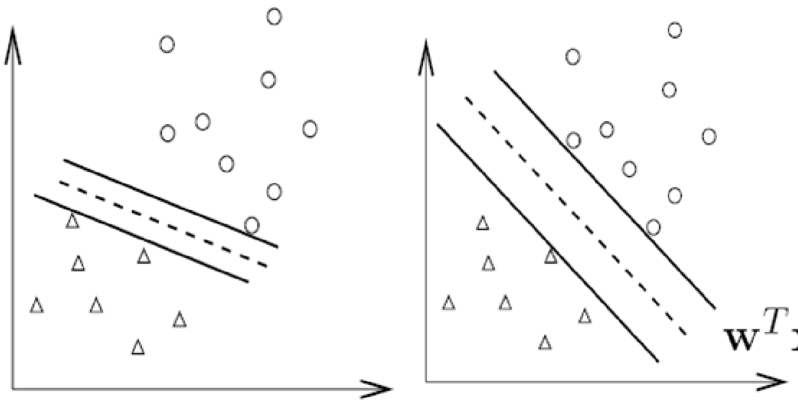
Support Vector Machine 簡介

By Biano

1. 資料分群 (Data Classification)

對於一群資料而言，有時候我們會希望依據資料的一些特性來將這群資料分為兩群。而就資料分群而言，我們已知有一些效果不錯的方法。例如：Nearest Neighbor、類神經網路(Neural Networks)、Decision Tree 等等方式，而如果在正確的使用的前提之下，這些方式的準確率相去不遠，然而，SVM 的優勢在於使用上較為容易。

2. Support Vector Machine 概念



$$\mathbf{w}^T\mathbf{x} + b = \begin{bmatrix} +1 \\ 0 \\ -1 \end{bmatrix}$$

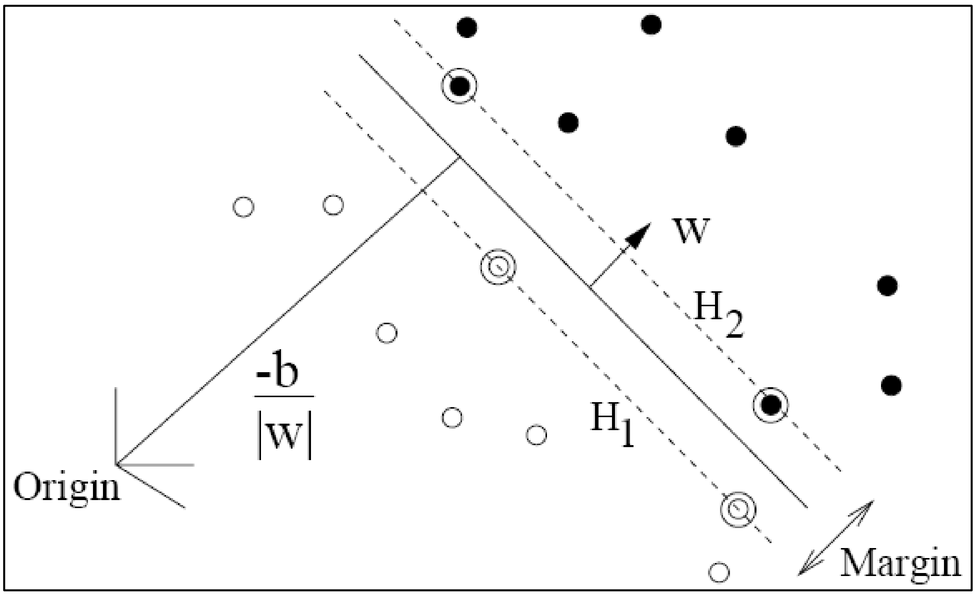
對於一群在 $\mathbf{R}^d$ 空間中的資料，我們希望能夠在該空間之中找出一 Hyper-plan，並且，希望此 Hyper-plan 可以將這群資料切成兩群(ie:群組 A、群組 B)。而屬於群組 A 的資料均位於 Hyper-plan 的同側，而群組 B 的資料均位於 Hyper-plan 的另一側，如下圖：

