Plug-in Classifier

Suppose $K = 2$.

Let the class densities be Gaussian that differ only in their means $\mathbf{m}_1$ and $\mathbf{m}_0$; let the covariance matrix be $\mathbf{\Sigma}$.

If we know these assumptions are valid, then estimate the unknown parameters and plug them into the classifier

Let $\mathbf{w} = \mathbf{\Sigma}^{-T}(\mathbf{m}_1 - \mathbf{m}_0)$ and $w_p = -(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{\Sigma}^{-1} \frac{(\mathbf{m}_1 + \mathbf{m}_0)}{2} + \log \frac{\pi_0}{\pi_1}$.

Define $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p$. The hyperplane $g(\mathbf{x}) = 0$ is the decision boundary.

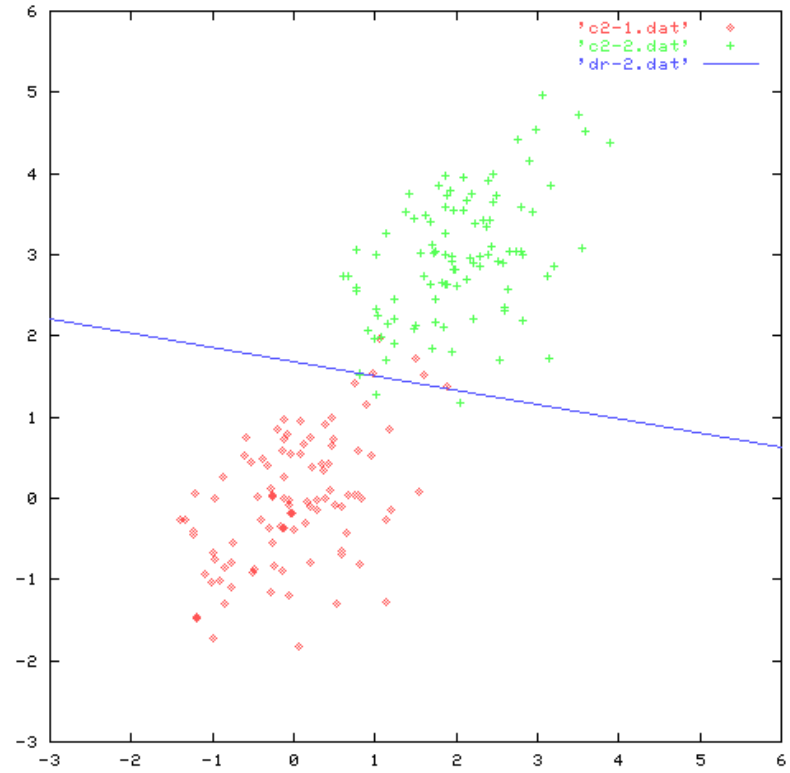
The optimal decision rule is to choose $l = 0$ if $g(\mathbf{x}) < 0$ and $l = 1$ if $g(\mathbf{x}) > 0$.

Linear Classifier

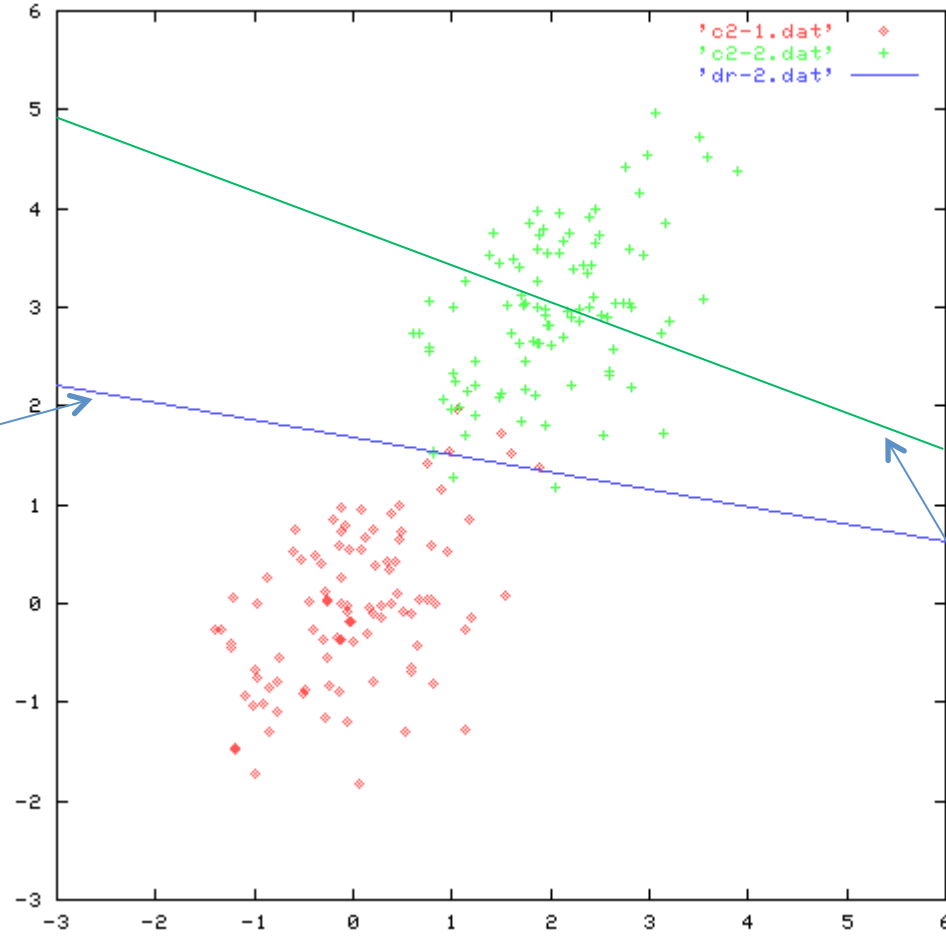
Let $\mathbf{w} = \boldsymbol{\Sigma}^{-T}(\mathbf{m}_1 - \mathbf{m}_0)$ and $w_p = -(\mathbf{m}_1 - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1} \frac{(\mathbf{m}_1 + \mathbf{m}_0)}{2} + \log \frac{\pi_0}{\pi_1}$.

Define $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p$. The hyperplane $g(\mathbf{x}) = 0$ is the decision boundary.

The optimal decision rule is to choose $l = 0$ if $g(\mathbf{x}) < 0$ and $l = 1$ if $g(\mathbf{x}) > 0$.



Linear Decision Boundaries



Optimal decision boundary

Arbitrary linear decision boundary

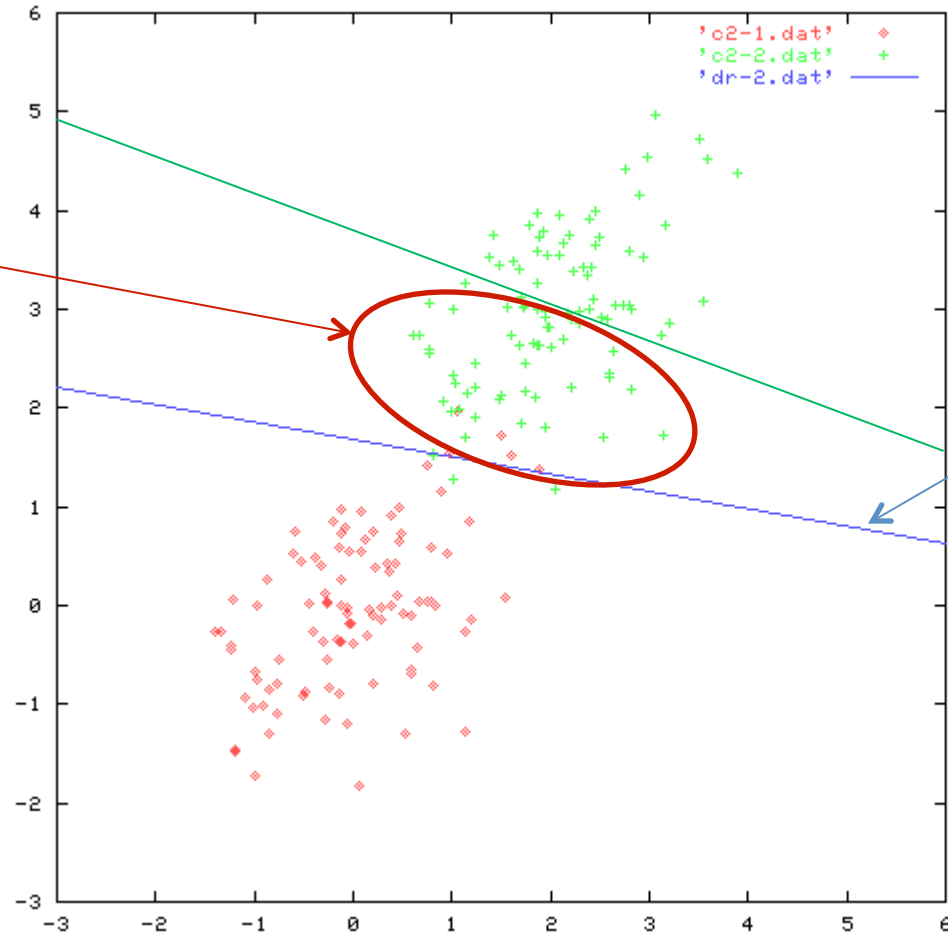
Cheaper to obtain but the performance may be unpredictable

Linear Decision Boundaries

Misclassifications

How can we correct our mistakes?

