# 支持向量機教學文件 (中文版)

李根逸

台灣大學通訊與多媒體實驗室

## Contents

# 1 支持向量機 (Support Vector Machine)

## 1.1 簡介

**支持向量機** (SVM: Support Vector Machine) 是一種可用來做**分類** (classfication)或**迴歸** (regression) 的方法。給予一群已經分類好的資料，支持向量機可以經由**訓練** (training) 獲得一組**模型** (model)。爾後，有尚未分類的資料時，支持向量機可以依據利用先前資料訓練出的模型去**預測** (predict) 這筆資料所屬的分類。

因為支持向量機在建立模型的時候，必須要先有已經分類好的資料作為訓練用，所以支持向量機是**監督式學習** (supervised learning) 的方法之一。

## 1.2 一個簡單的例子

以下我們舉一個極其簡化的例子來說明支持向量機的運作：

例一：已知六個人修習某門課的分數，以及是否通過的資訊：

| 座號 | 分數 | 通過與否 |
|---|---|---|
| 1 | 31 | 否 |
| 2 | 49 | 否 |
| 3 | 70 | 是 |
| 4 | 81 | 是 |
| 5 | 33 | 否 |
| 6 | 60 | 是 |

將這六個人的分數資料以及通過與否的分類結果送入支持向量機做為訓練用的資料，可以得到一組模型。這組模型在這個例子裡，就是以一個分數 59 來描述。爾後，我們可以使用這個模型來做預測，也就是說，分數大於 59 會被歸於通過這一類，反之則被歸於不通過這一類。將這組模型記錄下來後，以後當知道某個人修習這門課的成績時，便可以預測該人是否能夠通過這門課。

然而即便是如此簡單的例子，我們都不難產生以下的疑惑：

- 50 到 59 分都是可能用來做分類的分數，為何支持向量機會選出 59 而不是其他的分數？

- 因為訓練用的資料並沒有 50 到 59 分通過與否的資料，所以支持向量機怎麼知道 59 分就不會通過呢？
- 這個例子似乎假設教授只依照分數的高低來決定通過與否，如果教授只是使用電風扇或同時也參照了其他我們不知道的資料，只使用分數高低來預測教授的行為是否根本是錯的或者過於簡化？
- 同理，我們似乎假設了教授判斷的依據是固定不變的，這樣的假設是否太過強烈？

在思考這些問題之前，必須先體認到，支持向量機是一種**基於統計**的方法。也就是說，支持向量機只考慮它所得到的資料以及這些資料與分類結果之間的相關性。如果某類資料 (例: 分數) 與分類的結果 (例: 通過與否) 相關性很高的話，支持向量機就會基於統計的原則將該資料列為分類重要的依據。同樣地，支持向量機認為訓練用的資料跟做預測的資料有統計上共通的特性。支持向量機也的確無法知道教授是否以分數為標準，但因為擁有的資料只有分數而且分數跟分類結果之間有統計上的相關性，所以推論出分數 59 是統計上最有可能做為分類的依據[1]。換句話說，在這個例子裡, 如果座號跟分類結果有很大的相關性的話 (例: 座號大於 3 的人都通過，反之則沒有)，則將座號跟分數都當做資料送入支持向量機做訓練的話，支持向量機的確可能會將座號大於 3 與否當做預測分類的重要依據。而如果分數跟分類結果沒什麼相關性的話，甚至會完全不理會分數這種資料。

在下一個章節，我們會用比較正式的數學定義跟證明來說明支持向量機的運作原理，並舉一些比較複雜的例子。

# 2 支持向量機理論推導

## 2.1 簡介

支持向量機是一種**線性分類** (Linear Classification) 的方法，目的在於尋找一個可以使得所有訓練用而已經分類好的資料在**特徵空間** (feature space) 中可以將不同的**類別** (class)[2] 被清楚地分開的**超平面** (hyperplane)[3]。

在支持向量機中，所謂一筆分類好的資料包含了一個表示類別的整數和一組在特徵空間中的特徵向量。對例一來說，表示類別的整數可以用 +1 表示通過，-1 表示未通過，而特徵向量就是包含學生分數[4]的一維向量。我們

