2 Discriminant Functions

1. Consider a dichotomizer defined using the following linear discriminant function $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$ where $\mathbf{w} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $w_0 = -5$. Plot the decision boundary, and determine the class of the following feature vectors: $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 2 \\ 2 \end{pmatrix}$, and $\begin{pmatrix} 3 \\ 3 \end{pmatrix}$.

At decision boundary $g(\mathbf{x}) = 0$, therefore decision boundary is defined by

$$\mathbf{w}^t\mathbf{x} + w_0 = 0$$

$$(2 \;\; 1)\mathbf{x} - 5 = 0$$

$\mathbf{w}$ *has two elements, so we know we are working in a 2D feature space.*

We also know that we are dealing with a linear discriminant function, so the boundary is a straight line. We can plot a straight line if we know two points on that line.

Let's find where this line intercepts the axes.

$$(2 \;\; 1)\begin{pmatrix} x_1 \\ 0 \end{pmatrix} - 5 = 0$$

$$2x_1 - 5 = 0$$

$$x_1 = 2.5$$

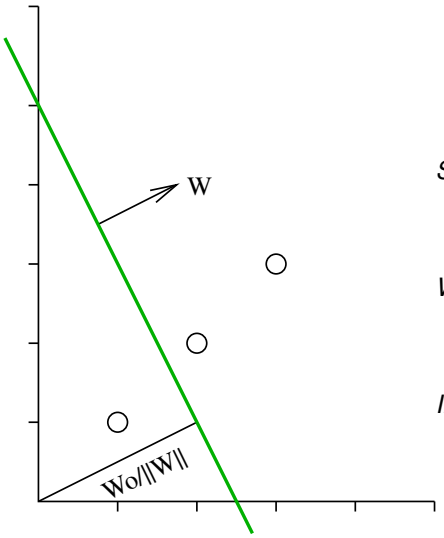
i.e., intercept at (2.5,0)

$$(2 \;\; 1)\begin{pmatrix} 0 \\ x_2 \end{pmatrix} - 5 = 0$$

$$x_2 - 5 = 0$$

$$x_2 = 5$$

i.e., intercept at (0,5)



So, hyperplane has equation:

$$x_2 = mx_1 + c = -2x_1 + 5$$

We could have got this by simply re-arranging

$$(2 \;\; 1)\mathbf{x} - 5 = 0$$

In general,

$$x_2 = mx_1 + c = \frac{-w_1}{w_2}x_1 + \frac{-w_0}{w_2}$$

To classify a point $\mathbf{x}$ *we calculate* $g(\mathbf{x})$*,* $\mathbf{x}$ *is in class 1 if* $g(\mathbf{x}) > 0$*.*

$(2\ \ 1)\begin{pmatrix} 1 \\ 1 \end{pmatrix} - 5 = 2 + 1 - 5 = -2$ *Therefore* $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ *is in class 2.*

$(2\ \ 1)\begin{pmatrix} 2 \\ 2 \end{pmatrix} - 5 = 4 + 2 - 5 = 1$ *Therefore* $\begin{pmatrix} 2 \\ 2 \end{pmatrix}$ *is in class 1.*

$(2\ \ 1)\begin{pmatrix} 3 \\ 3 \end{pmatrix} - 5 = 6 + 3 - 5 = 4$ *Therefore* $\begin{pmatrix} 3 \\ 3 \end{pmatrix}$ *is in class 1.*

Note, the vector normal to the hyperplane points towards class 1. The value of $g(\mathbf{x})$ *provides a measure of how far* $\mathbf{x}$ *is from the decision boundary. The actual distance is given by* $\frac{|g(\mathbf{x})|}{\|\mathbf{w}\|}$*.*

2. In augmented feature space, a dichotomizer is defined using the following linear discriminant function $g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$ **where** $\mathbf{a}^t = (-5, 2, 1)$ **and** $\mathbf{y} = \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix}$**. Determine the class of the following feature vectors:** $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$**,** $\begin{pmatrix} 2 \\ 2 \end{pmatrix}$**, and** $\begin{pmatrix} 3 \\ 3 \end{pmatrix}$**.**

To classify a point x we calculate $g(\mathbf{x})$*,* $\mathbf{x}$ *is in class 1 if* $g(\mathbf{x}) > 0$*.*

$(-5\ \ 2\ \ 1)\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = -5 + 2 + 1 = -2$ *Therefore* $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ *is in class 2.*

$(-5\ \ 2\ \ 1)\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} = -5 + 4 + 2 = 1$ *Therefore* $\begin{pmatrix} 2 \\ 2 \end{pmatrix}$ *is in class 1.*

$(-5\ \ 2\ \ 1)\begin{pmatrix} 1 \\ 3 \\ 3 \end{pmatrix} = -5 + 6 + 3 = 4$ *Therefore* $\begin{pmatrix} 3 \\ 3 \end{pmatrix}$ *is in class 1.*

Note, same as previous question, just using augmented vectors.

3. Consider a 3-dimensional feature space and quadratic discriminant function, $g(\mathbf{x})$**, where:**

$$g(\mathbf{x}) = x_1^2 - x_3^2 + 2x_2x_3 + 4x_1x_2 + 3x_1 - 2x_2 + 2$$

This discriminant function defines two classes, such that $g(\mathbf{x}) > 0$ **if** $\mathbf{x} \in \omega_1$ **and** $g(\mathbf{x}) \leq 0$ **if** $\mathbf{x} \in \omega_2$**. Determine the class of each of the following pattern vectors:** $(1\ \ 1\ \ 1)^t$**,** $(-1\ \ 0\ \ 3)^t$**, and** $(-1\ \ 0\ \ 0)^t$**.**

$g(\mathbf{x} = (1\ \ 1\ \ 1)^t) = 1^2 - 1^2 + 2 + 4 + 3 - 2 + 2 = 9$
hence $(1\ \ 1\ \ 1)^t$ *is in class 1.*