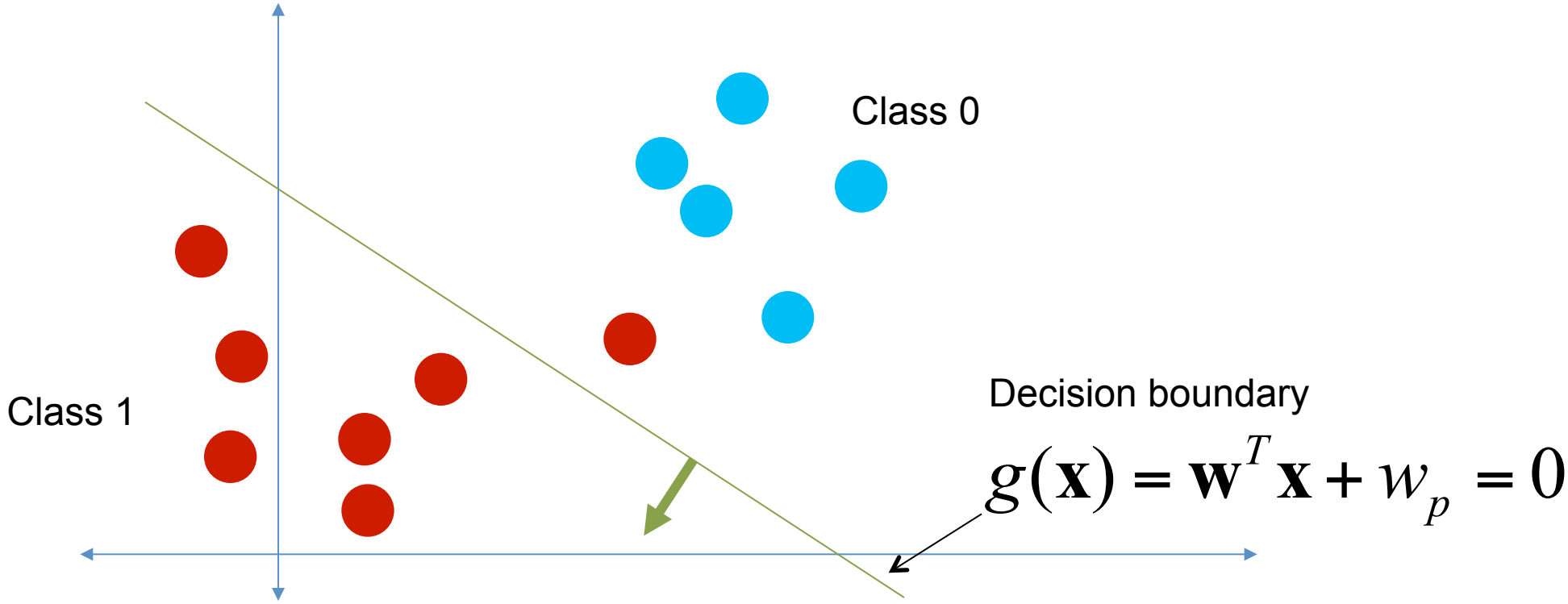
Move the (arbitrary) linear decision boundaries



Optimal decision boundary

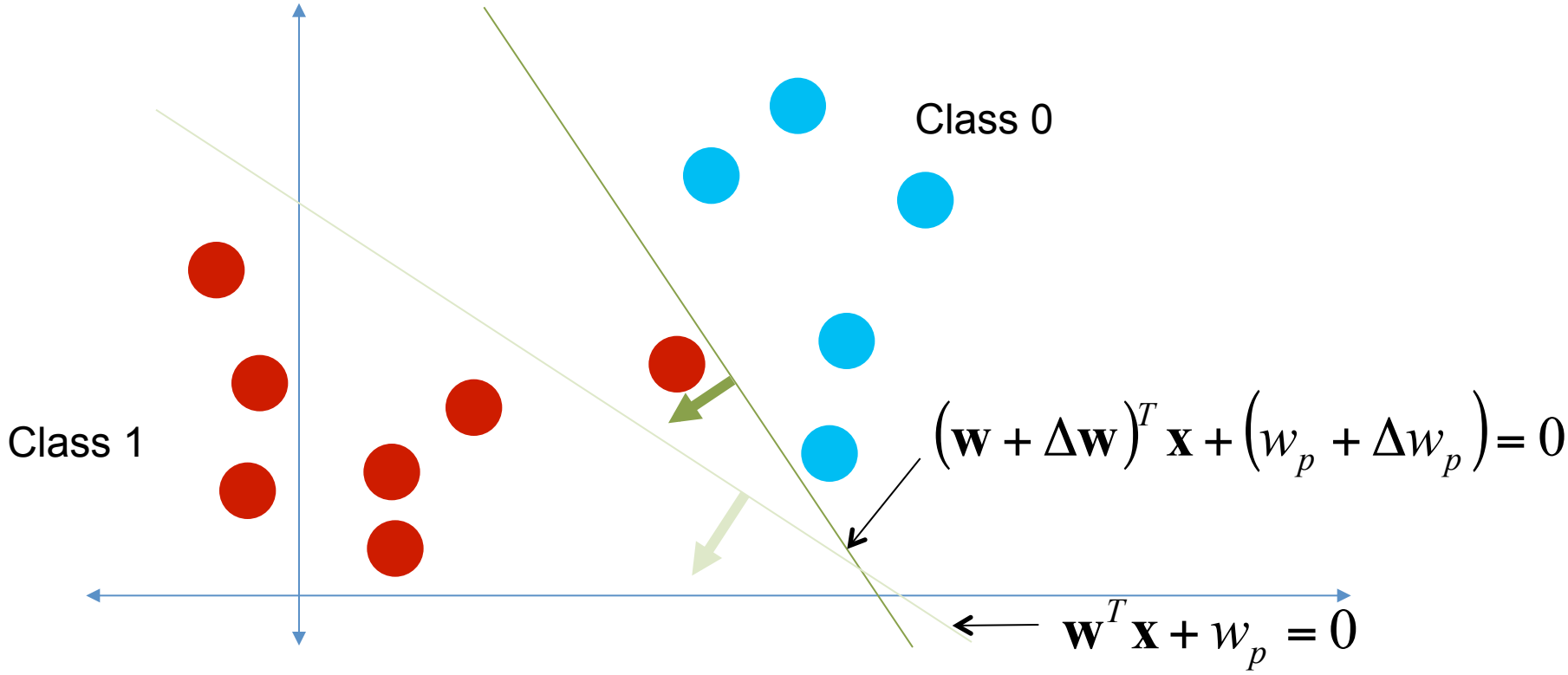
Linear Decision Boundaries

- How do we adjust the linear decision boundary to correct a misclassification?



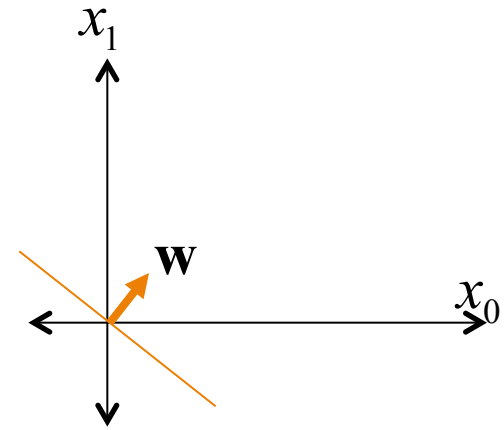
Linear Decision Boundaries

- How do we adjust the linear decision boundary to correct a misclassification?



Adjusting a Linear Boundary

- Think in terms of the weight space



The decision boundary is $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = 0$.
In the feature space (where the features $x_0, x_1, \cdots, x_{p-1}$ are the variables),
the decision boundary has a normal vector given by $\mathbf{w}$.

Suppose a vector $\mathbf{x}^{(0)}$ is observed. Since $\mathbf{w}$ is fixed, $g(\mathbf{x}^{(0)}) = \mathbf{w}^T\mathbf{x}^{(0)}$ is a scalar.

Because $\mathbf{w}^T\mathbf{x} = \mathbf{x}^T\mathbf{w} = 0$, we can think of the feature vector $\mathbf{x}$ as the surface normal
in a weight space, in which the weights $w_0, w_1, \cdots, w_p$ are the variables.

For a fixed (observed) vector $\mathbf{x}^{(0)}$, we have $\mathbf{w}^T\mathbf{x}^{(0)} = \mathbf{x}^{(0)^T}\mathbf{w} = 0$.

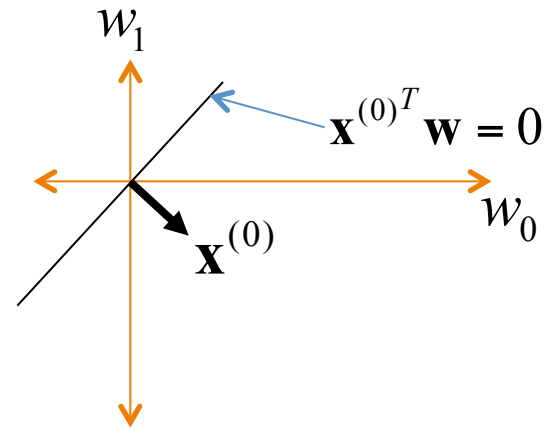
Adjusting a Linear Boundary

- Think in terms of the weight space

For a fixed (observed) vector $\mathbf{x}^{(0)}$, we have $\mathbf{w}^T\mathbf{x}^{(0)} = \mathbf{x}^{(0)T}\mathbf{w} = 0$.

If we think of the weights as variables, then $\mathbf{x}^{(0)T}\mathbf{w} = 0$ is a line/plane/hyperplane in the weight space.

The normal vector of $\mathbf{x}^{(0)T}\mathbf{w} = 0$ is $\mathbf{x}^{(0)}$.



Adjusting a Linear Boundary

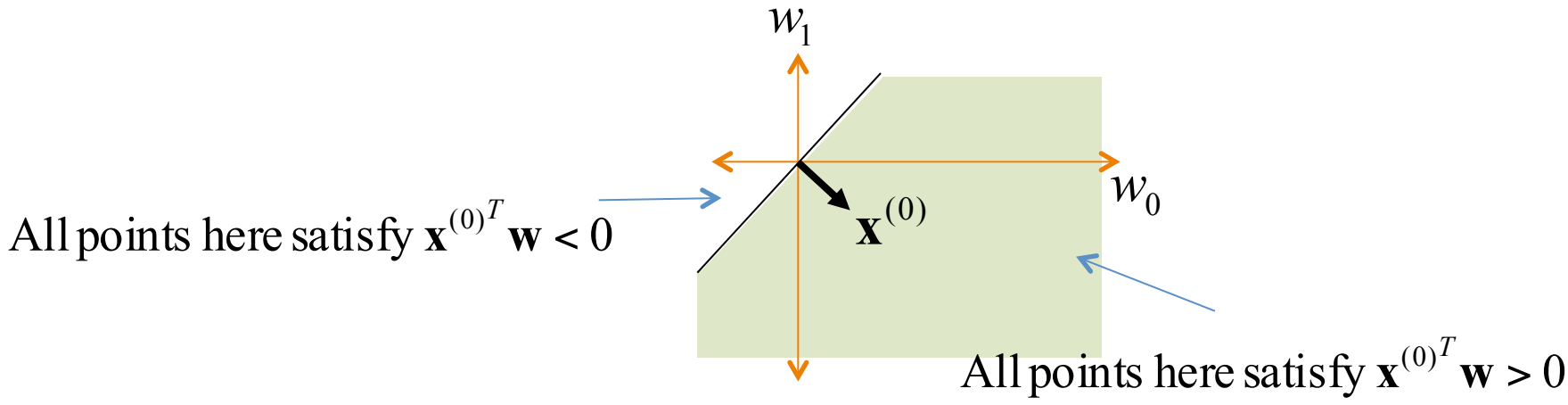
- Regions in the weight space

The line $\mathbf{x}^{(0)^T}\mathbf{w} = 0$ partitions the weight space into a positive and a negative region

The positive region is pointed to by the normal vector of $\mathbf{x}^{(0)^T}\mathbf{w} = 0$, which is $\mathbf{x}^{(0)}$.

All points $\mathbf{w}$ in the positive region satisfy $\mathbf{x}^{(0)^T}\mathbf{w} > 0$.