[1]何謂 "最有可能做為分類的依據"，在理論推導的時候有比較精確的說明

[2]注意此時我們討論的都只有兩個類別

[3]在 $n$ 維的向量空間裡，超平面即是一組 $n-1$ 維的線性方程式

[4]記得此時我們認為分類結果只跟分數有關 (跟座號或其它資料無關)

可以用下列的表來呈現這些資料：

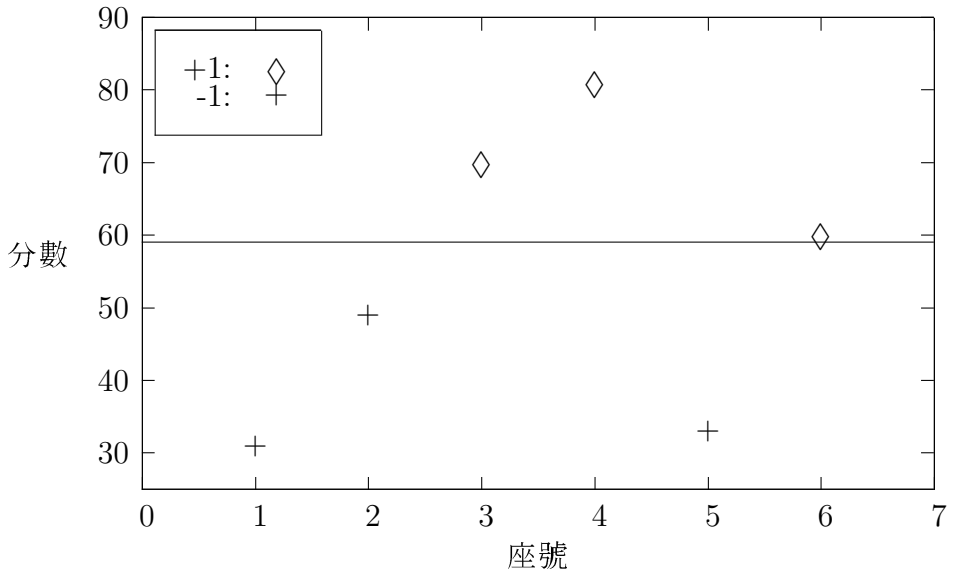
| 特徵向量 | 類別 |
|---|---|
| (31) | -1 |
| (49) | -1 |
| (70) | +1 |
| (81) | +1 |
| (33) | -1 |
| (60) | +1 |

在這個例子中，將這些當做訓練的資料送入支持向量機後找到的超平面就是 59 這個點。59 這個點可以在這個一維的特徵空間中將兩個類別分開。

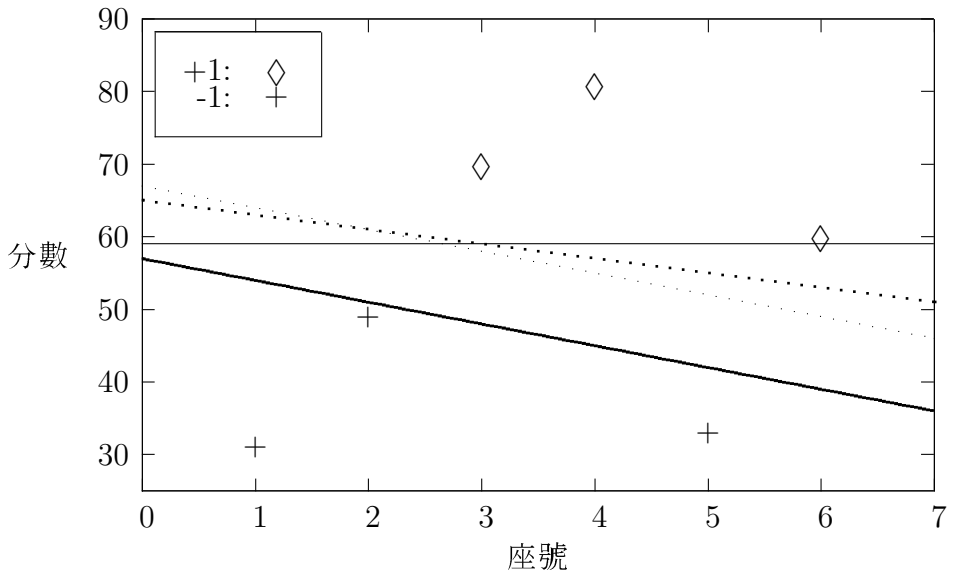
接著試著將座號也當做特徵向量的一部份我們可以用下列的表來呈現這些資料:

| 特徵向量 | 類別 |
|---|---|
| (1, 31) | -1 |
| (2, 49) | -1 |
| (3, 70) | +1 |
| (4, 81) | +1 |
| (5, 33) | -1 |
| (6, 60) | +1 |

將以上資料畫在圖上可以得到：



圖上的橫線代表分數為 59 分的分隔線。可以發現，即使將座號加入特徵向量，59 分仍然可以成功的將兩類資料完整分隔開。然而問題就出現了，其實有很多很多條直線[5]都可以成功的將這兩群資料分隔開。如下圖所示:




對於支持向量機而言，到底哪條線才是最好的呢？這將是我們所想要討論的。

## 2.2 理論推導

重新用正式的表示法描述我們的問題，我們的工作是預測一個測試的樣本是屬於兩個類別之中的哪一個。訓練用的樣本的格式是：$\{x_i, y_i\}, i = 1, ..., n$ 且 $x_i \in \mathbf{R}^d, y_i \in \{-1, +1\}$。其中，$\{x_i\}$ 稱為特徵向量而 $\{y_i\}$ 稱為**標籤** (label)。

如同之前的作法，我們先考慮一個非常簡單的例子。這個例子是**可線性分隔** (linear seperable) 的，也就是說，我們可以畫一條直線 $f(x) = w^T x - b$ 使得所有 $y_i = -1$ 的樣本會落在直線的一邊 ($f(x) > 0$) 而所有 $y_i = +1$ 的樣本會落在另一邊 ($f(x) < 0$)。於是，我們可以將新的測試樣本($x_{test}$)分類為 $y_{test} = sign(x_{test})$

然而如同之前所說的，有太多太多個這種可能的超平面。而在訓練時選擇不同的超平面會造成我們在測試時的準確度。舉例來說，我們可以選一條

[5]即是在二維空間中的超平面

非常靠近 $y_i = +1$ 類別裡所有成員的超平面。當測試時，直覺上，我們會很容易將 $y_i = -1$ 的樣本歸類為 $y_i = +1$ 而造成錯誤。因此，似乎我們找一個儘量可以將兩種類別樣本分隔成一樣遠的超平面會比較好。而顯然這個超平面剛好落在兩群樣本的中間。

幾何上，$w$ 向量跟 $w^T x = b$ 這超平面是**有向性正交** (directed orthogonal)。這點可以用下列敘述理解：首先讓 $b = 0$，則很明顯的所有向量 $x$ 跟 $w$ 向量的內積都為 0，也就是說，所有跟 $w$ 正交的向量都符合這條方程式。接著，將這超平面從原點移離向量 $a$，這超平面就會變成 $(x - a)^T w = 0$，也就是說，位移 $b = a^T w$ 就是 $a$ 在 $w$ 上的投影。

現在，我們定義另外兩個超平面，他們都平行於原本的分隔超平面。他們分別表示了訓練用樣本中，最靠近其中一種類別的超平面。我們將這兩個超平面稱為**支持超平面** (support hyperplane)，因為這兩個平面包含的資料向量支持這個平面。

定義從這兩個超平面到分隔超平面的距離分別為 $d_+$ 和 $d_-$，而**邊際** (margin) $\gamma$ 則為 $d_+ + d_-$。我們現在的目標是找到一個分隔超平面使得邊際最大，而此時，$d_+ = d_-$。

寫下這兩個支持超平面的方程式：

$$\mathbf{w}^T \mathbf{x} = b + \sigma$$
$$\mathbf{w}^T \mathbf{x} = b - \sigma$$

