9. Let the two sets of vectors be $\mathcal{S}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{S}_2 = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$. We assume S_1 and S_2 are linearly seperable, that is, there exists a linear discriminant function $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$ such that

$$\begin{aligned} g(\mathbf{x}) &> 0 \text{ implies } \mathbf{x} \in \mathcal{S}_1 \quad \text{and} \\ g(\mathbf{x}) &< 0 \text{ implies } \mathbf{x} \in \mathcal{S}_2. \end{aligned}$$

Consider a point $\mathbf{x}$ in the convex hull of $\mathcal{S}_1$, or

$$\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i,$$

where the α_i's are non-negative and sum to 1. The discriminant function evaluated at $\mathbf{x}$ is

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^t\mathbf{x} + w_0 \\ &= \mathbf{w}^t\left(\sum_{i=1}^{n} \alpha_i \mathbf{x}_i\right) + w_0 \\ &= \sum_{i=1}^{n} \alpha_i(\mathbf{w}^t\mathbf{x}_i + w_0) \\ &> 0, \end{aligned}$$

where we used the fact that $\mathbf{w}^t\mathbf{x}_i + w_0 > 0$ for $1 \le i \le n$ and $\sum_{i=1}^{n} \alpha_i = 1$.

Now let us assume that our point $\mathbf{x}$ is *also* in the convex hull of S_2, or

$$\mathbf{x} = \sum_{j=1}^{m} \beta_j \mathbf{y}_j,$$

where the β_j's are non-negative and sum to 1. We follow the approach immediately above and find

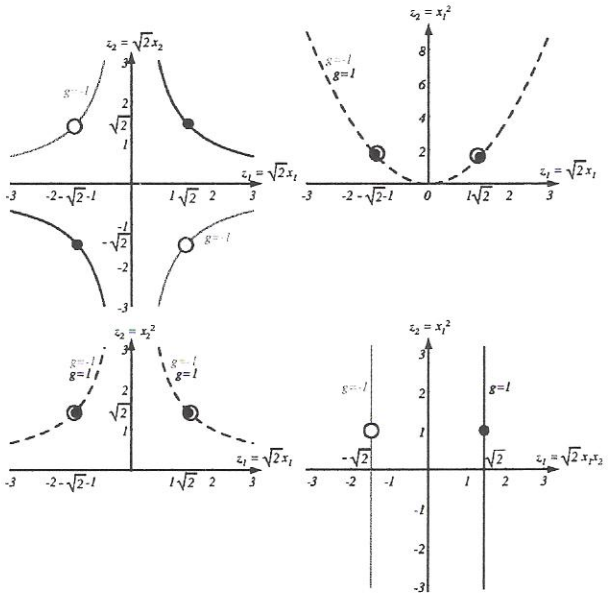
$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^t\mathbf{x} + w_0 \\ &= \mathbf{w}^t\left(\sum_{j=1}^{m} \beta_j \mathbf{y}_j\right) + w_0 \\ &= \sum_{j=1}^{m} \beta_j \underbrace{(\mathbf{w}^t\mathbf{y}_j + w_0)}_{g(\mathbf{y}_j)<0} \\ &< 0, \end{aligned}$$

where the last step comes from the realization that $g(\mathbf{y}_j) = \mathbf{w}^t\mathbf{y}_j + w_0 < 0$ for each $\mathbf{y}_i$, since they are each in $\mathcal{S}_2$. Thus, we have a contradiction: $g(\mathbf{x}) > 0$ and $g(\mathbf{x}) < 0$, and hence clearly the intersection is empty. In short, either two sets of vectors are either linearly separable or their convex hulls intersects.

10. Consider a piecewise linear machine.

(a) The discriminiant functions have the form

$$g_i(\mathbf{x}) = \max_{j=1,\ldots,n_i} g_{ij}(\mathbf{x}),$$



as shown in the figure.
The margins are not the same, simply because the real margin is the distance of the support vectors to the optimal hyperplane in $\mathbf{R}^6$ space, and their projection to lower dimensional subspaces does not necessarily preserve the margin.
31. The Support Vector Machine algorithm can be written:

Algorithm 0 (SVM)

```
1  begin initialize a; worst1 ← ∞; worst2 ← ∞; b ← ∞
2    i ← 0
3    do i ← i + 1
4       if z_i = −1 and a^t y_i z_i < worst1, then worst1 ← a^t y_i z_i; kworst1 ← k
5       if z_i = 1 and a^t y_i z_i < worst2, then worst2 ← a^t y_i z_i; kworst2 ← k
6    until i = n
7    a ← a + y_kworst2 − y_kworst1
8    a_0 ← a^t(y_kworst2 + y_kworst1)/2
9    oldb ← b; b ← a^t y_kworst1 / || a ||
10 until | b − oldb | < ε
11 return a_0, a
12 end
```