Similarly, all points $\mathbf{w}$ in the negative region satisfy $\mathbf{x}^{(0)^T}\mathbf{w} < 0$.



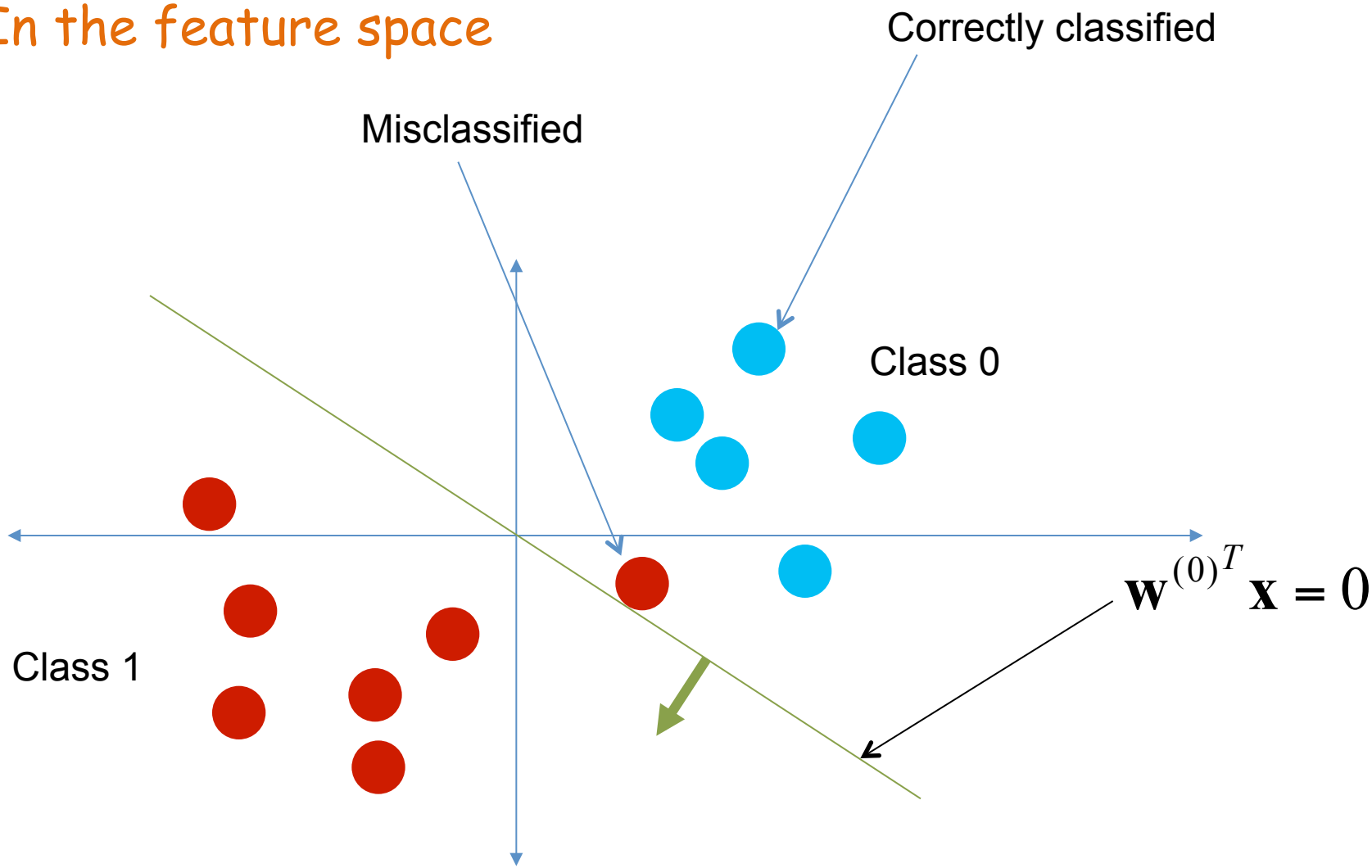
Adjusting a Linear Boundary

Suppose a sample $\left(\mathbf{x}^{(0)}, t^{(0)}\right)$ is given.

Suppose we have a decision boundary specified by $\mathbf{w}^{(0)}$.

Adjusting a Linear Decision Boundary

In the feature space

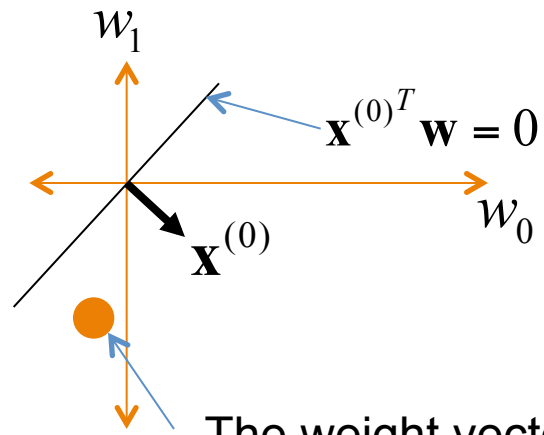


Adjusting a Linear Boundary

Suppose a sample $\left(\mathbf{x}^{(0)}, t^{(0)}\right)$ is given.

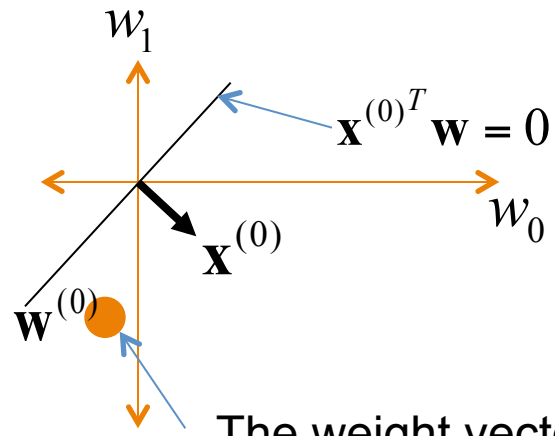
Suppose we have a decision boundary specified by $\mathbf{w}^{(0)}$.

In the weight space



The weight vector is a point in the weight space

In the weight space



The weight vector is a point in the weight space

If $t^{(0)}$ is 1, then the weight vector correctly classified $\mathbf{x}^{(0)}$.

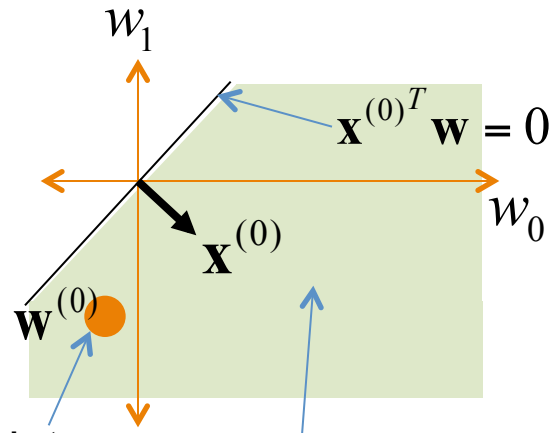
Otherwise, the sample is incorrectly classified.

Now we can think of adjusting the weight vector as moving the weight in the weight space.

Suppose a sample $\left(\mathbf{x}^{(0)}, t^{(0)}\right)$ is given.

Suppose we have a decision boundary specified by $\mathbf{w}$.

In the weight space



The weight vector $\mathbf{w}$ is a point in the weight space

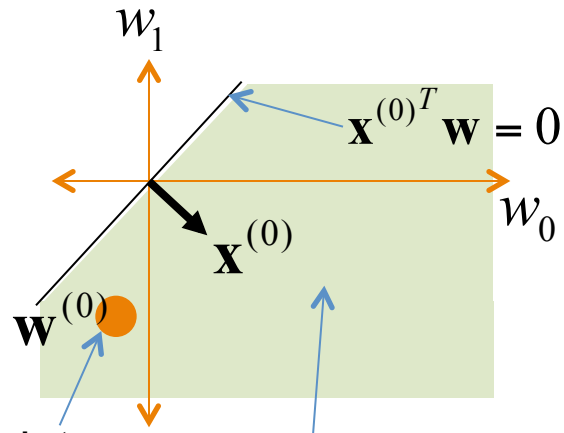
If $t^{(0)}$ is 1, then any weight vector $\mathbf{w}$ in this region correctly classifies $\mathbf{x}^{(0)}$ because

$\mathbf{x}^{(0)T}\mathbf{w} = \mathbf{w}^T\mathbf{x}^{(0)} > 0.$

Suppose a sample $\left(\mathbf{x}^{(0)}, t^{(0)}\right)$ is given.

Suppose we have a decision boundary specified by $\mathbf{w}$.

In the weight space



The weight vector $\mathbf{w}$ is a point in the weight space

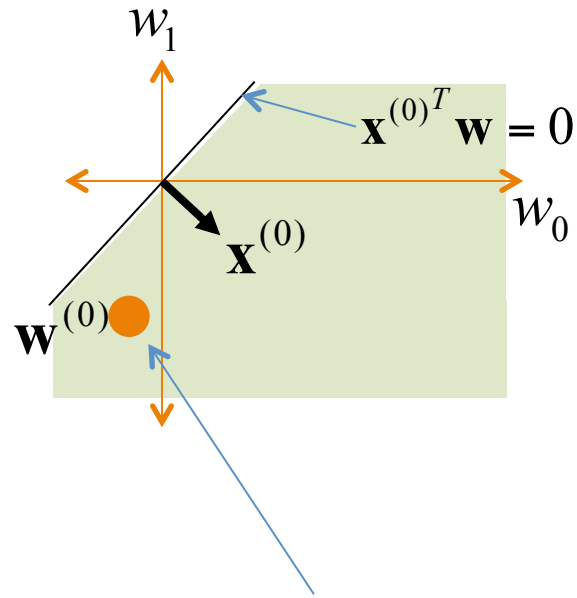
If $t^{(0)}$ is 0, then any weight vector $\mathbf{w}$ in this region misclassifies $\mathbf{x}^{(0)}$ because

$\mathbf{x}^{(0)T}\mathbf{w} = \mathbf{w}^T\mathbf{x}^{(0)} > 0.$

Suppose a sample $\left(\mathbf{x}^{(0)}, 0\right)$ is given; i.e., the feature $\mathbf{x}^{(0)}$ is in Class 0.

Suppose we have a decision boundary specified by $\mathbf{w}$.