比較上圖左以及右，我們可以發現左圖所找出的 Hyper-plan(虛線)，其兩平行且與兩組資料點相切的 Hyper-plan(實線)之間距離較近，而右圖則具有較大的 Margin。而由於我們希望可以找出將兩群資料點分的較開的 Hyper-plan，因次我們認為右圖所找出的是比較好的 Hyper-plan。

因此，將問題簡述如下：

已知 Training Data Sets:：$\{x_i, y_i\}, i = 1, 2, ..., n$ ， $x_i \in \mathbf{R}^d, y_i \in \{1, -1\}$ ，我們希望利用 Training Data 找出一最佳 Hyper-plan H，以利將未知的 Xi 歸類。

3. SVM 理論說明：Preliminaries



由上圖，實線為我們找出的 Hyper-plan，而我們將 H1 與 H2 稱之為 Support Hyper-plans，而我們希望能夠找出最佳的 Classification Hyper-plan 使兩 Support Hyper-plans 之間有最大的 Margin。

將 Classification Hyper-plan(實線)定義為 $W^TX = -b$ ($W^TX + b = 0$)。因此，我們可以把 Support Hyper-plan(H1、H2)寫為：

$$H1: w^T x + b + \delta$$

$$H2: w^T x + b - \delta$$

而我們可以利用一個常數將 W、b 與 delta 做 Scaling，因此可以把上兩式重寫為：

$$H1: w^T x + b = 1$$

$$H2: w^T x + b = -1$$

而同時，從 H1 到原點的距離爲：$|1-b|/\|w\|$；H2 到原點的距離爲：$|-1-b|/\|w\|$，因此 H1 與 H2 之間的距離爲：$\frac{2}{\|w\|}$。

利用以上的等式，而對於在 $\mathbf{R}^d$ 的空間來說，資料點必須滿足：

$\left(w^T x_i\right)+b\geq 1$，if $y_i=1$

$\left(w^T x_i\right)+b\geq 1$，if $y_i=-1$

我們可以將以上二式改寫爲 $y_i\left(\left(w^T x_i\right)+b\right)\geq 1$

同時，因爲希望兩個 Support Hyper-plans 之間距離爲最大，因此希望 $\frac{2}{\|w\|}$ 爲 Maximize，亦即：Minimize $\frac{\|w\|}{2}$

4. SVM 理論說明

由以上第三點，則 SVM 的問題可以歸納如下：

$$\begin{cases} \min_{w,b} \frac{1}{2} w^T w \\ y_i\left(\left(w^T x_i\right)+b\right)\geq 1 \end{cases}$$

若已知以上的限制式，雖然我們可以利用 Quadratic Programming 解出，然而，有時 Xi 可能本身維度很高(甚至有可能是無限多維的)，而這時我們會需要利用 Kernel($K(x,y)=\phi(x)^T\phi(y)$)以求得解。因此，我們利用 Lagrangian 將以上兩式轉爲：$\mathcal{L}$

$$\mathcal{L}(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i\left[y_i\left(w^T x_i+b\right)-1\right]$$

而爲符合原問題條件，必須滿足 $\min_w \max_\alpha \mathcal{L}(w,\alpha)$ ，因此我們分別對 w 以及 b 微分，並且得到以下：

- ✓ 對 w 微分：$w-\sum_i \alpha_i y_i x_i = 0$ ➔ $w^* = \sum_i \alpha_i y_i x_i$
- ✓ 對 b 微分：$\sum_i \alpha_i y_i$

並且將之代回 Lagrangian，則可得到如下的 Dual Problem：

- ✓ Maximize: $\mathcal{L}_D = \sum_{i=1} \alpha_i - \frac{1}{2}\sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j$
- ✓ Subject to: $\sum_i \alpha_i y_i = 0$
  $\alpha_i \geq 0 \forall i$