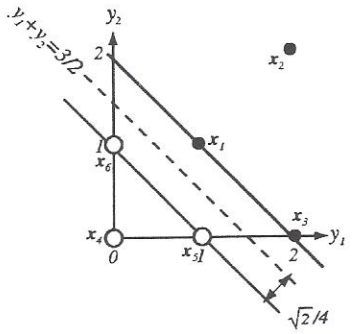
Note that the algorithm picks the worst classified patterns from each class and adjusts $\mathbf{a}$ such that the hyperplane moves toward the center of the worst patterns and rotates so that the angle between the hyperplane and the vector connecting the worst points increases. Once the hyperplane separates the classes, all the updates will involve support vectors, since the **if** statements can only pick the vectors with the smallest $|\mathbf{a}^t\mathbf{y}_i|$.
32. Consider Support Vector Machines for classification.

(a) We are given the following six points in two categories:

$$\omega_1 \quad : \quad \mathbf{x}_1 = \begin{pmatrix}1\\1\end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix}2\\2\end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix}2\\0\end{pmatrix}$$

$$\omega_2 \quad : \quad \mathbf{x}_4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \mathbf{x}_6 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

with $z_1 = z_2 = z_3 = -1$ and $z_4 = z_5 = z_6 = +1$.



The optimal hyperplane is $y_1 + y_2 = 3/2$, or $(3/2 \ -1 \ -1)^t(1 \ y_1 \ y_2) = 0$. To ensure $z_k \mathbf{a}^t \mathbf{y} \geq 1$, we have to scale $(3/2 \ -1 \ -1)^t$ by 2, and thus the weight vector is $(3 \ -2 \ -2)^t$. The optimal margin is the shortest distance from the patterns to the optimal hyperplane, which is $\sqrt{2}/4$, as can be seen in the figure.

(b) Support vectors are the samples on the margin, that is, the ones with the shortest distance to the separating hyperplane. In this case, the support vectors are $\{\mathbf{x}_1, \mathbf{x}_3, \mathbf{x}_5, \mathbf{x}_6\} = \{(1,1)^t, (2,0)^t, (1,0)^t, (0,1)^t\}$.

(c) We seek to maximize the criterion given in Eq. 109 in the text,

$$L(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \alpha_k - \frac{1}{2} \sum_{k,j}^{n} \alpha_k \alpha_j z_k z_j \mathbf{y}_j^t \mathbf{y}_k$$

subject to the constraints

$$\sum_{k=1}^{n} z_k \alpha_k = 0$$

for $\alpha_k \geq 0$. Using the constraint, we can substitute $\alpha_6 = \alpha_1 + \alpha_2 + \alpha_3 - \alpha_4 - \alpha_5$ in the expression for $L(\boldsymbol{\alpha})$. Then we can get a system of linear equations by setting the partial derivatives, $\partial L / \partial \alpha_i$ to zero. This yields:

$$\begin{bmatrix} -1 & -2 & -2 & 0 & 1 \\ -2 & -5 & 2 & -1 & 1 \\ -2 & 2 & -5 & 1 & 1 \\ 0 & -1 & 1 & -1 & -1 \\ 1 & 1 & 3 & -1 & -2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} -2 \\ -2 \\ -2 \\ 0 \\ 0 \end{bmatrix}.$$

Unfortunately, this is an inconsistent set of equations. Therefore, the maxima must be achieved on the boundary (where some α_i vanish). We try each $\alpha_i = 0$ and solve $\partial L / \partial \alpha_i = 0$:

$$\frac{\partial L(0, \alpha_2, \alpha_3, \alpha_4, \alpha_5)}{\partial \alpha_i} = 0$$

implies $\boldsymbol{\alpha} = 1/5(0, -2\ -2\ 8\ -8\ -4)^t$, which violates the constraint $\alpha_i \geq 0$. Next, both of the following vanishing derivatives,

$$\frac{\partial L(\alpha_1, 0, \alpha_3, \alpha_4, \alpha_5)}{\partial \alpha_i} = \frac{\partial L(\alpha_1, \alpha_2, 0, \alpha_4, \alpha_5)}{\partial \alpha_i} = 0$$

lead to inconsistent equations. Then the derivative

$$\frac{\partial L(\alpha_1, \alpha_2, \alpha_3, 0, \alpha_5)}{\partial \alpha_i} = 0$$

implies $\boldsymbol{\alpha} = 1/5(16\ 0\ 4\ 0\ 14\ 6)^t$, which does not violate the constraint $\alpha_i \geq 0$. In this case the criterion function is $L(\boldsymbol{\alpha}) = 4$. Finally, we have

$$\frac{\partial L(\alpha_1, \alpha_2, \alpha_3, \alpha_4, 0)}{\partial \alpha_i} = 0$$

which implies $\boldsymbol{\alpha} = 1/5(2\ 2\ 2\ 0\ 0\ 6)^t$, and the constraint $\alpha_i \geq 0$ is obeyed. In this case the criterion function is $L(\boldsymbol{\alpha}) = 1.2$.