$g(\mathbf{x} = (-1 \;\; 0 \;\; 3)^t) = (-1)^2 - 3^2 + 0 + 0 - 3 - 0 + 2 = -9$
hence $(-1 \;\; 0 \;\; 3)^t$ *is in class 2.*

$g(\mathbf{x} = (-1 \;\; 0 \;\; 0)^t) = (-1)^2 - 0 + 0 + 0 - 3 - 0 + 2 = 0$
hence $(-1 \;\; 0 \;\; 0)^t$ *is in class 2.*

4. Consider a dichotomizer defined in a 2-dimensional feature space using a quadratic discriminant function, $g(\mathbf{x})$, where:

$$g(\mathbf{x}) = \mathbf{x}^t\mathbf{A}\mathbf{x} + \mathbf{x}^t\mathbf{b} + c$$

Classify the following feature vectors: $(0, -1)^t$, and $(1, 1)^t$, when:

i) $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, **and** $c = -3$.

ii) $\mathbf{A} = \begin{pmatrix} -2 & 5 \\ 5 & -8 \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, **and** $c = -3$.

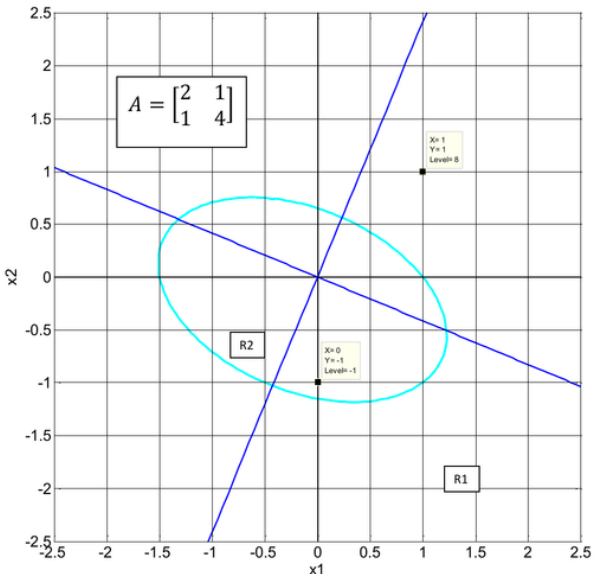
i)

$$\begin{aligned} g(\mathbf{x}) &= (x_1, x_2)\begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1, x_2)\begin{pmatrix} 1 \\ 2 \end{pmatrix} - 3 \\ &= (2x_1 + x_2, x_1 + 4x_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + x_1 + 2x_2 - 3 \\ &= 2x_1^2 + x_1x_2 + x_1x_2 + 4x_2^2 + x_1 + 2x_2 - 3 \\ &= 2x_1^2 + 4x_2^2 + 2x_1x_2 + x_1 + 2x_2 - 3 \end{aligned}$$

When $\mathbf{x} = (0, -1)^t$, $g(\mathbf{x}) = 0 + 4(-1)^2 + 0 + 0 + 2(-1) - 3 = -1$
$g(\mathbf{x}) \leq 0$ *so* $\mathbf{x}$ *is in class 2.*

When $\mathbf{x} = (1, 1)^t$, $g(\mathbf{x}) = 2 + 4 + 2 + 1 + 2 - 3 = 8$
$g(\mathbf{x}) > 0$ *so* $\mathbf{x}$ *is in class 1.*

The decision surface looks like this:



ii)

$$g(\mathbf{x}) = (x_1, x_2)\begin{pmatrix} -2 & 5 \\ 5 & -8 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1, x_2)\begin{pmatrix} 1 \\ 2 \end{pmatrix} - 3$$

$$= (-2x_1 + 5x_2, 5x_1 - 8x_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + x_1 + 2x_2 - 3$$

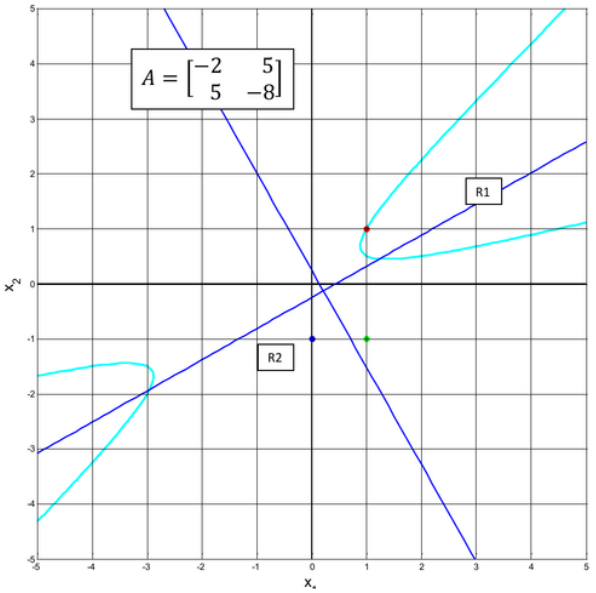
$$= -2x_1^2 + 5x_1x_2 + 5x_1x_2 - 8x_2^2 + x_1 + 2x_2 - 3$$

$$= -2x_1^2 - 8x_2^2 + 10x_1x_2 + x_1 + 2x_2 - 3$$

When $\mathbf{x} = (0, -1)^t$, $g(\mathbf{x}) = 0 - 8(-1)^2 + 0 + 0 + 2(-1) - 3 = -13$
$g(\mathbf{x}) \leq 0$ *so* $\mathbf{x}$ *is in class 2.*

When $\mathbf{x} = (1, 1)^t$, $g(\mathbf{x}) = -2 - 8 + 10 + 1 + 2 - 3 = 0$
$g(\mathbf{x}) = 0$ *so* $\mathbf{x}$ *is in class 2.*

The decision surface looks like this:



5. In augmented feature space, a dichotomizer is defined using the following linear discriminant function $g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$ where $\mathbf{a}^t = (-3, 1, 2, 2, 2, 4)$ and $\mathbf{y}^t = (1, \mathbf{x}^t)$. Determine the class of the following feature vectors, $\mathbf{x}$: $\begin{pmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$.