注意到我們現在已經**過度參數化** (over-parameterized) 這個問題：如果將 $\mathbf{w}$, $b$, 和 $\sigma$ 都乘上一個常數，這個 $x$ 的方程式將依然成立。為了去除模稜兩可的情況，我們需要讓 $\sigma = 1$。

我們現在也可以計算 $d_+ = (||b + 1| - b||)/||\mathbf{w}|| = 1/||\mathbf{w}||$ [6]。因此邊際大小等於此值的兩倍：$\gamma = 2/||\mathbf{w}||$

由上面對於支持超平面的定義，我們可以寫下分隔超平面需要符合的**限制**(contraint)：

$$\mathbf{w}^T \mathbf{x_i} - b \leq -1 \qquad \forall y_i = -1$$
$$\mathbf{w}^T \mathbf{x_i} - b \geq +1 \qquad \forall y_i = +1$$

或者直接寫成一個方程式：

$$y_i(\mathbf{w}^T \mathbf{x_i} - b) - 1 \geq 0$$

[6] 只有當 $b \notin (-1, 0)$ 時成立，否則原點並不會落在兩個超平面之間。如果 $b \in (-1, 0)$, 應該讓 $d_+ = (||b + 1| + |b||)/||w|| = 1/||\mathbf{w}||$

我們可以使支持向量機的**原始問題**(primal problem) 公式化變成：

$$\begin{array}{ll} \textbf{最小化} & \frac{1}{2}||\mathbf{w}||^2 \\ \textbf{並符合} & y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 \geq 0 \ \forall i \end{array}$$

也就是說，在所有訓練用樣本落在支持超平面某邊的限制下，我們最大化邊際大小。而剛好落在支持超平面上的樣本向量就被稱為**支持向量** (support vector)，因為它們支持這些超平面而且決定這個問題的解。

這個原始問題可以用**二次規劃** (quadratic program) 來解。但是這樣將難以**核化** (kernelized)，因為在核化的時候，不僅僅只是做向量內積而已。因此，我們將這問題經由**拉格朗日法** (Lagrangian) 變換為對偶問題 (dual problem)：

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

這個對偶問題的解是 $\min_{\mathbf{w}} \max_{\alpha} \mathcal{L}(\mathbf{w}, \alpha)$，也就是個求鞍點的問題。當原本的**目標函數** (object function) 是**凸** (convex)的，我們可以將最大化跟最小化先後順序交換。然後，我們可以找到在鞍點時 $\mathbf{w}$ 需符合的條件。這可以用分別對 $\mathbf{w}$ 和 $b$ 微分後求得：

$$\mathbf{w} - \sum_i \alpha_i y_i \mathbf{x_i} = 0 \implies \mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x_i}$$
$$\sum_i \alpha_i y_i = 0$$

將這些條件代回原式可以得到對偶問題 (dual problem)：

$$\begin{array}{ll} \textbf{最大化} & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T \mathbf{x_j} \\ \textbf{並符合} & \sum_i \alpha_i y_i = 0 \\ & \alpha_i \geq 0 \ \forall i \end{array}$$

這個對偶問題依然是個二次規劃的問題，但是可以注意到此時我們的變數 $\alpha_i$ 的個數跟訓練用樣本的個數 $N$ 一樣多。

最重要的一點是這個問題只跟 $\mathbf{x_i}$ 的內積，$\mathbf{x_i}^T\mathbf{x_j}$，有關。這樣就可以經由 $\mathbf{x_i}^T\mathbf{x_j} \to k(x_i, x_j)$ 做核化。

對偶性理論在原始問題為凸的時一定會成立，此時對偶問題則會是**凹** (concave) 的。更重要的是，原始問題的唯一解必相對於對偶問題的唯一解。事實上，$\mathcal{L}_{\mathcal{P}}(\mathbf{w}^*) = \mathcal{L}_{\mathcal{D}}(\alpha^*)$，也就是說**對偶性差距** (duality-gap) 為零。