Thus $\boldsymbol{\alpha} = 1/5(16\ 0\ 4\ 0\ 14\ 6)^t$ is where the criterion function L reaches its maximum within the constraints. Now we seek the weight vector $\mathbf{a}$. We seek to minimize $L(\mathbf{a}, \boldsymbol{\alpha})$ of Eq. 108 in the text,

$$L(\mathbf{a}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{a}\|^2 - \sum_{k=1}^{n} \alpha_k [z_k \mathbf{a}^t \mathbf{y}_k - 1],$$

with respect to $\mathbf{a}$. We take the derivative of the criterion function,

$$\frac{\partial L}{\partial \mathbf{a}} = \mathbf{a} - \sum_{k=1}^{n} \alpha_k z_k \mathbf{y}_k = \mathbf{0},$$

which for the α_k found above has solution

$$\begin{aligned} \mathbf{a} &= -(16/5)\mathbf{y}_1 - 0\mathbf{y}_2 - 4/5\mathbf{y}_3 + 0\mathbf{y}_4 + 14/5\mathbf{y}_5 + 6/5\mathbf{y}_6 \\ &= \begin{pmatrix} 0 \\ -2 \\ -2 \end{pmatrix}. \end{aligned}$$

Note that $\partial L/\partial \mathbf{a} = 0$ here is not sufficient to allow us to find the bias a_0 directly since the $\|\mathbf{a}\|^2$ term does not include the augmented vector $\mathbf{a}$ and $\sum_k \alpha_k z_k = 0$. We determine a_0, then, by using one of the support vectors, for instance $\mathbf{y}_1 = (1\ 1\ 1)^t$. Since $\mathbf{y}_1$ is a support vector, $\mathbf{a}^t \mathbf{y}_1 z_1 = 1$ holds, and thus

$$-\begin{pmatrix} 0 \\ -2 \\ -2 \end{pmatrix} (1\ 1\ 1) = -a_0 + 4 = 1.$$

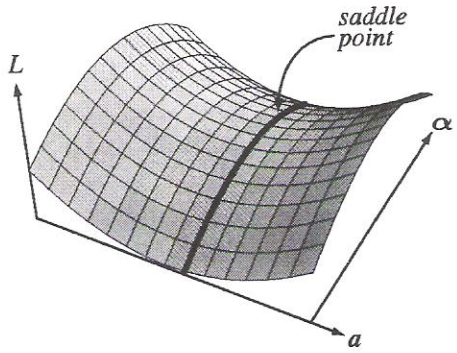
This, then, means $a_0 = 3$, and the full weight vector is $\mathbf{a} = (3\ -2\ -2)^t$.

33. Consider the Kuhn-Tucker theorem and the conversion of a constrained optimization problem for support vector machines to an unconstrained one.

(a) Here the relevant functional is given by Eq. 108 in the text, that is,

$$L(\mathbf{a}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{a}\|^2 - \sum_{k=1}^{n} \alpha_k [z_k \mathbf{a}^t \mathbf{y}_k - 1].$$

We seek to maximize L with respect to $\boldsymbol{\alpha}$ to guarantee that all the patterns are correctly classified, that is $z_k \mathbf{a}^t \mathbf{y}_k \geq 1$, and we want to minimize L with respect to the (un-augmented) $\mathbf{a}$. This will give us the optimal hyperplane. This solution corresponds to a saddle point in $\boldsymbol{\alpha}$–$\mathbf{a}$ space, as shown in the figure.



(b) We write the augmented vector $\mathbf{a}$ as $(a_0 \ \mathbf{a}_r)^t$, where a_0 is the augmented bias. Then we have

$$L(\mathbf{a}_r, \boldsymbol{\alpha}, a_0) = \frac{1}{2}\|\mathbf{a}_r\|^2 - \sum_{k=1}^{n} \alpha_k [z_k \mathbf{a}_r^t \mathbf{y}_k + z_k a_0 - 1].$$

