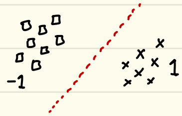
Support Vector Machines (SVMs)

among the best off-the-shelf supervised learning algorithm

They rely on the notion of 'margin' and try to find a classifier with a large 'margin'

SVM is closely related to logistic regression in that both try to find a hyperplane that separates the data

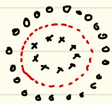


The differences, however, are : (1) The cost functions that are optimized in SVM and LR are different:

SVM : maximizes the worst 'margin'

LR : maximizes the likelihood

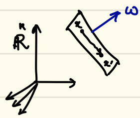
(2) SVM is more versatile and can handle data that are not linearly separable, instead can be separated by a nonlinear decision boundary.



* Given a dataset $D = \{(x^1, y^1), \dots, (x^n, y^n)\}$ where $x^i \in \mathbb{R}^K$ and $y^i \in \{-1, +1\}$, find a decision boundary $w^T x + b = 0$ that best separate the data with 'maximum worst margin'.

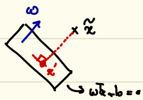
For now, let's assume only two classes, and data separable by a hyperplane.

(I) $\underset{\mathbb{R}^K}{w}^T \underset{\mathbb{R}^K}{x} + \underset{\mathbb{R}}{b} = 0 \longrightarrow$ specifies the equation of a hyperplane :



Assume x, x' are on the hyperplane $\begin{cases} w^T x + b = 0 \\ w^T x' + b = 0 \end{cases} \Rightarrow w^T(x - x') = 0 \rightarrow$ Thus, w is the normal to the hyperplane and b is the offset from origin.

(II) what is the distance of a point $\tilde{x}$ to hyperplane $w^T x + b = 0$



$$\begin{cases} \tilde{x} - x' \text{ is orthogonal to the plane} \longrightarrow \tilde{x} - x' \text{ is parallel to } w \rightarrow \tilde{x} - x' = \alpha w \\ x' \text{ is on the plane} \longrightarrow w^T x' + b = 0 \quad (2) \end{cases}$$

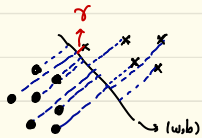
$$x' = \tilde{x} - \alpha w \quad (1)$$

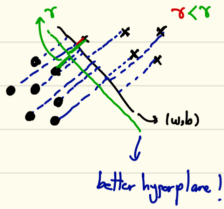
$$(1),(2) \Longrightarrow w^T(\tilde{x} - \alpha w) + b = 0 \longrightarrow w^T \tilde{x} + b = \alpha \|w\|^2 \longrightarrow \alpha = \frac{w^T \tilde{x} + b}{\|w\|^2}$$

5

The distance of $\tilde{x}$ to $\omega^T x + b = 0$ is $\|\tilde{x} - x'\|_2 = \|\alpha\,\omega\|_2 = |\alpha|\,\|\omega\|_2 = \dfrac{|\omega^T \tilde{x} + b|}{\|\omega\|_2}$

* Given $D = \{(x^1, y^1), \ldots, (x^N, y^N)\}$, we have





Distance of each x^i from $\omega^T x + b = 0$ is $\dfrac{|\omega^T x^i + b|}{\|\omega\|_2}$

Let $\gamma(\omega, b) \triangleq \min\limits_{i=1,\ldots,N} \dfrac{|\omega^T x^i + b|}{\|\omega\|_2} \implies \forall i = 1, \ldots, N : \dfrac{y^i(\omega^T x^i + b)}{\|\omega\|_2} \geqslant \gamma$

We want to find a hyperplane (ω, b) for which $\gamma(\omega, b)$ (the minimum margin is maximized)

$$\begin{cases} \max\limits_{\omega, b, \gamma} \ \gamma \\ \text{s.t.} \ \dfrac{y^i(\omega^T x^i + b)}{\|\omega\|_2} \geqslant \gamma \qquad , \forall i = 1, \ldots, N \end{cases}$$

Let $\bar{\gamma} \triangleq \|\omega\|_2 \, \gamma$

We can write the above maximization in the following equivalent form:

$$\begin{cases} \max\limits_{\omega, b, \bar{\gamma}} \ \dfrac{\bar{\gamma}}{\|\omega\|_2} \\ \text{s.t.} \ y^i(\omega^T x^i + b) \geqslant \bar{\gamma} \ , \forall i = 1, \ldots, N \end{cases}$$

If $(\omega^*, b^*, \bar{\gamma}^*)$ a solution, then $(\alpha \bar{\gamma}^*, \alpha \omega^*, \alpha b^*)$ also a solution for any $\alpha > 0$.

We can fix this by enforcing $\bar{\gamma}^* = 1 \longrightarrow (\omega^*, b^*, 1) \longrightarrow$ can not choose any $\alpha \neq 1$.

Vanilla SVM

$$\longrightarrow \begin{cases} \max\limits_{(\omega, b)} \ \dfrac{1}{\|\omega\|_2} \\ \text{s.t.} \ y^i(\omega^T x^i + b) \geqslant 1 \ , \forall i = 1, \ldots, N \end{cases} \longrightarrow \begin{cases} \min\limits_{\omega, b} \ \frac{1}{2}\|\omega\|_2^2 \\ \text{s.t.} \ y^i(\omega^T x^i + b) \geqslant 1 \ , \forall i = 1, \ldots, N \end{cases}$$