In the weight space



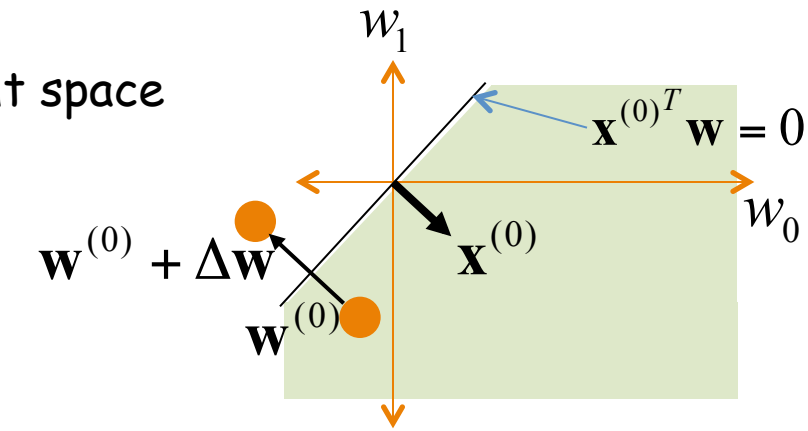
Because $t^{(0)}$ is 0, the weight vector $\mathbf{w}^{(0)}$ misclassifies $\mathbf{x}^{(0)}$ because

$\mathbf{x}^{(0)T}\mathbf{w}^{(0)} = \mathbf{w}^{(0)T}\mathbf{x}^{(0)} > 0.$

Suppose a sample $\left(\mathbf{x}^{(0)}, 0\right)$ is given; i.e., the feature $\mathbf{x}^{(0)}$ is in Class 0.

Suppose we have a decision boundary specified by $\mathbf{w}$.

In the weight space



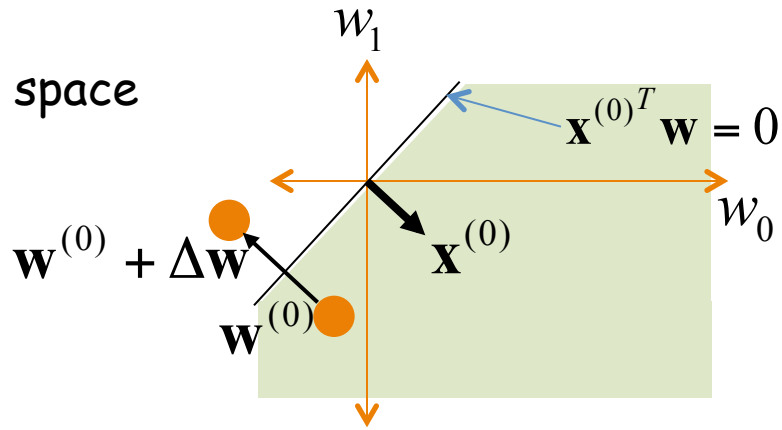
The weight vector $\mathbf{w}^{(0)}$ is adjusted to $\left(\mathbf{w}^{(0)} + \Delta\mathbf{w}\right)$ which is now on the negative side of the boundary defined by $\mathbf{x}^{(0)}$.

Because $\mathbf{x}^{(0)T}\left(\mathbf{w}^{(0)} + \Delta\mathbf{w}\right) = \left(\mathbf{w}^{(0)} + \Delta\mathbf{w}\right)^T \mathbf{x}^{(0)} < 0$, the adjusted weight defines a new decision boundary in the feature space that correctly classifies the feature point $\mathbf{x}^{(0)}$.

Suppose a sample $(\mathbf{x}^{(0)}, 0)$ is given; i.e., the feature $\mathbf{x}^{(0)}$ is in Class 0.

Suppose we have a decision boundary specified by $\mathbf{w}$.

In the weight space



The direction of $\Delta\mathbf{w}$ is given by $\mathbf{x}^{(0)}$.

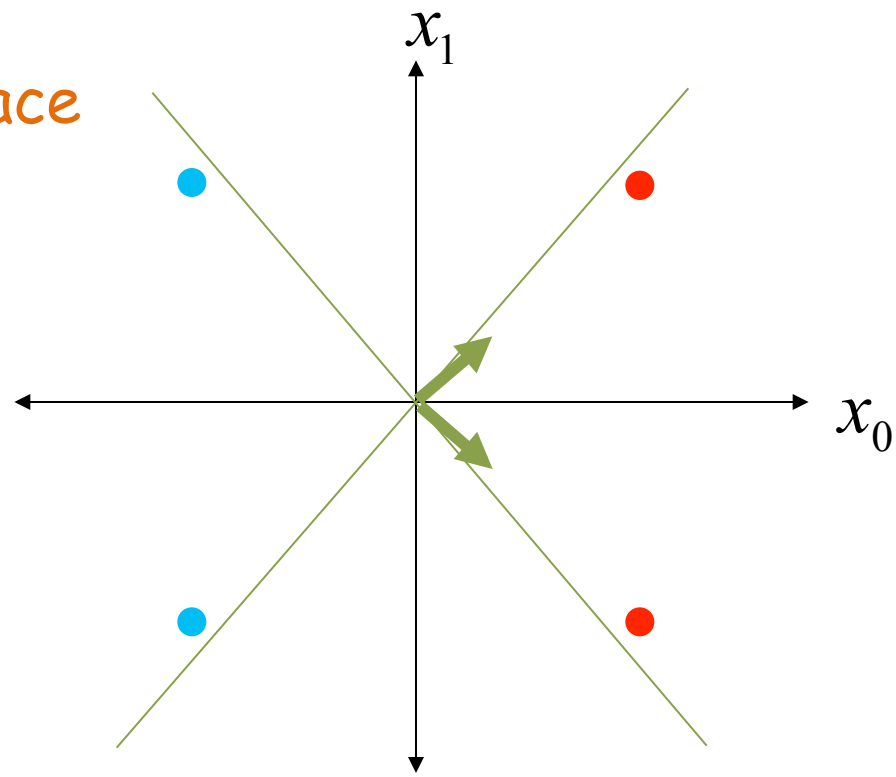
The distance of $\mathbf{w}$ from $\mathbf{x}^{(0)T}\mathbf{w} = 0$ is

Because $\mathbf{x}^{(0)T}(\mathbf{w} + \Delta\mathbf{w}) = (\mathbf{w} + \Delta\mathbf{w})^T\mathbf{x}^{(0)} < 0$, the adjusted weight defines a new decision boundary in the feature space that correctly classifies the feature point $\mathbf{x}^{(0)}$.

An Example

Let $\left\{\left(\begin{bmatrix}-1\\-1\end{bmatrix},0\right),\left(\begin{bmatrix}-1\\+1\end{bmatrix},0\right),\left(\begin{bmatrix}+1\\-1\end{bmatrix},1\right),\left(\begin{bmatrix}+1\\+1\end{bmatrix},1\right)\right\}$ be a training set.

Feature Space



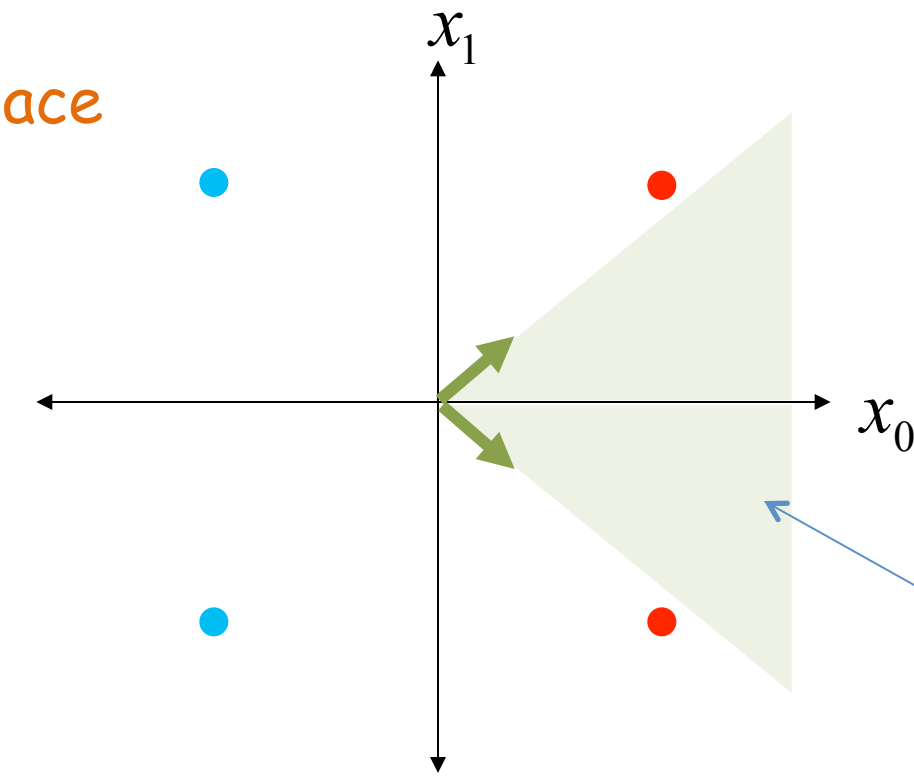
Some observations:

- Linearly separable set
- Many solutions exist
- "Optimal" decision boundary?

An Example

Let $\left\{\left(\begin{bmatrix}-1\\-1\end{bmatrix},0\right),\left(\begin{bmatrix}-1\\+1\end{bmatrix},0\right),\left(\begin{bmatrix}+1\\-1\end{bmatrix},1\right),\left(\begin{bmatrix}+1\\+1\end{bmatrix},1\right)\right\}$ be a training set.

Feature Space



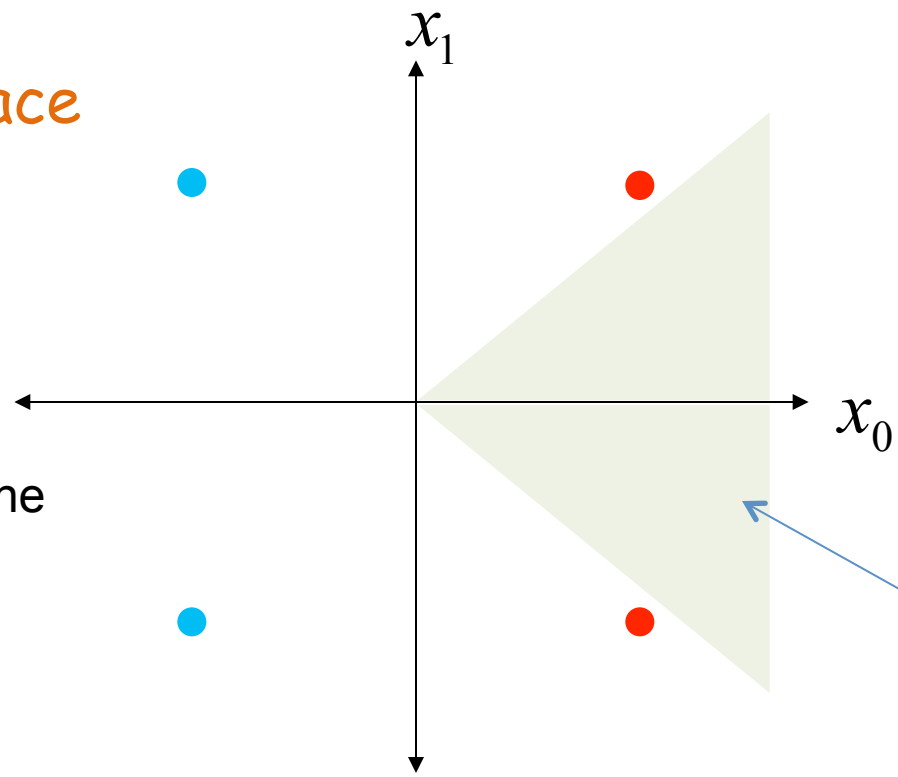
Some observations:
- Linearly separable set
- Many solutions exist
- "Optimal" decision boundary?