以上的 Dual Problem 同樣的也是一個 Quadratic Problem，同時，α_i 的數量會與 data-cases 的數量相同。此外，由上列兩式所表示的問題，其中所有跟 x_i 相關的係數，均為 Inner Product 的形式，也就是 $x_i^T x_j$。而這樣的形式有助於 Kernal 的運用。

5. KKT (Karush-Kuhn-Tucker) Conditions

由以上我們可知，由於 SVM 的問題是一個 Convex Problem，而對於一個 Convex Problem 來說，KKT conditions are necessary and sufficient to solve w、b、α。因此，對我們來說，求解上述的 SVM 的問題，及等同於下列的 KKT Condtions。因此我們依據以上推導的式子列出 KKT Conditions：

- ✓ $\partial_w \mathcal{L}_P = 0 \rightarrow w - \sum_i \alpha_i y_i x_i = 0$
- ✓ $\partial_b \mathcal{L}_P = 0 \rightarrow \sum_i \alpha_i y_i = 0$
- ✓ Constraint: $y_i\left(w^T x_i - b\right) - 1 \geq 0$
- ✓ Multiplier Condition: $\alpha_i \geq 0$
- ✓ Complementary Slackness: $\alpha_i\left[y_i\left(w^T x_i - b\right) - 1\right] = 0$

6. Support Vector and Overview

回到 SVM 最初的目的，我們希望能夠找出一個 Function， $f(g)$ ，而這個 Function 可以用來決定某筆資料的分類。也就是說，當我們已經找出一個如此的 $f(g)$ 時，只要代入某一筆未知分類的資料，即可辨識出該筆資料的類別。而由以上的推導，我們得知 $f(g)$ 的決定取決於 w 、 b ，而我們可以把 $f(g)$ 寫成如下：

$$y_i = f(x) = w^T x + b = \sum_i \alpha_i y_i x_i^T x + b$$

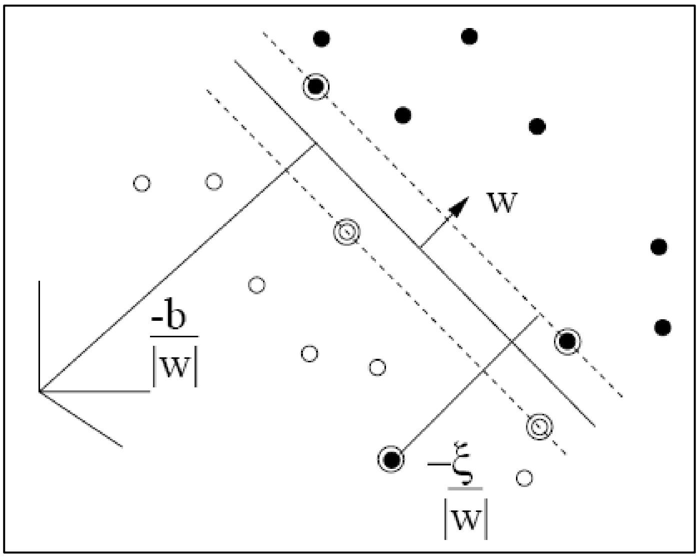
也可以把把上式寫成如下：

$$-b = \sum_j \alpha_i y_i x_j^T x_i - y_i$$

而，因此我們稱 Training Cases 中在 Support Hyper-plan 上的點作爲 Support Vectors，而同時也是這些點去決定 w 以及 b 的值，也就是最後求出的分類 Function。同時，由上兩式，我們也可以發現，只有 $\alpha_i > 0$ 的狀況之下，該資料才會對於 w 以及 b 的值有決定性的影響。

7. Non- Separable Cases

以上所有的推導，皆是狀況在較爲理想的狀態之下，也就是：我們可以完全的將兩類的資料分成兩群。然而，並非所有的分類問題狀況皆如此的單純，有時也會碰上無法將資料分類的問題。



如右圖：可以看到有一點黑點，雖然應該屬於黑色的分群，然而卻沒有辦法利用已知的 Hyper-plan(實線)做有效的分類，而因此，面對這樣的狀況我們通常會在解 SVM 問題時多加入一項考慮誤差的變數：ξ 。