At the saddle point, $\partial L/\partial a_0 = 0$ and $\partial L/\partial \mathbf{a}_r = \mathbf{0}$. The first of these derivative vanishing implies

$$\sum_{k=1}^{n} \alpha_k^* z_k = 0.$$

(c) The second derivative vanishing implies

$$\frac{\partial L}{\partial \mathbf{a}_r} = \mathbf{a}_r - \sum_{k=1}^{n} \alpha_k^* z_k \mathbf{y}_k$$

and thus

$$\mathbf{a}_r = \sum_{k=1}^{n} \alpha_k^* z_k \mathbf{y}_k.$$

Since $\sum_{k=1}^{n} \alpha_k z_k = 0$, we can thus write the solution in augmented form as

$$\mathbf{a} = \sum_{k=1}^{n} \alpha_k^* z_k \mathbf{y}_k.$$

(d) If $\alpha_k^*(z_k\mathbf{a}^{*t}\mathbf{y}_k - 1) = 0$ and if $z_k\mathbf{a}^{*t}\mathbf{y}_k \neq 0$, then $\boldsymbol{\alpha}_k^*$ must be zero. Respectively, call the predicates as h, NOTp and q, then the above states h $AND\ NOT$ p $\rightarrow$ q, which is equivalent to h $AND\ NOT$ q $\rightarrow$ p. In other words, given the expression

$$\alpha_k^*(z_k\mathbf{a}^{*t}\mathbf{y} - 1) = 0,$$

then $\boldsymbol{\alpha}^*$ is non-zero if and only if $z_k\mathbf{a}^{*t}\mathbf{y}_k = 1$.

(e) Here we have

$$\begin{aligned}\bar{L} &= \frac{1}{2}\left\|\sum_{k=1}^{n}\alpha_k z_k\mathbf{y}_k\right\|^2 - \sum_{k=1}^{n}\alpha_k\left[z_k\left(\sum_{l=1}^{n}\alpha_l z_l\mathbf{y}_l\right)\mathbf{y}_k - 1\right] \\ &= \frac{1}{2}\left(\sum_{k=1}^{n}\alpha_k z_k\mathbf{y}_k\right)^t\left(\sum_{k=1}^{n}\alpha_k z_k\mathbf{y}_k\right) - \sum_{kl}\alpha_k\alpha_l z_k z_l\mathbf{y}_k^t\mathbf{y}_l + \sum_{k=1}^{n}\alpha_k.\end{aligned}$$

Thus we have

$$\bar{L} = \sum_{k=1}^{n}\alpha_k - \frac{1}{2}\sum_{kl}\alpha_k\alpha_l z_k z_l\mathbf{y}_k^t\mathbf{y}_l.$$

(f) See part (e).

34. We repeat Example 2 in the text but with the following four points:

$$\mathbf{y}_1 = (1\ \sqrt{2}\ 5\sqrt{2}\ 5\sqrt{2}\ 1\ 25)^t,\quad \mathbf{y}_2 = (1\ -2\sqrt{2}\ -4\sqrt{2}\ 8\sqrt{2}\ 4\ 16)^t, z_1 = z_2 = -1$$
$$\mathbf{y}_3 = (1\ \sqrt{2}\ 3\sqrt{2}\ 6\sqrt{2}\ 4\ 9)^t,\quad \mathbf{y}_4 = (1\ -2\sqrt{2}\ 5\sqrt{2}\ -5\sqrt{2}\ 1\ 25)^t, z_3 = z_4 = +1$$

We seek the optimal hyperplane, and thus want to maximize the functional given by Eq. 109 in the text:

$$L(\boldsymbol{\alpha}) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\sum_{kl}^{4}\alpha_l\alpha_k z_k z_l\mathbf{y}_k^t\mathbf{y}_l,$$

with constraints $\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4$ and $\alpha_1 \geq 0$. We substitute $\alpha_4 = \alpha_1 + \alpha_2 - \alpha_3$ into $L(\boldsymbol{\alpha})$ and take the partial derivatives with respect to α_1, α_2 and α_3 and set the derivatives to zero:

$$\begin{aligned}\frac{\partial L}{\partial\alpha_1} &= 2 - 208\alpha_1 - 256\alpha_2 + 232\alpha_3 = 0 \\ \frac{\partial L}{\partial\alpha_2} &= 2 - 256\alpha_1 - 592\alpha_2 + 496\alpha_3 = 0 \\ \frac{\partial L}{\partial\alpha_3} &= 232\alpha_1 + 496\alpha_2 - 533\alpha_3 = 0.\end{aligned}$$

The solution to these equations — $\alpha_1 = 0.0154$, $\alpha_2 = 0.0067$, $\alpha_3 = 0.0126$ — indeed satisfy the constraint $\alpha_i \geq 0$, as required.

Now we compute $\mathbf{a}$ using Eq. 108 in the text:

$$\frac{\partial L}{\partial\mathbf{a}} = \mathbf{a} - \sum_{k=1}^{4}\alpha_k z_k\mathbf{y}_k = 0,$$