To classify a point $\mathbf{x}$ we calculate $g(\mathbf{x})$, $\mathbf{x}$ is in class 1 if $g(\mathbf{x}) > 0$.

$$(-3 \;\; 1 \;\; 2 \;\; 2 \;\; 2 \;\; 4) \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = -3 + 0 - 2 + 0 + 0 + 4 = -1$$

Therefore class 2.

$$(-3 \;\; 1 \;\; 2 \;\; 2 \;\; 2 \;\; 4) \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = -3 + 1 + 2 + 2 + 2 + 4 = 8$$

Therefore class 1.

Note, same as part (i) of previous question, just using generalised linear discriminant function where $\mathbf{y}^t = (1 \;\; x_1 \;\; x_2 \;\; x_1x_2 \;\; x_1^2 \;\; x_2^2)$.

6. A Linear Discriminant Function is used to define a Dichotomizer, such that $\mathbf{x}$ is assigned to class 1 if $g(\mathbf{x}) > 0$, and $\mathbf{x}$ is assigned to class 2 otherwise. Use the Batch Perceptron Learning Algorithm (with augmented notation and sample normalisation), to find appropriate parameters for the linear discriminant function, when the data set is as shown.

$\mathbf{x}$	$class$
$(1, 5)^t$	1
$(2, 5)^t$	1
$(4, 1)^t$	2
$(5, 1)^t$	2

Assume an initial values of $\mathbf{a} = (w_0, \mathbf{w}^t)^t = (-25, 6, 3)^t$, and use a learning rate of 1.

Using Augmented notation and sample normalisation, dataset is:

$\mathbf{x}$	$\mathbf{y}$
$(1, 5)^t$	$(1, 1, 5)^t$
$(2, 5)^t$	$(1, 2, 5)^t$
$(4, 1)^t$	$(-1, -4, -1)^t$
$(5, 1)^t$	$(-1, -5, -1)^t$

For the Batch Perceptron Learning Algorithm, weights are updated such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta \sum_{y \in \chi} \mathbf{y}$. *Here,* $\eta = 1$.

Epoch 1: initial $\mathbf{a} = (-25, 6, 3)^t$

$\mathbf{y}$	$g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$	*misclassified (i.e., $g(\mathbf{x}) \leq 0$)?*
$(1, 1, 5)^t$	$(-25 \times 1) + (6 \times 1) + (3 \times 5) = -4$	*yes*
$(1, 2, 5)^t$	$(-25 \times 1) + (6 \times 2) + (3 \times 5) = 2$	*no*
$(-1, -4, -1)^t$	$(-25 \times -1) + (6 \times -4) + (3 \times -1) = -2$	*yes*
$(-1, -5, -1)^t$	$(-25 \times -1) + (6 \times -5) + (3 \times -1) = -8$	*yes*

$\mathbf{a} \leftarrow (-25, 6, 3)^t + (1, 1, 5)^t + (-1, -4, -1)^t + (-1, -5, -1)^t = (-26, -2, 6)^t$

Epoch 2: $\mathbf{a} = (-26, -2, 6)^t$

$\mathbf{y}$	$g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$	*misclassified (i.e., $g(\mathbf{x}) \leq 0$)?*
$(1, 1, 5)^t$	$(-26 \times 1) + (-2 \times 1) + (6 \times 5) = 2$	*no*
$(1, 2, 5)^t$	$(-26 \times 1) + (-2 \times 2) + (6 \times 5) = 0$	*yes*
$(-1, -4, -1)^t$	$(-26 \times -1) + (-2 \times -4) + (6 \times -1) = 28$	*no*
$(-1, -5, -1)^t$	$(-26 \times -1) + (-2 \times -5) + (6 \times -1) = 30$	*no*

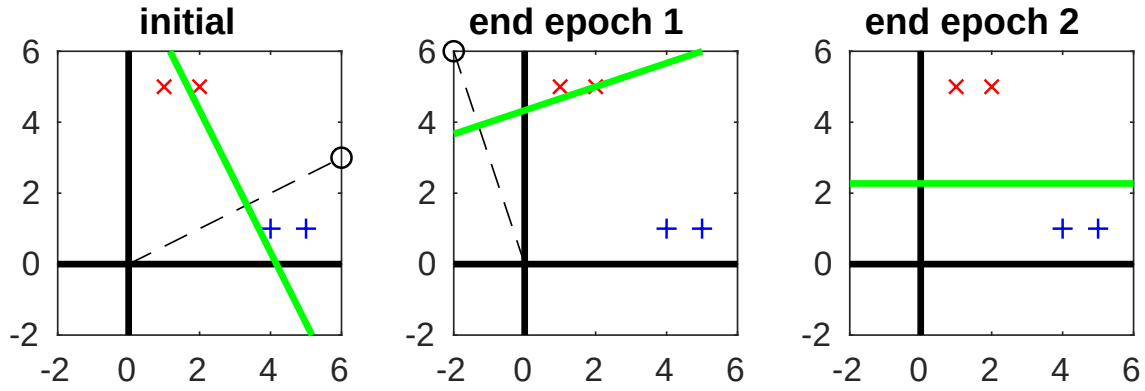
$\mathbf{a} \leftarrow (-26, -2, 6)^t + (1, 2, 5)^t = (-25, 0, 11)^t$

Epoch 3: $\mathbf{a} = (-25, 0, 11)^t$

$\mathbf{y}$	$g(\mathbf{x}) = \mathbf{a}^t\mathbf{t}$	*misclassified (i.e., $g(\mathbf{x}) \leq 0$)?*
$(1, 1, 5)^t$	$(-25 \times 1) + (0 \times 1) + (11 \times 5) = 30$	*no*
$(1, 2, 5)^t$	$(-25 \times 1) + (0 \times 2) + (11 \times 5) = 30$	*no*
$(-1, -4, -1)^t$	$(-25 \times -1) + (0 \times -4) + (11 \times -1) = 14$	*no*
$(-1, -5, -1)^t$	$(-25 \times -1) + (0 \times -5) + (11 \times -1) = 14$	*no*

Learning has converged, so required parameters are $\mathbf{a} = (-25, 0, 11)^t$.