Region of acceptable solutions

An Example

Let $\left\{\left(\begin{bmatrix}-1\\-1\end{bmatrix},0\right),\left(\begin{bmatrix}-1\\+1\end{bmatrix},0\right),\left(\begin{bmatrix}+1\\-1\end{bmatrix},1\right),\left(\begin{bmatrix}+1\\+1\end{bmatrix},1\right)\right\}$ be a training set.

Feature Space



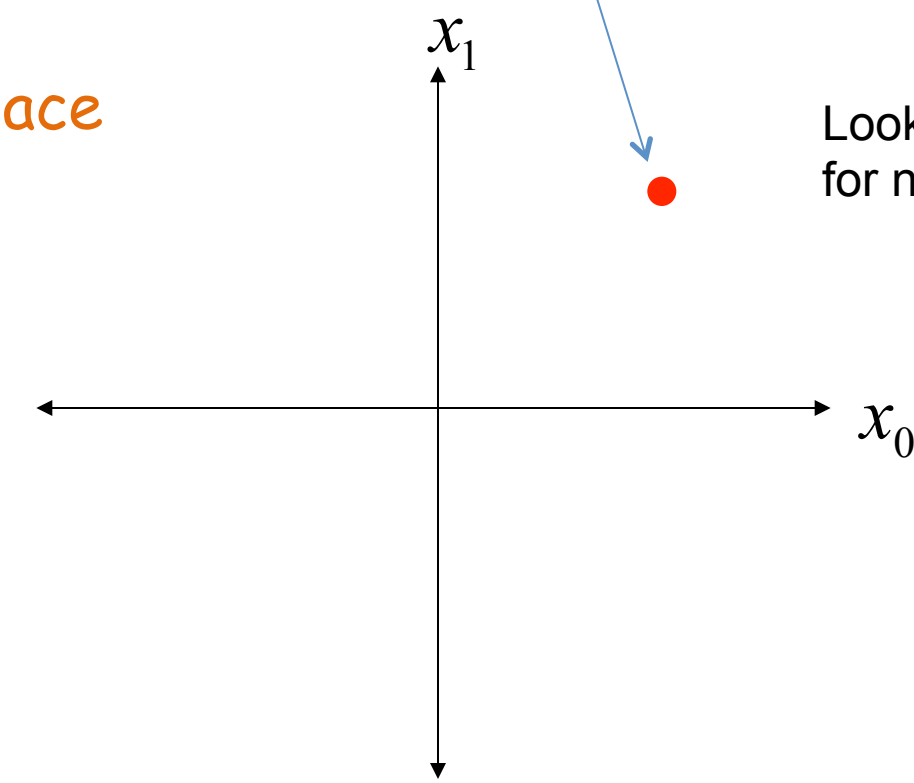
How does this look in the weight space?

Region of acceptable solutions

An Example

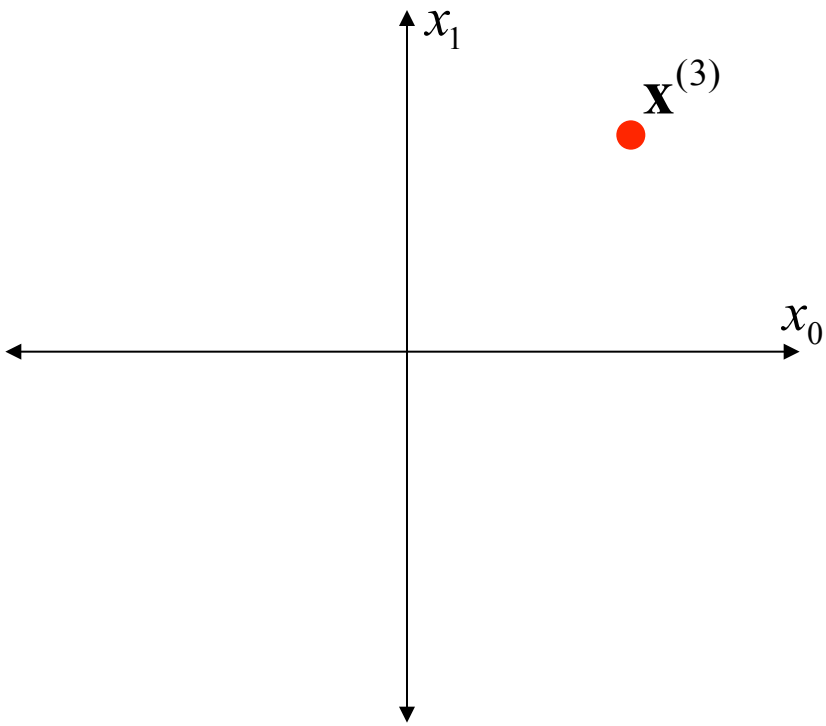
Let $\left\{\left(\begin{bmatrix}-1\\-1\end{bmatrix},0\right),\left(\begin{bmatrix}-1\\+1\end{bmatrix},0\right),\left(\begin{bmatrix}+1\\-1\end{bmatrix},1\right),\left(\begin{bmatrix}+1\\+1\end{bmatrix},1\right)\right\}$ be a training set.

Feature Space



Look at one training sample for now

Feature Space



Let $\left(\begin{bmatrix} +1 \\ +1 \end{bmatrix}, 1 \right)$ be a training sample;

in this example, $\mathbf{x}^{(3)} = \begin{bmatrix} +1 \\ +1 \end{bmatrix}$ and $t^{(3)} = 1$.

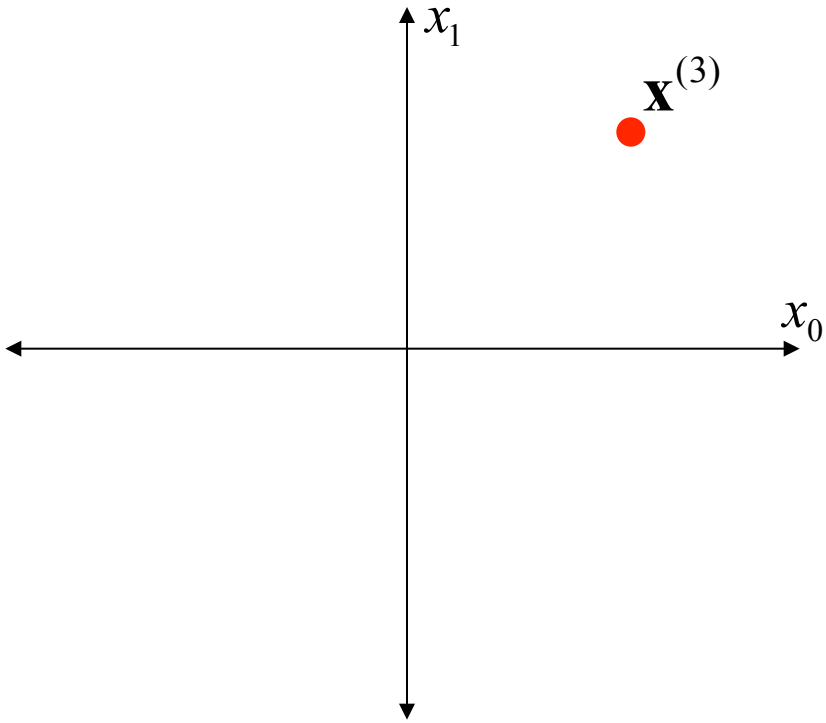
Recall that a decision boundary is given by $\mathbf{w}^T \mathbf{x} = 0$.

In the feature space, the weights form a vector and $\mathbf{x}$ is a point, such as $\mathbf{x}^{(3)}$.

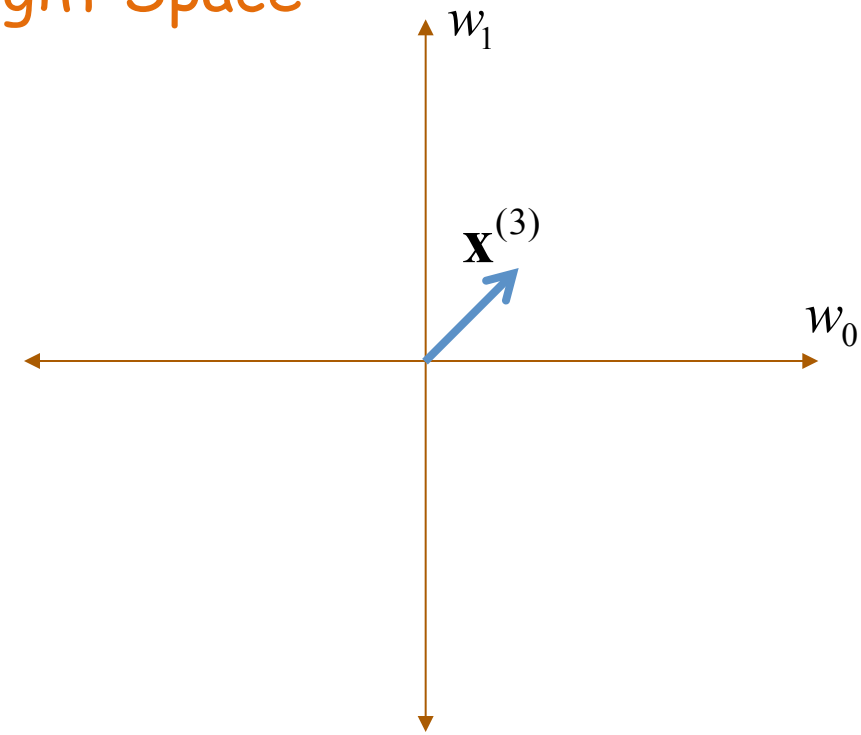
Recall that a decision boundary $\mathbf{w}^T\mathbf{x} = 0$ can be written as $\mathbf{x}^T\mathbf{w} = 0$.

In the weight space, the features form a vector, such as $\mathbf{x}^{(3)}$, and $\mathbf{w}$ is a point.