下去我們開始討論這個問題的解，也就是鞍點，需要符合的條件。這些條件通常被稱為 KKT [7] 條件。這些條件在一般情況下是必要條件，而在凸的最佳化問題中則也是充分條件。這些條件由在原始問題中對 $\mathbf{w}$ 微分後讓微分的結果等於零得到。然而這只是條件的一部份，另外一部份則來自於拉格朗日法對於不等式限制的條件，也就是**拉式乘數** (Lagrangian multiplier) 必須不是負數。最後還有一個稱為**互補寬鬆度** (complementary slackness) 的條件需要符合。

$$\begin{aligned}
\partial_{\mathbf{w}}\mathcal{L}_{\mathcal{P}} = 0 \rightarrow \quad & \mathbf{w} - \textstyle\sum_i \alpha_i y_i \mathbf{x_i} = 0 \\
\partial_b \mathcal{L}_{\mathcal{P}} = 0 \rightarrow \quad & \textstyle\sum_i \alpha_i y_i = 0 \\
\textbf{限制-1} \quad & y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 \geq 0 \\
\textbf{拉式乘數條件} \quad & \alpha_i \geq 0 \\
\textbf{互補寬鬆度} \quad & \alpha_i[y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1] = 0
\end{aligned}$$

由 "限制-1" 跟 "互補寬鬆度" 兩組條件來看，意味著當 $y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 > 0$ 時，$\alpha_i = 0$。反之，$y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 = 0$時，$alpha_i \geq 0$。訓練用且 $\alpha_i > 0$ 的樣本決定了支持超平面的位置，因此被稱為支持向量。這些向量會位於支持超平面上而且它們決定了這個問題的解。一般情況下，它們的個數很少，所以被稱為 "**稀疏的**" (sparse) 的解 (大部分的$\alpha_i = 0$)。

我們真正有興趣的是可以用來分類測試用樣本的函數 $f()$,

$$f(x) = \mathbf{w}^{*T}\mathbf{x} - b^* = \sum_i \alpha_i y_i \mathbf{x_i}^T\mathbf{x} - b^*$$

利用 KTT 條件中的 "互補寬鬆度"，我們可以得到一個 $b^*$ 的解：

$$b^* = (\sum_j \alpha_j y_j \mathbf{x_j}^T\mathbf{x_i} - y_i)$$

其中 $i$ 是支持向量且$y_i^2 = 1$。所以，使用任何一組支持向量就可以決定 $b$, 但是為了數值的穩定性，比較好的方法是將它們全部平均起來 (雖然它們理論上都要一樣)。

再提一次最重要的結論是這個函數 $f$ 可以用 $\mathbf{x_i}^T\mathbf{x_j}$ 這樣的內積方式來表示。於是我們可以用核函數 $k(\mathbf{x_i}, \mathbf{x_j})$ 取代這個內積進而應用在高維度的非線性空間中使用。而且，既然 $\alpha$ 一般情況下非常稀疏，我們不需要在給定新的 $\mathbf{x}$ 來預測類別時計算很多核函數值。

[7] Karush-Kuhn-Tucker 的縮寫

## 2.3 不可完全分隔的情況

很明顯地，不是所有的訓練用樣本集合都可以被線性的分隔，因此我們需要去改變之前的推導來允許這些情況的發生。而不難發現我們的問題在於原本的限制可能無法永遠被滿足。因此，加入一個 "寬鬆變數" (slack variable) $\xi_i$ 來放鬆這些限制：

$$\begin{aligned} \mathbf{w}^T\mathbf{x_i} - b &\leq -1 + \xi_i \quad \forall y_i = -1 \\ \mathbf{w}^T\mathbf{x_i} - b &\geq +1 - \xi_i \quad \forall y_i = +1 \\ \xi_i &\geq 0 \qquad \forall i \end{aligned}$$

變數 $\xi_i$ 允許我們去違反原本的限制。我們應該在目標函數中對這些違反限制的情況加予處罰，否則這些限制就顯得毫無意義。處罰函數的形式為 $C(\sum_i \xi_i)^k$ 時，只要 $k$ 為正整數，可以依然使這個問題是凸的最佳化問題。當 $k = 1, 2$ 時，這仍是一個二次規劃的問題。在下面的討論中，我們選擇 $k = 1$。$C$ 的值將控制處罰跟邊際的平衡。