In feature space, the decision surface found at each epoch looks like this:



7. Repeat the previous question using the Sequential Perceptron Learning Algorithm (with augmented notation and sample normalisation).

Using Augmented notation and sample normalisation, dataset is:

$\mathbf{x}$	$\mathbf{y}$
$(1,5)^t$	$(1,1,5)^t$
$(2,5)^t$	$(1,2,5)^t$
$(4,1)^t$	$(-1,-4,-1)^t$
$(5,1)^t$	$(-1,-5,-1)^t$

For the Sequential Perceptron Learning Algorithm, weights are updated such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta \mathbf{y}_k$*, where* $\mathbf{y}_k$ *is a misclassified exemplar. Here,* $\eta = 1$.

Epoch 1: initial $\mathbf{a} = (-25, 6, 3)^t$

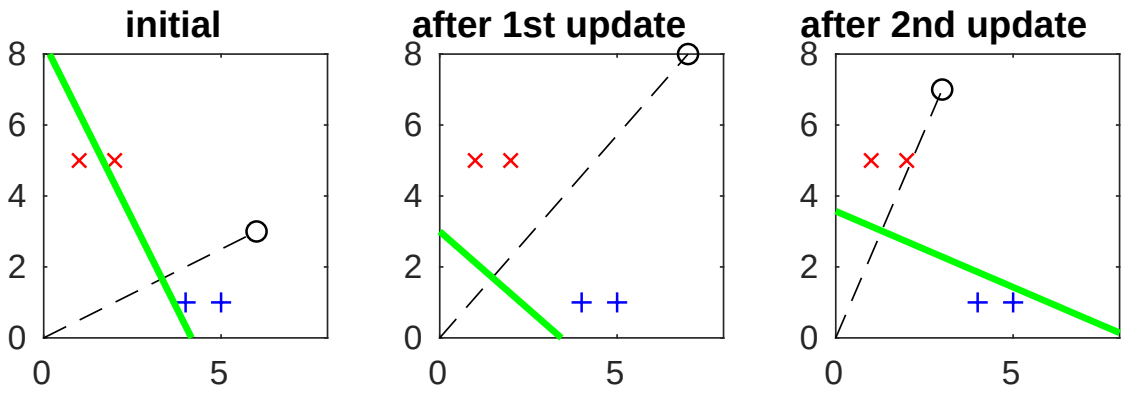
$\mathbf{y}^t$	$g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$	*updated* $\mathbf{a}^t$
$(1,1,5)$	$(-25 \times 1) + (6 \times 1) + (3 \times 5) = -4$	$(-25,6,3) + (1,1,5) = (-24,7,8)$
$(1,2,5)$	$(-24 \times 1) + (7 \times 2) + (8 \times 5) = 30$	$(-24,7,8)$
$(-1,-4,-1)$	$(-24 \times -1) + (7 \times -4) + (8 \times -1) = -12$	$(-24,7,8) + (-1,-4,-1) = (-25,3,7)$
$(-1,-5,-1)$	$(-25 \times -1) + (3 \times -5) + (7 \times -1) = 3$	$(-25,3,7)$

Epoch 2: $\mathbf{a} = (-25, 3, 7)^t$

$\mathbf{y}^t$	$g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$	*updated* $\mathbf{a}^t$
$(1,1,5)$	$(-25 \times 1) + (3 \times 1) + (7 \times 5) = 13$	$(-25,3,7)$
$(1,2,5)$	$(-25 \times 1) + (3 \times 2) + (7 \times 5) = 16$	$(-25,3,7)$
$(-1,-4,-1)$	$(-25 \times -1) + (3 \times -4) + (7 \times -1) = 6$	$(-25,3,7)$
$(-1,-5,-1)$	$(-25 \times -1) + (3 \times -5) + (7 \times -1) = 3$	$(-25,3,7)$

Learning has converged, so required parameters are $\mathbf{a} = (-25, 3, 7)^t$.

In feature space, the decision surface found at each update looks like this:



8. Write pseudo-code for the sequential Perceptron Learning Algorithm

Augment and apply sample normalisation to the feature vectors.
Initialise $\mathbf{a}$ *and set the learning rate.*
For each sample, $\mathbf{y}_k$*, in the data-set:*
    Calculate $g(\mathbf{x}) = \mathbf{a}^T\mathbf{y}$
    If $\mathbf{y}_k$ *is misclassified (i.e., if* $g(\mathbf{x}) \leq 0$*)*
        Update solution such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta\mathbf{y}_k$
Repeat until $\mathbf{a}$ *unchanged by all samples (i.e. all samples correctly classified).*

OR
Augment the feature vectors.
Set $\omega_k = 1$ *for samples in class 1, and* $\omega_k = -1$ *for samples in class 2.*
Initialise $\mathbf{a}$ *and set the learning rate.*
For each sample, $\mathbf{y}_k$*, in the data-set:*
    Calculate $g(\mathbf{x}) = \mathbf{a}^T\mathbf{y}$
    If $\mathbf{y}_k$ *is misclassified (i.e., if* $sign(g(\mathbf{x})) \neq \omega_k$*)*
        Update solution such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta\omega_k\mathbf{y}_k$
Repeat until $\mathbf{a}$ *unchanged by all samples (i.e. all samples correctly classified).*