Feature Space



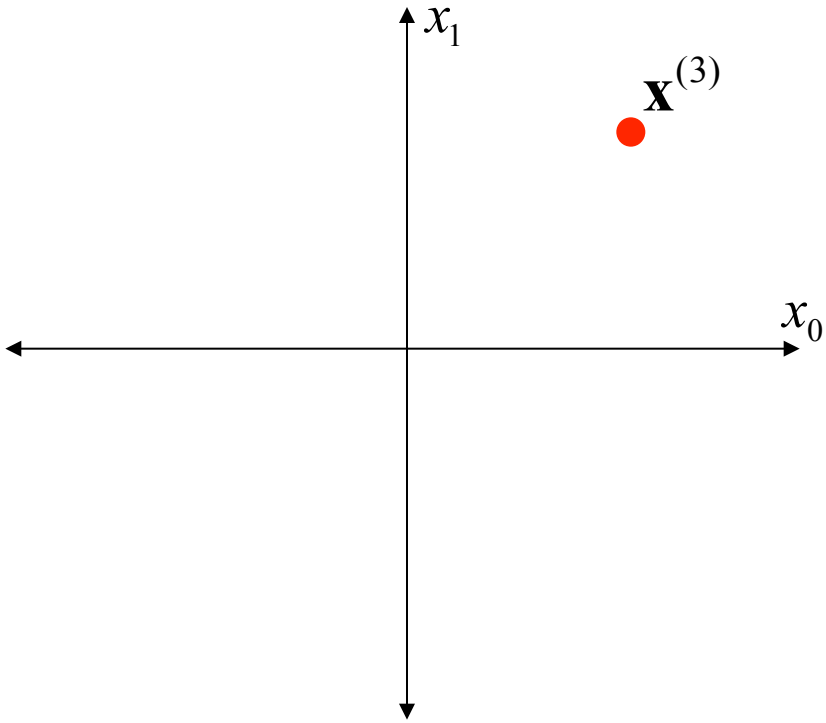
Weight Space



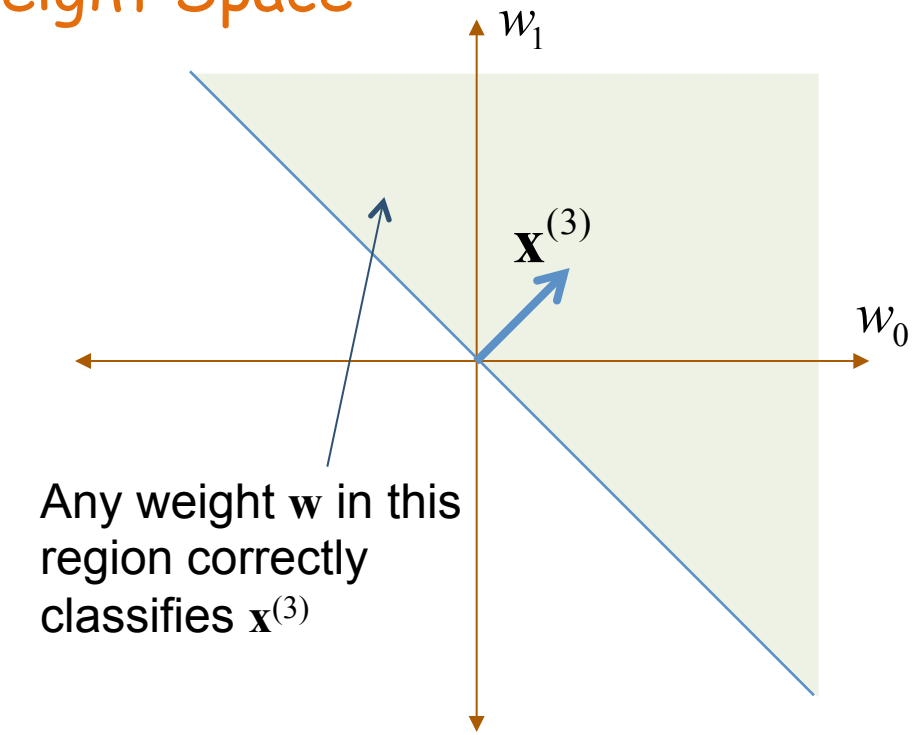
In the weight space, the features form a vector, such as $\mathbf{x}^{(3)}$, and $\mathbf{w}$ is a point.

In the weight space, $\mathbf{x}^{(3)^T}\mathbf{w} = 0$ is a boundary separating those weights $\mathbf{w}$ that satisfy $\mathbf{x}^{(3)^T}\mathbf{w} > 0$ from those that satisfy $\mathbf{x}^{(3)^T}\mathbf{w} < 0$.

Feature Space

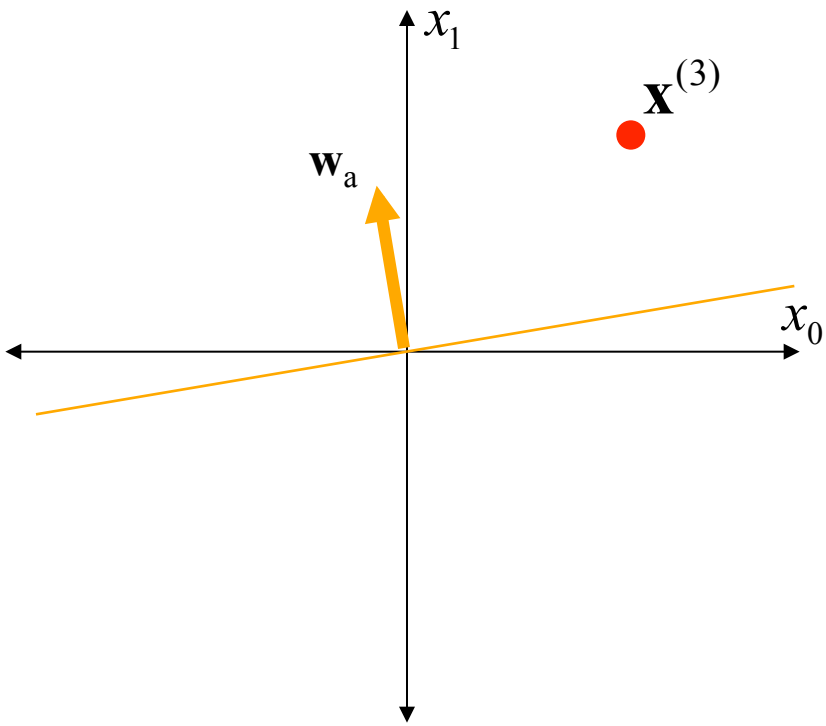


Weight Space

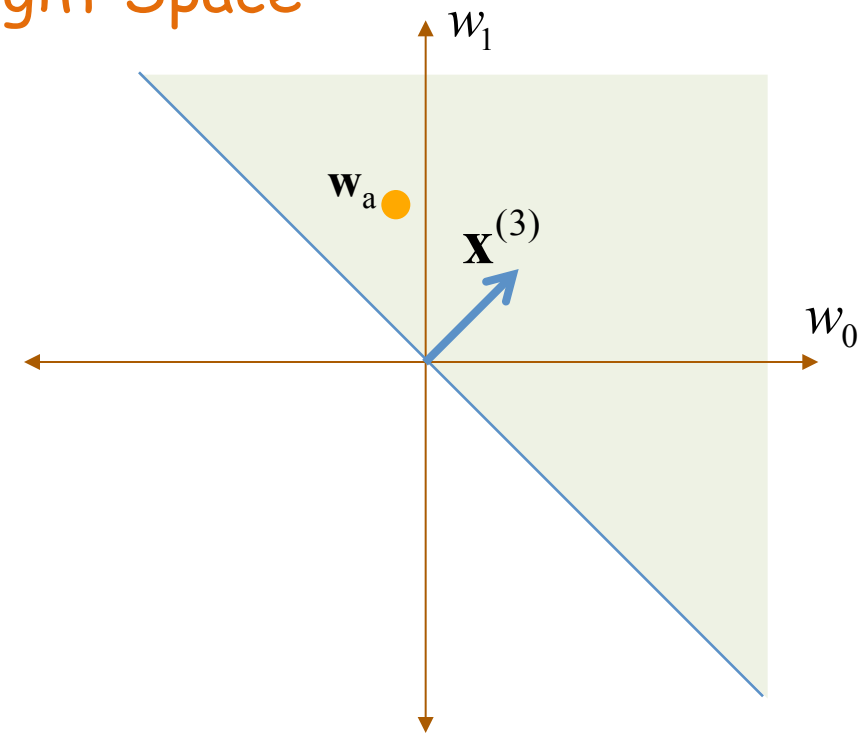


A point $\mathbf{w}_a$ on the positive side of $\mathbf{x}^{(3)\,T}\mathbf{w}=0$ in the weight space correctly classifies the point $\mathbf{x}^{(3)}$ in the feature space because $\mathbf{x}^{(3)}$ is on the positive side of $\mathbf{w}_a{}^T\mathbf{x}=0$ in the feature space.

Feature Space



Weight Space

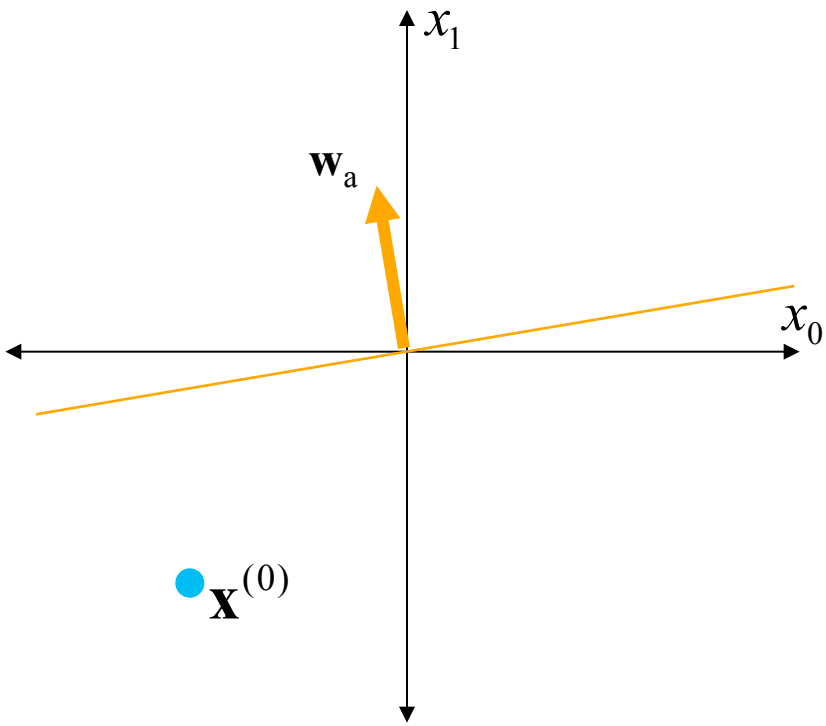


A point $\mathbf{w}_a$ on the negative side of $\mathbf{x}^{(0)\,T}\mathbf{w}=0$ in the weight space correctly classifies the point $\mathbf{x}^{(0)}$ in the feature space because $\mathbf{x}^{(0)}$ is on the negative side of $\mathbf{w}_a{}^T\mathbf{x}=0$ in the feature space.

