5 Support Vector Machine

Motivated by the statistical learning theory that Vapnik and Chervonenkis developed, Vapnik and his collaborators proposed the optimal separating hyperplane and its nonlinear generalization for pattern recognition in the early 90's. This new classification technique is called the support vector machine (SVM). Its theoretical motivation lies in the control of such capacity measures of a class of decision rules as the V-C dimension to ensure better generalization of an estimated rule.

Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. Note that ± 1 symmetric class labels are used instead of 1 or 0. We begin with hyperplanes for discrimination of the two classes. We will first consider the situation where the training data are linearly separable. The non-separable case will be considered in the following section. The SVM looks for an "optimal" separating hyperplane $\beta' x + \beta_0 = 0$. The associated rule in this case is $f(x) = \text{sign}(\beta' x + \beta_0)$.

5.1 Perceptron Algorithm

As a precursor of the SVM, the *perceptron algorithm* finds a separating hyperplane by sequentially updating β and β_0 of a linear classifier, $f(x) = \text{sign}(\beta' x + \beta_0)$. Given $\mathcal{D}_n = \{(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\} : i = 1, \ldots, n\}$, it consists of two steps:

Step 1. Initialize $\beta^{(0)} = 0$ and $\beta_0^{(0)} = 0$.

Step 2. While there is a misclassified point such that $y_i({\beta^{(m-1)}}' x_i + \beta_0^{(m-1)}) \leq 0$ for $m = 1, 2, \ldots,$ repeat the following.

- Choose a misclassified point (x_i, y_i).
- Update $\beta^{(m)} = \beta^{(m-1)} + y_i x_i$ and $\beta_0^{(m)} = \beta_0^{(m-1)} + y_i$.

Novikoff (1962) proved the following theorem on the bound of the number of corrections of the perceptron algorithm.

Theorem 11. *Let $R = \max_i ||x_i||$. Suppose that $\mathcal{D}_n$ is separable so that for some $w \in \mathbb{R}^d$ with $||w|| = 1$ and $b \in \mathbb{R}$, $\delta = \min_i y_i(w' x_i + b) > 0$. Then the algorithm terminates within $\lfloor (R^2 + 1)(b^2 + 1)/\delta^2 \rfloor$ iterations, where z is the largest integer that does not exceed $\lfloor z \rfloor$.*

5.2 Separable Case

If the training data $\mathcal{D}_n$ are linearly separable, then there exist $\delta > 0$, β_0 and β such that

$$\begin{aligned} \beta' x_i + \beta_0 &\geq \delta \quad \text{for } y_i = 1 \text{ and} \\ \beta' x_i + \beta_0 &\leq -\delta \quad \text{for } y_i = -1. \end{aligned}$$

Since β_0 and β can be normalized, without loss of generality, δ is set to 1. Then we have the following *separability condition.*

$$y_i(\beta' x_i + \beta_0) \geq 1 \quad \text{for all } i = 1, \ldots, n \tag{5.1}$$

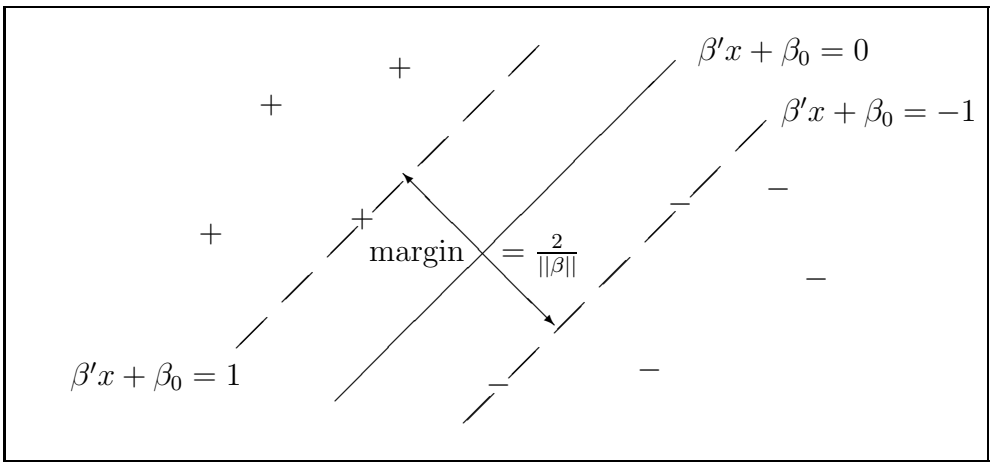

Figure 4: Separating hyperplane

The distance of a point $x_0 \in \mathbb{R}^d$ from a hyperplane $\beta'x + \beta_0 = 0$ is given by

$$\frac{|\beta'x_0 + \beta_0|}{||\beta||}.$$

Given the separating hyperplane, take the distance between the two convex hulls determined by x_i's with class labels 1 and -1, respectively. The distance is the same as the sum of the distances from the nearest x_i's with $y_i = \pm 1$ to the hyperplane. It is called the geometric margin between the two classes. Under the specified normalization of the separating hyperplane, the margin is given by

$$\frac{1}{||\beta||} + \frac{1}{||\beta||} = \frac{2}{||\beta||}.$$

See Figure 4. Among infinitely many separating hyperplanes, the linear SVM finds the hyperplane with the maximum margin. In other words, the optimal hyperplane for the SVM is defined to

$$\underset{\beta_0,\beta}{\text{maximize}} \ \frac{2}{||\beta||} \ \text{ subject to } y_i(\beta'x_i + \beta_0) \geq 1 \text{ for all } i = 1, \ldots, n; \tag{5.2}$$

or equivalently

$$\underset{\beta_0,\beta}{\text{minimize}} \ \frac{1}{2}||\beta||^2 \ \text{ subject to } y_i(\beta'x_i + \beta_0) \geq 1 \text{ for all } i = 1, \ldots, n. \tag{5.3}$$

Once we have $(\hat{\beta}_0, \hat{\beta})$, the SVM solution to the optimization problem is given as

$$\hat{f}(x) = \text{sign}(\hat{\beta}'x + \hat{\beta}_0).$$

So, why should we maximize the margin? To state a theoretical justification for that in the context of the V-C theory, consider a class of rules, say, $\mathcal{F}$, corresponding to δ-margin separating hyperplanes with $||\beta|| = 1$:

$$f(x) = \begin{cases} 1 & \text{if } \beta'x + \beta_0 \geq \delta \\ -1 & \text{if } \beta'x + \beta_0 \leq -\delta. \end{cases} \tag{5.4}$$