9. Consider the following linearly separable data set.

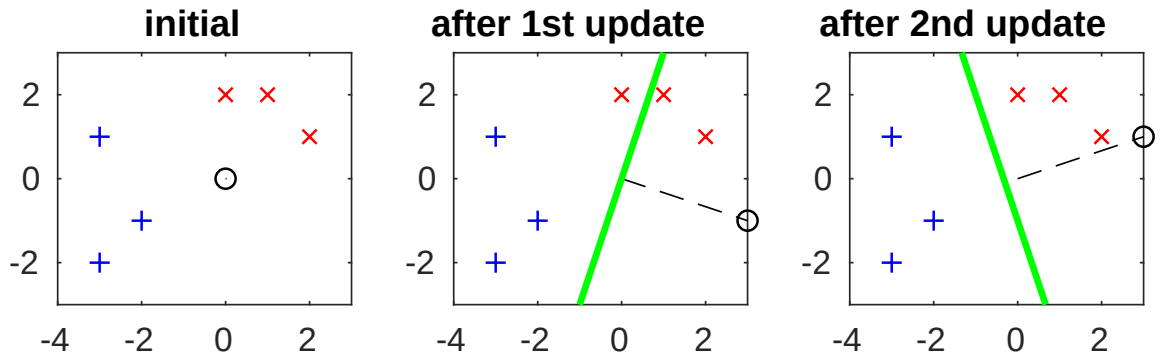
$\mathbf{x}$	$class(\omega)$
$(0,2)^t$	1
$(1,2)^t$	1
$(2,1)^t$	1
$(-3,1)^t$	-1
$(-2,-1)^t$	-1
$(-3,-2)^t$	-1

Apply the Sequential Perceptron Learning Algorithm to determine the parameters of a linear discriminant function that will correctly classify this data. Assume an initial values of $\mathbf{a} = (w_0, \mathbf{w}^t)^t = (1,0,0)^t$**, and use a learning rate of 1.**

For the Sequential Perceptron Learning Algorithm, weights are updated such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta\omega_k\mathbf{y}_k$*, where* $\mathbf{y}_k$ *is a misclassified exemplar, and* ω_k *is the class label associated with* $\mathbf{y}_k$*. Here,* $\eta = 1$*.*

iteration	$\mathbf{a}^t_{old}$	$\mathbf{y}^t$	$g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$	ω_k	$\mathbf{a}^t_{new} = \mathbf{a}^t_{old} + \omega_k\mathbf{y}^t$ *if misclassified*
1	$[1,0,0]$	$[1,0,2]$	1	1	$[1,0,0]$
2	$[1,0,0]$	$[1,1,2]$	1	1	$[1,0,0]$
3	$[1,0,0]$	$[1,2,1]$	1	1	$[1,0,0]$
4	$[1,0,0]$	$[1,-3,1]$	1	-1	$[0,3,-1]$
5	$[0,3,-1]$	$[1,-2,-1]$	-5	-1	$[0,3,-1]$
6	$[0,3,-1]$	$[1,-3,-2]$	-7	-1	$[0,3,-1]$
7	$[0,3,-1]$	$[1,0,2]$	-2	1	$[1,3,1]$
8	$[1,3,1]$	$[1,1,2]$	6	1	$[1,3,1]$
9	$[1,3,1]$	$[1,2,1]$	8	1	$[1,3,1]$
10	$[1,3,1]$	$[1,-3,1]$	-7	-1	$[1,3,1]$
11	$[1,3,1]$	$[1,-2,-1]$	-6	-1	$[1,3,1]$
12	$[1,3,1]$	$[1,-3,-2]$	-10	-1	$[1,3,1]$
13	$[1,3,1]$	$[1,0,2]$	3	1	$[1,3,1]$

Learning has converged (we have gone through all the data without needing to update the parameters), so required parameters are $\mathbf{a} = (1,3,1)^t$. *In feature space, the decision surface found after each update looks like this:*



10. Repeat previous question using the sample normalisation method of implementing the Sequential Perceptron Learning Algorithm.

Using Augmented notation and sample normalisation, dataset is:

$\mathbf{x}^t$	$\mathbf{y}^t$
$(0,2)$	$(1,0,2)$
$(1,2)$	$(1,1,2)$
$(2,1)$	$(1,2,1)$
$(-3,1)$	$(-1,3,-1)$
$(-2,-1)$	$(-1,2,1)$
$(-3,-2)$	$(-1,3,2)$

For the Sequential Perceptron Learning Algorithm, weights are updated such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta\mathbf{y}_k$, *where* $\mathbf{y}_k$ *is a misclassified exemplar. Here,* $\eta = 1$.

iteration	$\mathbf{a}^t_{old}$	$\mathbf{y}^t$	$g(\mathbf{x}) = \mathbf{a}^t\mathbf{y}$	$\mathbf{a}^t_{new} = \mathbf{a}^t_{old} + \mathbf{y}^t$ *if misclassified*
1	$[1,0,0]$	$[1,0,2]$	1	$[1,0,0]$
2	$[1,0,0]$	$[1,1,2]$	1	$[1,0,0]$
3	$[1,0,0]$	$[1,2,1]$	1	$[1,0,0]$
4	$[1,0,0]$	$[-1,3,-1]$	-1	$[0,3,-1]$
5	$[0,3,-1]$	$[-1,2,1]$	5	$[0,3,-1]$
6	$[0,3,-1]$	$[-1,3,2]$	7	$[0,3,-1]$
7	$[0,3,-1]$	$[1,0,2]$	-2	$[1,3,1]$
8	$[1,3,1]$	$[1,1,2]$	6	$[1,3,1]$
9	$[1,3,1]$	$[1,2,1]$	8	$[1,3,1]$
10	$[1,3,1]$	$[-1,3,-1]$	7	$[1,3,1]$
11	$[1,3,1]$	$[-1,2,1]$	6	$[1,3,1]$
12	$[1,3,1]$	$[-1,3,2]$	10	$[1,3,1]$
13	$[1,3,1]$	$[1,0,2]$	3	$[1,3,1]$

Learning has converged, so required parameters are $\mathbf{a} = (1,3,1)^t$.