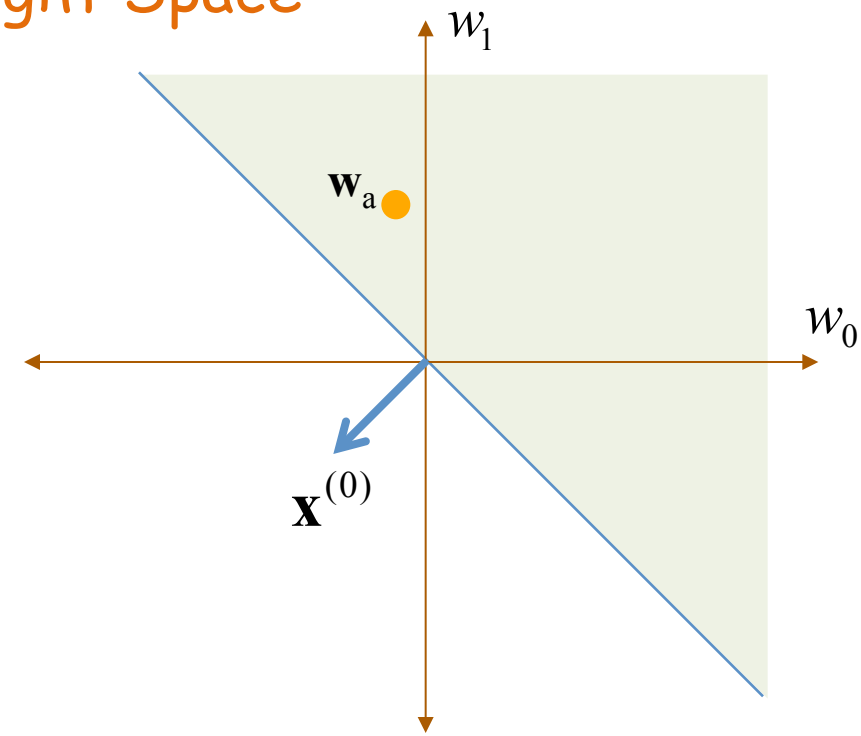
But what about the other training samples?

Let $\left\{\left(\begin{bmatrix}-1\\-1\end{bmatrix},0\right),\left(\begin{bmatrix}-1\\+1\end{bmatrix},0\right),\left(\begin{bmatrix}+1\\-1\end{bmatrix},1\right),\left(\begin{bmatrix}+1\\+1\end{bmatrix},1\right)\right\}$ be a training set.

Feature Space



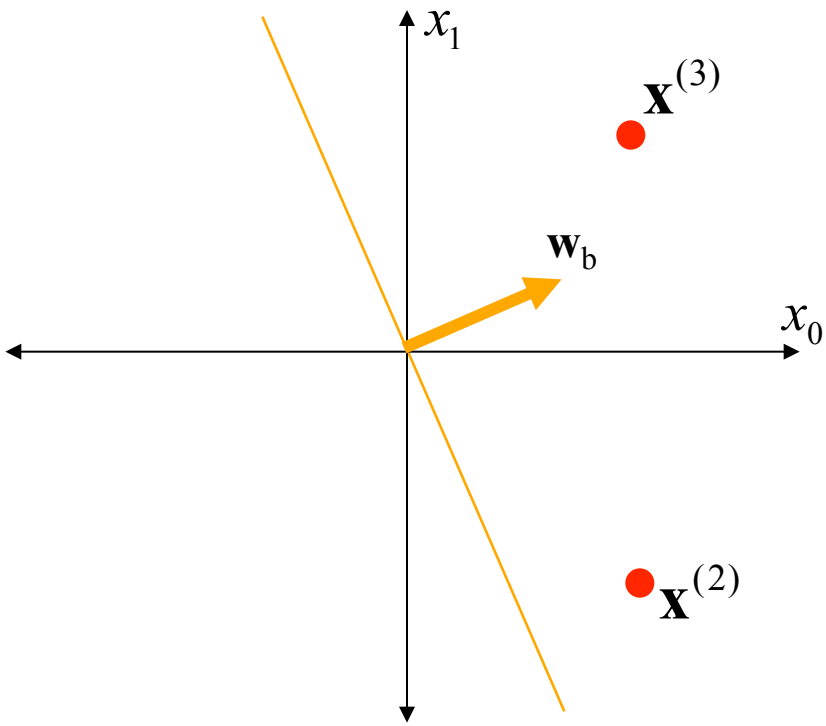
Weight Space



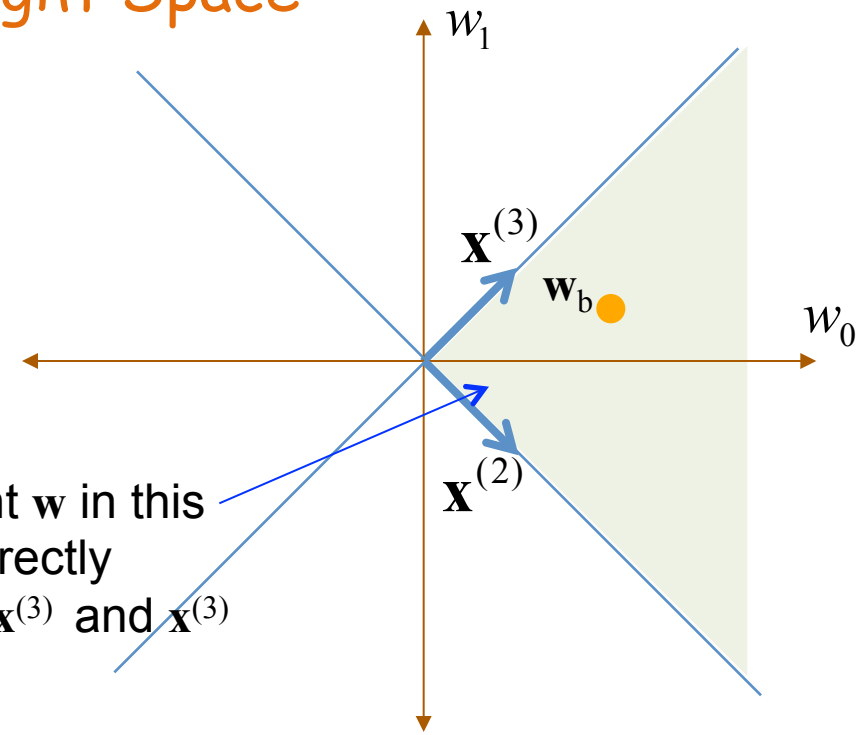
Look at two training samples at the same time

Let $\left\{\left(\begin{bmatrix}-1\\-1\end{bmatrix},0\right),\left(\begin{bmatrix}-1\\+1\end{bmatrix},0\right),\left(\begin{bmatrix}+1\\-1\end{bmatrix},1\right),\left(\begin{bmatrix}+1\\+1\end{bmatrix},1\right)\right\}$ be a training set.

Feature Space



Weight Space

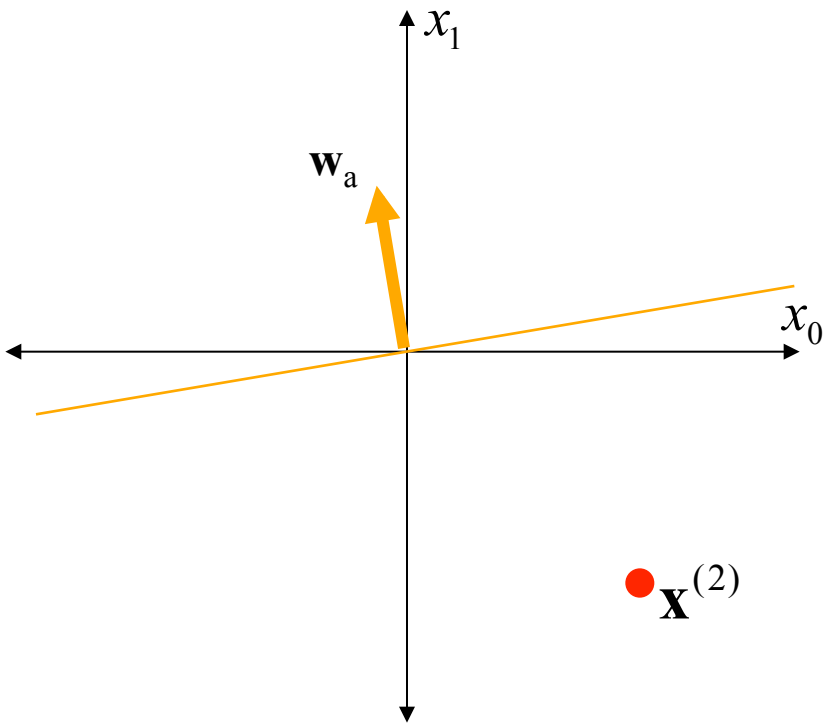


Any weight $\mathbf{w}$ in this region correctly classifies $\mathbf{x}^{(3)}$ and $\mathbf{x}^{(3)}$

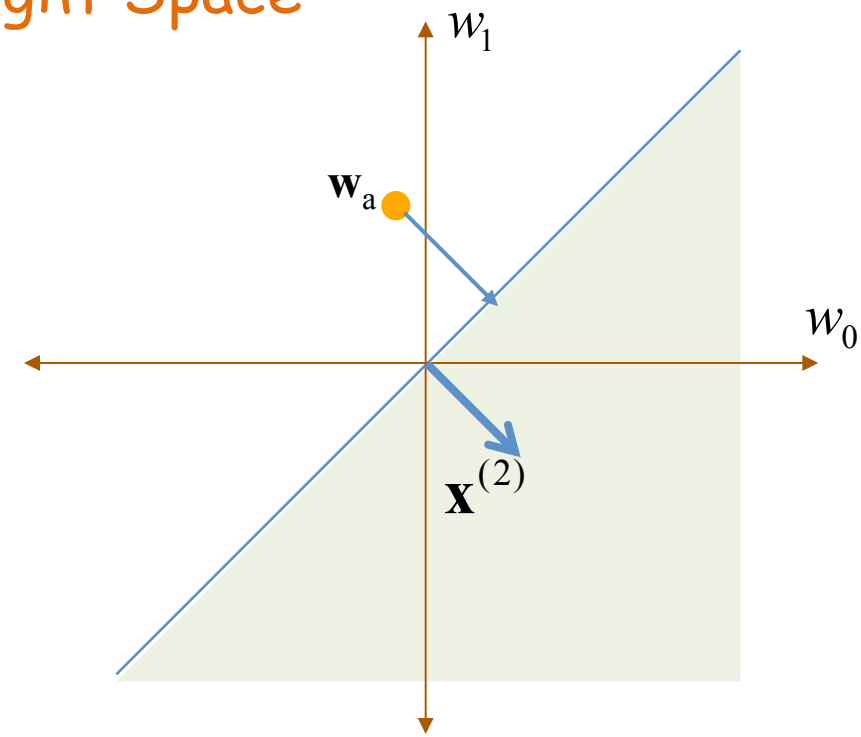
How do we correct $\mathbf{w}_a$ by moving it to the positive side of $\mathbf{x}^{(2)\,T}\mathbf{w}=0$ in the weight space?