要將一個樣本放在分隔超平面錯誤的一邊需要讓 $\xi > 1$。因此，$\sum_i \xi_i$ 的和可以解釋為這些違反限制的情況有多嚴重，同時也是違法限制的個數的上限。

這個新的原始問題就變成：

$$\begin{aligned} \textbf{最小化} \quad & \tfrac{1}{2}||\mathbf{w}||^2 + C\textstyle\sum_i \xi_i \\ \textbf{並符合} \quad & y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 + \xi_i \geq 0 \ \forall i \\ & \xi \geq 0 \ \forall i \end{aligned}$$

使用拉格朗日法：

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_i \xi_i - \sum_{i=1}^{N} \alpha_i[y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 + \xi_i] - \sum_{i=1}^{N} \mu\xi_i$$

再由此獲得 KKT 條件：

$$\begin{aligned} \partial_{\mathbf{w}}\mathcal{L}_{\mathcal{P}} = 0 \rightarrow \quad & \mathbf{w} - \textstyle\sum_i \alpha_i y_i \mathbf{x_i} = 0 \\ \partial_b\mathcal{L}_{\mathcal{P}} = 0 \rightarrow \quad & \textstyle\sum_i \alpha_i y_i = 0 \\ \partial_\xi\mathcal{L}_{\mathcal{P}} = 0 \rightarrow \quad & C - \alpha_i - \mu_i = 0 \\ \textbf{限制-1} \quad & y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 + \xi \geq 0 \\ \textbf{限制-2} \quad & \xi_i \geq 0 \\ \textbf{拉氏乘數-1} \quad & \alpha_i \geq 0 \\ \textbf{拉氏乘數-2} \quad & \mu_i \geq 0 \\ \textbf{互補寬鬆度-1} \quad & \alpha[y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 + \xi_i] = 0 \\ \textbf{互補寬鬆度-2} \quad & \mu_i\xi_i = 0 \end{aligned}$$

我們可以從以上的條件推論出下列事實。如果我們假設 $\xi_i > 0$ 則 $\mu_i = 0$, 因此 $\alpha_i = C$ 且 $\xi_i = 1 - y_i(\mathbf{w}^T\mathbf{x_i} - b)$。同樣地，當 $\xi_i = 0$ 則 $\mu_i > 0$, 因此 $\alpha_i < C$。除非 $\xi = 0$，我們同樣可以得到：當 $y_i(\mathbf{w}^T\mathbf{x} - b) - 1 = 0$ 時，$\alpha_i > 0$，否則如果 $y_i(\mathbf{w}^T\mathbf{x} - b) - 1 > 0$ 時，$\alpha_i = 0$。總而言之，跟之前一樣，如果樣本不在支持超平面而且在正確的一邊時，$\xi_i = \alpha_i = 0$。如果在支持超平面上時，我們同樣得到 $\xi = 0$，只是這時 $\alpha_i > 0$。最後，對於那些在錯誤的一邊的樣本，$\alpha_i$最大可以到 $C$，而 $\xi$ 會平衡被違反的限制使得 $y_i(\mathbf{w}^T\mathbf{x_i} - b) - 1 + \xi_i = 0$

幾何上，我們可以計算出支持超平面到違反限制的樣本的距離為 $\xi_i/||\mathbf{w}$。

最後，我們需要將這個原始問題轉化成對偶問題使得我們可以有效率的解它和將它核化。再一次地，我們使用 KKT 條件將 $\mathbf{w}$，$b$ 跟 $\xi$ 代換掉：

$$\begin{array}{ll} \textbf{最大化} & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T \mathbf{x_j} \\ \textbf{並符合} & \sum_i \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \ \forall i \end{array}$$

令人驚訝地，除了 $\alpha_i$ 現在有了上界外，這幾乎跟原本的式子是一樣地。同時，注意到這式子依然只跟 $\mathbf{x_i}^T\mathbf{x_j}$ 這樣的內積有關，可以被核化。