11. A data-set consists of exemplars from three classes.
Class 1: $\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}2\\0\end{pmatrix}$. **Class 2:** $\begin{pmatrix}0\\2\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}$. **Class 3:** $\begin{pmatrix}-1\\-1\end{pmatrix}$. **Use the Sequential Multiclass Perceptron Learning algorithm to find the parameters for three linear discriminant functions that will correctly classify this data. Assume initial values for all parameters are zero, and use a learning rate of 1. If more than one discriminant function produces the maximum output, choose the function with the highest index (i.e., the one that represents the largest class label).**

Sequential Multiclass Perceptron Learning algorithm:

- *Initialise* $\mathbf{a}_j$ *for each class.*
- *For each exemplar* $(\mathbf{y}_k, \omega_k)$ *in turn*
  - *Find predicted class* $j = \arg\max_{j'}(\mathbf{a}^T_{j'}\mathbf{y}_k)$
  - *If predicted class is not true class (i.e.,* $j \neq \omega_k$*), update weights:*
    $\mathbf{a}_{\omega_k} = \mathbf{a}_{\omega_k} + \eta\mathbf{y}_k$
    $\mathbf{a}_j = \mathbf{a}_j - \eta\mathbf{y}_k$
- *repeat until weights stop changing*

Using Augmented notation dataset is:

$\mathbf{y}^t$	ω
$(1,1,1)$	1
$(1,2,0)$	1
$(1,0,2)$	2
$(1,-1,1)$	2
$(1,-1,-1)$	3

it	*old parameters*			$\mathbf{y}^t$	$g_i(\mathbf{x}) = \mathbf{a}_i^t\mathbf{y}$			ω	*new parameters*		
	$\mathbf{a}_1^t$	$\mathbf{a}_2^t$	$\mathbf{a}_3^t$		g_1	g_2	g_3		$\mathbf{a}_1^t$	$\mathbf{a}_2^t$	$\mathbf{a}_3^t$
1	[0, 0, 0]	[0, 0, 0]	[0, 0, 0]	[1, 1, 1]	0	0	0	1	[1, 1, 1]	[0, 0, 0]	[-1, -1, -1]
2	[1, 1, 1]	[0, 0, 0]	[-1, -1, -1]	[1, 2, 0]	3	0	-3	1	[1, 1, 1]	[0, 0, 0]	[-1, -1, -1]
3	[1, 1, 1]	[0, 0, 0]	[-1, -1, -1]	[1, 0, 2]	3	0	-3	2	[0, 1, -1]	[1, 0, 2]	[-1, -1, -1]
4	[0, 1, -1]	[1, 0, 2]	[-1, -1, -1]	[1, -1, 1]	-2	3	-1	2	[0, 1, -1]	[1, 0, 2]	[-1, -1, -1]
5	[0, 1, -1]	[1, 0, 2]	[-1, -1, -1]	[1, -1, -1]	0	-1	1	3	[0, 1, -1]	[1, 0, 2]	[-1, -1, -1]
6	[0, 1, -1]	[1, 0, 2]	[-1, -1, -1]	[1, 1, 1]	0	3	-3	1	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]
7	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]	[1, 2, 0]	5	-2	-3	1	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]
8	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]	[1, 0, 2]	1	2	-3	2	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]
9	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]	[1, -1, 1]	-1	2	-1	2	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]
10	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]	[1, -1, -1]	-1	0	1	3	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]
11	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]	[1, 1, 1]	3	0	-3	1	[1, 2, 0]	[0, -1, 1]	[-1, -1, -1]

Learning has converged, so required parameters are $\mathbf{a}_1 = (1, 2, 0)^t, \mathbf{a}_2 = (0, -1, 1)^t, \mathbf{a}_3 = (-1, -1, -1)^t$.

12. Consider the following linearly separable data set.

$\mathbf{x}$	$class$
$(0, 2)^t$	1
$(1, 2)^t$	1
$(2, 1)^t$	1
$(-3, 1)^t$	-1
$(-2, -1)^t$	-1
$(-3, -2)^t$	-1

Use the pseudoinverse (pinv in MATLAB) to calculate the parameters of a linear discriminant function that can be used to classify this data. Use an arbitrary margin vector $\mathbf{b} = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]^t$**.**

Using Augmented notation and sample normalisation, dataset is:

$\mathbf{x}^t$	$\mathbf{y}^t$
$(0, 2)$	$(1, 0, 2)$
$(1, 2)$	$(1, 1, 2)$
$(2, 1)$	$(1, 2, 1)$
$(-3, 1)$	$(-1, 3, -1)$
$(-2, -1)$	$(-1, 2, 1)$
$(-3, -2)$	$(-1, 3, 2)$

$\mathbf{Ya} = \mathbf{b}$ *where* $\mathbf{Y} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 1 & 2 \\ 1 & 2 & 1 \\ -1 & 3 & -1 \\ -1 & 2 & 1 \\ -1 & 3 & 2 \end{pmatrix}$

Find the pseudo-inverse of $\mathbf{Y}$ *(*$\mathbf{Y}^\dagger = (\mathbf{Y}^t\mathbf{Y})^{-1}\mathbf{Y}^t$*) using MATLAB command pinv:*

$$\mathbf{Y}^\dagger = \begin{pmatrix} 0.0682 & 0.1648 & 0.3807 & 0.1023 & -0.2330 & -0.2557 \\ -0.0341 & 0.0426 & 0.1847 & 0.1989 & -0.0085 & 0.0028 \\ 0.1402 & 0.0748 & -0.1203 & -0.2064 & 0.1184 & 0.1828 \end{pmatrix}$$

Thus, $\mathbf{a} = \mathbf{Y}^\dagger \mathbf{b} = \begin{pmatrix} 0.0682 & 0.1648 & 0.3807 & 0.1023 & -0.2330 & -0.2557 \\ -0.0341 & 0.0426 & 0.1847 & 0.1989 & -0.0085 & 0.0028 \\ 0.1402 & 0.0748 & -0.1203 & -0.2064 & 0.1184 & 0.1828 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

