_i, \frac{w}{||w||} \rangle + b \geq \frac{1}{||w||} \\ y_i = - : & \langle x_i, \frac{w}{||w||} \rangle + b \leq -\frac{1}{||w||} \end{cases}$$

With this normalization, the length of the margin is  $\frac{1}{||w||}$ . Therefore to maximize the margin, we should minimize  $||w||$ .

## 2.1 Dual Problem

Considering the margin maximization problem as the primal, we can construct the dual problem in this way:

1. Find the convex hull of the positive and negative points
2. Find the two points, each in one of the convex shapes, nearest to each other.
3. Cut the of the line segment as the margin.

A polygon can be formulated as  $\sum c_i x_i$ , if  $x_i$  are all the corner points. Therefore the dual problem is:

$$\min_{c_i, c_j} \left\| \sum_{i \in +} c_i x_i - \sum_{j \in -} c_j x_j \right\|^2$$

subject to:

$$\begin{aligned} c_i &\geq 0, & \sum_{i \in +} c_i &= 1 \\ c_j &\geq 0, & \sum_{j \in -} c_j &= 1 \end{aligned}$$

## 2.2 Geometric Argument

Here is a geometric argument that the dual and the primal problems are equivalent.

- If we have two parallel separating planes of distance  $M$ , the distance of any  $+$  point with any  $-$  point is at least  $M$ . Therefore:

$$\text{min distance} \geq \text{max margin}$$

- If we have two points of minimum distance, one in each polygon, then all the other points of the two polygons are behind the planes that go through these two points. Therefore:

$$\text{max margin} \geq \text{min distance}$$

## 3 Karush-Kuhn-Tucker Formulation

Originally SVM was developed using the Karush-Kuhn-Tucker (KKT) method. Here we present the general formulation of the problem, when the points are not completely separable.

Geometrically, when the positive and negative points are not completely separable, we can shrink the polygons. This can be achieved through using a set of *slack variables*.

$$\min \frac{1}{2} \|w\|^2 + c \sum_{i=1}^n \xi_i$$

Subject to:

$$y_i (\langle x_i, w \rangle + b) \geq 1 - \xi_i, \quad i = 1, \dots, n$$

To derive the dual problem using KKT saddle points condition:

$$\frac{1}{2} \|w\|^2 + c \sum_i \xi - \sum_{i=1}^n \alpha_i ((y_i \langle w, x_i \rangle + b) - 1 + \xi_i) - \sum_{i=1}^n \beta_i \xi_i$$

In short, this formula is a Lagrangian with both the primal and dual variables:

$$\text{Lagrangian}(\underbrace{w, b, \xi_i}_{\text{primal}}, \underbrace{\alpha_i, \beta_i}_{\text{dual}}), \quad \alpha_i, \beta_i \geq 0$$

To get the dual formulation, lets separate the primal variables:

$$L = \frac{1}{2} \|w\|^2 - \langle w, \sum_{i=1}^n \alpha_i y_i x_i \rangle - b \sum_{i=1}^n \alpha_i y_i + \sum_{i=1}^n \xi_i (c - \alpha_i - \beta_i) + \sum_{i=1}^n \alpha_i$$

By taking the partial derivatives we find the value of the primal variables at the optimal point:

$$\frac{\partial L}{\partial w} : w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \rightarrow w = \sum_{i=1}^n \alpha_i y_i x_i$$

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

$$\frac{\partial L}{\partial \xi_i} : c - \alpha_i - \beta_i = 0$$

By substituting  $w = \sum_{i=1}^n \alpha_i y_i x_i$  we get:

$$\begin{aligned} L &= \frac{1}{2} \|w\|^2 - \langle w, w \rangle + 0 + 0 + \sum_i \alpha_i \\ &= -\frac{1}{2} \left\| \sum_i \alpha_i y_i x_i \right\|^2 + \sum_i \alpha_i \end{aligned}$$

Therefore the dual formulation is:

$$\max L = -\frac{1}{2} \left\| \sum_i \alpha_i y_i x_i \right\|^2 + \sum_i \alpha_i$$

subject to:

$$\begin{aligned} \sum \alpha_i y_i &= 0 \\ 0 &\leq \alpha_i \leq c \end{aligned}$$

## 4 Expanding the Features

Sometimes the training data is not separable with any hyperplane in the feature space. In this case, expanding the feature space is helpful.

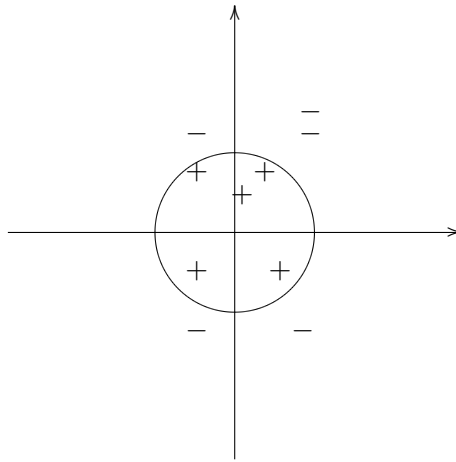


Figure 3: The training points may not be separable in the original feature space. But if we consider  $x_1^2$  and  $x_2^2$  SVM works.

One good way to expand the features,  $x_1, x_2, \dots, x_n$ , to all the possible combinations, e.g.,  $x_1^2, x_1 x_2, \dots$ , using  $(1 + \langle x_i, x \rangle)^2$  instead of simply  $\langle x_i, x \rangle$ .

In this case the SVM formulation would be:

$$SVM : \text{sgn}(\sum \alpha_i y_i (1 + \langle x_i, x \rangle)^2 + b)$$

This is commonly written as:

$$SVM : \text{sgn}(\sum \alpha_i y_i k(x_i, x) + b)$$

A more general and common kernel form is the Gaussian Radial basis function:

$$k(x_i, x) = e^{-\frac{1}{2\sigma^2} \|x - x_i\|^2}$$

This is a form of similarity measure. In other words, SVM is using the support vectors as a committee of voters. The method compares the new

point with the “important points” (i.e., support vector points), and sums their weighted votes. The overall vote is reported as the final result. In this respect, SVM is not learning much about the data, and is considered a black-box learner.

## **5 Acknowledgment**

This note is based on Prof. Ying Nian Wu’s lectures on Theoretical Statistics at UCLA. Figures 1 and 2 are from the Internet.