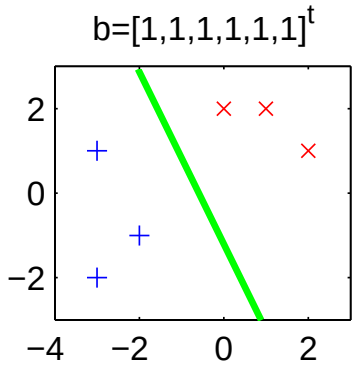
$$\mathbf{a} = \begin{pmatrix} 0.2273 \\ 0.3864 \\ 0.1894 \end{pmatrix}$$

Check:

$\mathbf{y}^t$	$g(\mathbf{x}) = \mathbf{a}^t\mathbf{y} = (0.2273, 0.3864, 0.1894)\mathbf{y}$
$(1, 0, 2)$	0.6061
$(1, 1, 2)$	0.9924
$(1, 2, 1)$	1.1894
$(-1, 3, -1)$	0.7424
$(-1, 2, 1)$	0.7348
$(-1, 3, 2)$	1.3106

All positive, so discriminant function provides correct classification.

In feature space, the decision surface looks like this:

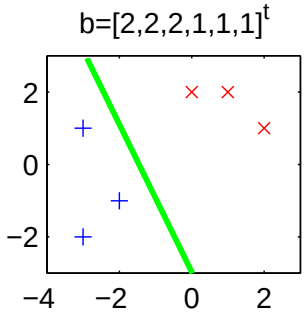


13. Repeat the previous question using (a) $\mathbf{b} = [2, 2, 2, 1, 1, 1]^t$**, and (b)** $\mathbf{b} = [1, 1, 1, 2, 2, 2]^t$**.**

a) $\mathbf{a} = \mathbf{Y}^\dagger \mathbf{b} = \begin{pmatrix} 0.0682 & 0.1648 & 0.3807 & 0.1023 & -0.2330 & -0.2557 \\ -0.0341 & 0.0426 & 0.1847 & 0.1989 & -0.0085 & 0.0028 \\ 0.1402 & 0.0748 & -0.1203 & -0.2064 & 0.1184 & 0.1828 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

$$\mathbf{a} = \begin{pmatrix} 0.8409 \\ 0.5795 \\ 0.2841 \end{pmatrix}$$

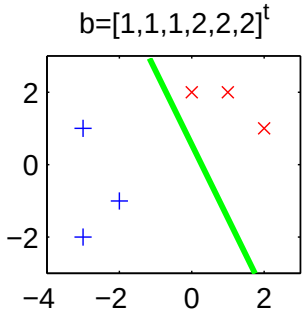
In feature space, the decision surface looks like this:



b) $\mathbf{a} = \mathbf{Y}^{\dagger}\mathbf{b} = \begin{pmatrix} 0.0682 & 0.1648 & 0.3807 & 0.1023 & -0.2330 & -0.2557 \\ -0.0341 & 0.0426 & 0.1847 & 0.1989 & -0.0085 & 0.0028 \\ 0.1402 & 0.0748 & -0.1203 & -0.2064 & 0.1184 & 0.1828 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{pmatrix}$

$$\mathbf{a} = \begin{pmatrix} -0.1591 \\ 0.5795 \\ 0.2841 \end{pmatrix}$$

In feature space, the decision surface looks like this:



Each element of $\mathbf{b}$ *corresponds to a data point. Increasing the value of* $\mathbf{b}$ *increases the margin (i.e. distance) of the hyperplane from the corresponding data point.*

14. For the same dataset used in the preceding question, apply 12 iterations of the Sequential Widrow-Hoff Learning Algorithm. Assume an initial values of $\mathbf{a} = (w_0, \mathbf{w}^t)^t = (1, 0, 0)^t$ **, use a margin vector** $\mathbf{b} = [111111]^t$**, and a learning rate of 0.1.**

Using Augmented notation and sample normalisation, dataset is:

$\mathbf{x}^t$	$\mathbf{y}^t$
$(0, 2)$	$(1, 0, 2)$
$(1, 2)$	$(1, 1, 2)$
$(2, 1)$	$(1, 2, 1)$
$(-3, 1)$	$(-1, 3, -1)$
$(-2, -1)$	$(-1, 2, 1)$
$(-3, -2)$	$(-1, 3, 2)$

For the Sequential Widrow-Hoff Learning Algorithm, weights are updated such that: $\mathbf{a} \leftarrow \mathbf{a} + \eta(b_k - \mathbf{a}^t\mathbf{y}_k)\mathbf{y}_k$. *Here,* $\eta = 0.1$.

it	$\mathbf{a}^t$	$\mathbf{y}_k^t$	$\mathbf{a}^t\mathbf{y}_k$	$\mathbf{a}_{new}^t = \mathbf{a}^t + 0.1(b_k - \mathbf{a}^t\mathbf{y}_k)\mathbf{y}_k^t$
1	$[1, 0, 0]$	$[1, 0, 2]$	1	$[1, 0, 0] + 0.1(1\text{-}1)[1, 0, 2] = [1, 0, 0]$
2	$[1, 0, 0]$	$[1, 1, 2]$	1	$[1, 0, 0] + 0.1(1\text{-}1)[1, 1, 2] = [1, 0, 0]$
3	$[1, 0, 0]$	$[1, 2, 1]$	1	$[1, 0, 0] + 0.1(1\text{-}1)[1, 2, 1] = [1, 0, 0]$
4	$[1, 0, 0]$	$[\text{-}1, 3, \text{-}1]$	-1	$[1, 0, 0] + 0.1(1\text{-}(\text{-}1))[\text{-}1, 3, \text{-}1] = [0.8, 0.6, \text{-}0.2]$
5	[0.8,0.6,–0.2]	$[\text{-}1, 2, 1]$	0.2	[0.8,0.6,–0.2]+0.1(1–0.2)[–1,2,1]=[0.72,0.76,–0.12]
6	[0.72,0.76,–0.12]	$[\text{-}1, 3, 2]$	1.32	[0.72,0.76,–0.12]+0.1(1–1.32)[–1,3,2]=[0.752,0.664,–0.184]
7	[0.752,0.664,–0.184]	$[1, 0, 2]$	0.384	[0.752,0.664,–0.184]+0.1(1–0.384)[[1,0,2]=[0.8136,0.6640,–0.0608]
8	[0.8136,0.6640,–0.0608]	$[1, 1, 2]$	1.356	[0.8136,0.6640,–0.0608]+0.1(1–1.356)[1,1,2]=[0.778,0.6284,–0.1320]
9	[0.778,0.6284,–0.1320]	$[1, 2, 1]$	1.9028	[0.778,0.6284,–0.1320]+0.1(1–1.9028)[1,2,1]=[0.6877,0.4478,–0.2223]
10	[0.6877,0.4478,–0.2223]	$[\text{-}1, 3, \text{-}1]$	0.8781	[0.6877,0.4478,–0.2223]+0.1(1–0.8781)[–1,3,–1]=[0.6755,0.4844,–0.2345]
11	[0.6755,0.4844,–0.2345]	$[\text{-}1, 2, 1]$	0.0588	[0.6755,0.4844,–0.2345]+0.1(1–0.0588)[–1,2,1]=[0.5814,0.6726,–0.1404]
12	[0.5814,0.6726,–0.1404]	$[\text{-}1, 3, 2]$	1.1558	[0.5814,0.6726,–0.1404]+0.1(1–1.1558)[–1,3,2]=[0.597,0.6259,–0.1715]

15. The table shows the training data set for a simple classification problem.

Class	feature vector
1	**(0.15,0.35)**
2	**(0.15,0.28)**
2	**(0.12,0.2)**
3	**(0.1,0.32)**
3	**(0.06,0.25)**

Use the k-nearest-neighbour classifier to determine the class of a new feature vector $\mathrm{x} = (0.1, 0.25)$. Use Euclidean distance and
a) k=1
b) k=3
How would these results be affected if the first dimension of the feature space was scaled by a factor of two?

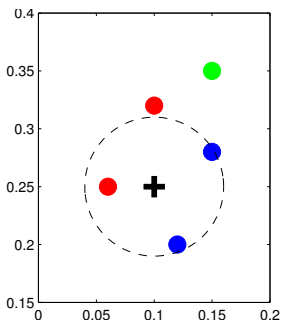
Calculate distance of each sample to x:

Class	*feature vector*	*Euclidean distance to (0.1,0.25)*
1	*(0.15,0.35)*	$\sqrt{(0.1-0.15)^2 + (0.25-0.35)^2} = 0.1118$
2	*(0.15,0.28)*	$\sqrt{(0.1-0.15)^2 + (0.25-0.28)^2} = 0.0583$
2	*(0.12,0.2)*	$\sqrt{(0.1-0.12)^2 + (0.25-0.2)^2} = 0.0539$
3	*(0.1,0.32)*	$\sqrt{(0.1-0.1)^2 + (0.25-0.32)^2} = 0.0700$
3	*(0.06,0.25)*	$\sqrt{(0.1-0.06)^2 + (0.25-0.25)^2} = 0.0400$

a) The nearest neighbour has class label 3. Therefore class new sample as 3.

b) The three nearest neighbours have class labels 3, 2, and 2. Therefore class new sample as 2.

Note, the feature space looks like this:



If we scale the first dimension by a factor of two:

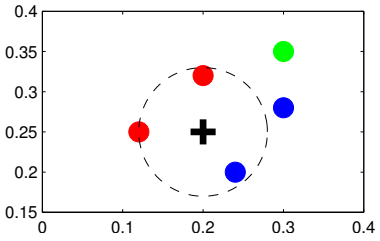
Class	*feature vector*	*Euclidean distance to (0.2,0.25)*
1	*(0.3,0.35)*	$\sqrt{(0.2-0.3)^2+(0.25-0.35)^2}=0.1414$
2	*(0.3,0.28)*	$\sqrt{(0.2-0.3)^2+(0.25-0.28)^2}=0.1044$
2	*(0.24,0.2)*	$\sqrt{(0.2-0.24)^2+(0.25-0.2)^2}=0.0640$
3	*(0.2,0.32)*	$\sqrt{(0.2-0.2)^2+(0.25-0.32)^2}=0.0700$
3	*(0.12,0.25)*	$\sqrt{(0.2-0.12)^2+(0.25-0.25)^2}=0.0800$

For k=1, class of new sample is 2

For k=3, class of new sample is 3

Note, this is the opposite of the result we got previously. kNN is very sensitive to the scale of the feature dimensions!

The feature space now looks like this:

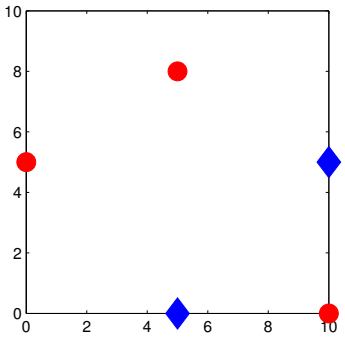


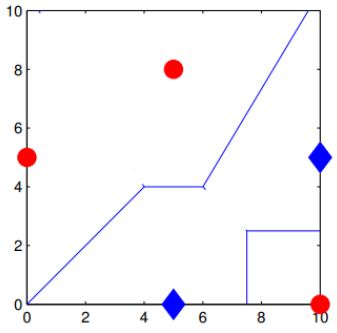
16. a) Plot the decision boundaries that result from applying a nearest neighbour classifier (i.e. a kNN classifier with k=1), to the data shown in the table. Assume Euclidean distance is used to define distance.

Class	$\mathbf{x}_1$	$\mathbf{x}_2$
1	**0**	**5**
	5	**8**
	10	**0**
2	**5**	**0**
	10	**5**

b) For the same data, plot the decision boundaries that result from applying a nearest mean classifier (i.e. one in which a new feature vector, $\mathbf{x}$, is classified by assigning it to the same category as nearest sample mean).

a)





b)