Add $c\mathbf{x}^{(2)}$ to it: $\Delta\mathbf{w} = c\mathbf{x}^{(2)}$

Feature Space



Weight Space



After the adjustment $\Delta \mathbf{w} = c\mathbf{x}^{(2)}$, we write

$$\mathbf{w}_{a2} = \mathbf{w}_a + c\mathbf{x}^{(2)}.$$

In the feature space, we have

$$\mathbf{w}_{a2}{}^T \mathbf{x}^{(2)} = \left(\mathbf{w}_a + c\mathbf{x}^{(2)}\right)^T \mathbf{x}^{(2)} = \mathbf{w}_a{}^T \mathbf{x}^{(2)} + c\mathbf{x}^{(2)T}\mathbf{x}^{(2)}.$$

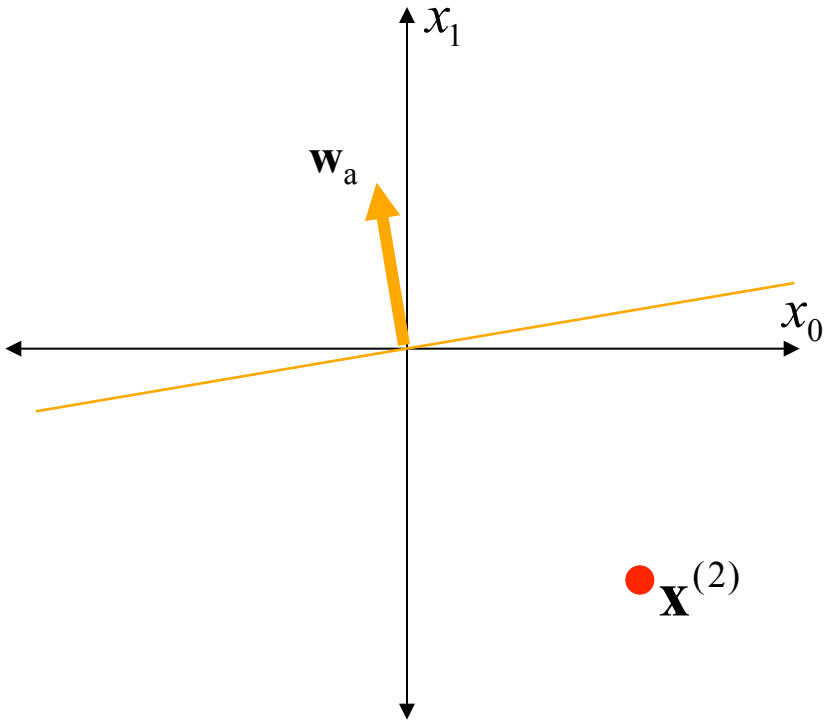
To correctly classify $\mathbf{x}^{(2)}$, we need

$$\mathbf{w}_a{}^T \mathbf{x}^{(2)} + c\mathbf{x}^{(2)T}\mathbf{x}^{(2)} > 0,$$

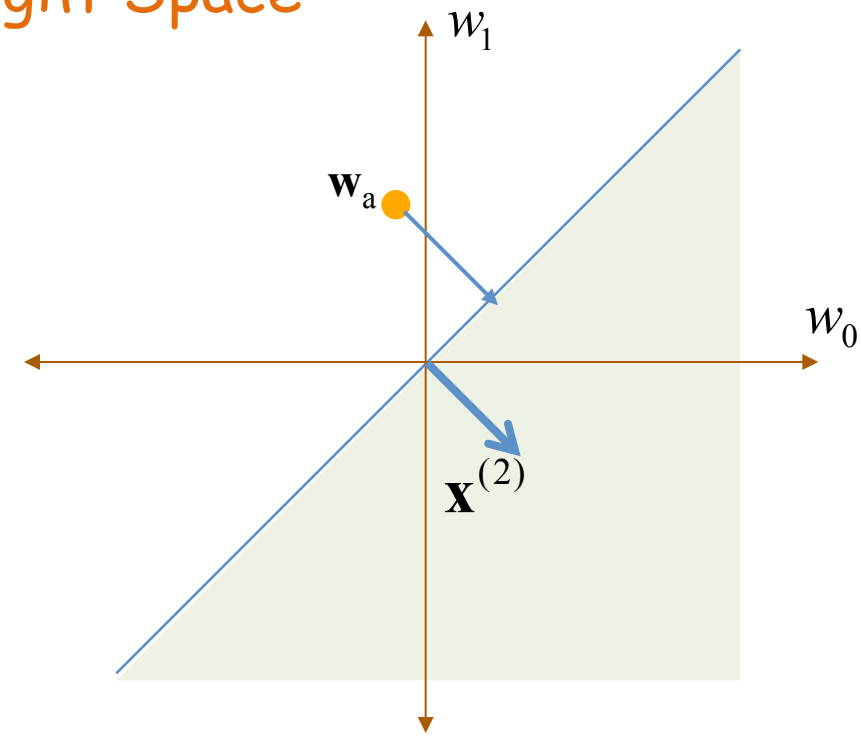
so that

$$c > \frac{-\mathbf{w}_a{}^T \mathbf{x}^{(2)}}{\left\|\mathbf{x}^{(2)}\right\|^2}.$$

Feature Space



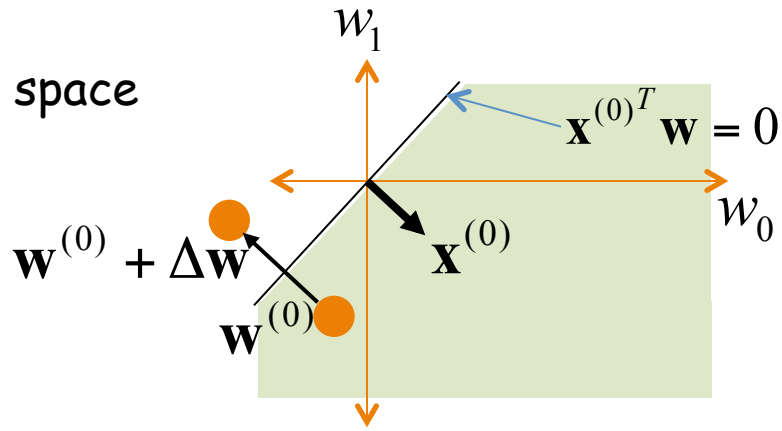
Weight Space



Suppose a sample $\left(\mathbf{x}^{(0)},0\right)$ is given; i.e., the feature $\mathbf{x}^{(0)}$ is in Class 0.

Suppose we have a decision boundary specified by $\mathbf{w}$.

In the weight space



The direction of $\Delta\mathbf{w}$ is given by $\mathbf{x}^{(0)}$.

We want $\mathbf{x}^{(0)T}\left(\mathbf{w}+\Delta\mathbf{w}\right) = \left(\mathbf{w}+\Delta\mathbf{w}\right)^T\mathbf{x}^{(0)} < 0$, so that the adjusted weight defines a new decision boundary in the feature space that correctly classifies the feature point $\mathbf{x}^{(0)}$.

Let the adjustment $\Delta \mathbf{w} = c\mathbf{x}^{(0)}$.

In the feature space, we have

$$\left(\mathbf{w} + c\mathbf{x}^{(0)}\right)^T \mathbf{x}^{(0)} = \mathbf{w}^T \mathbf{x}^{(0)} + c\mathbf{x}^{(0)^T} \mathbf{x}^{(0)}.$$

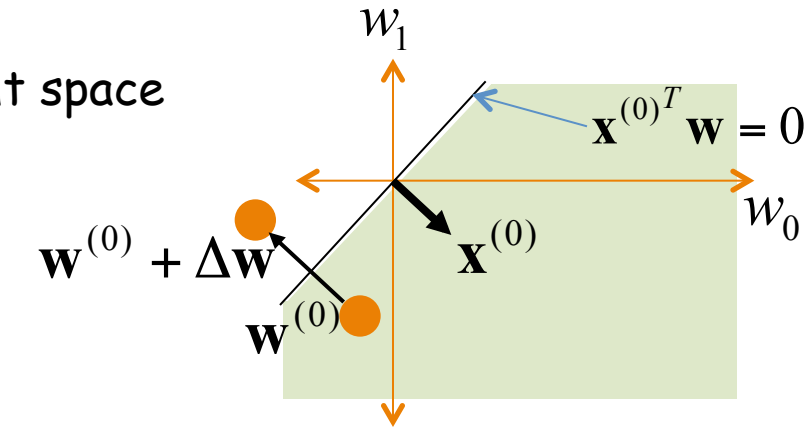
To correctly classify $\mathbf{x}^{(0)}$, we need

$$\mathbf{w}^T \mathbf{x}^{(0)} + c\mathbf{x}^{(0)^T} \mathbf{x}^{(0)} < 0,$$

so that

$$c < \frac{-\mathbf{w}^T \mathbf{x}^{(0)}}{\left\|\mathbf{x}^{(0)}\right\|^2}.$$

In the weight space



Rewrite $\Delta \mathbf{w} = -|c|\mathbf{x}^{(0)}$.

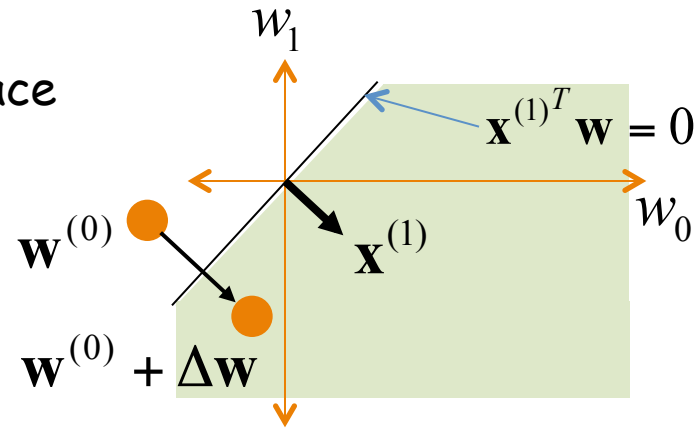
Then

$$|c| > \frac{\mathbf{w}^T \mathbf{x}^{(0)}}{\left\|\mathbf{x}^{(0)}\right\|^2}.$$

Suppose a sample $(\mathbf{x}^{(1)}, 1)$ is given; i.e., the feature $\mathbf{x}^{(1)}$ is in Class 1

Suppose we have a decision boundary specified by $\mathbf{w}$.

In the weight space



The direction of $\Delta\mathbf{w}$ is given by $\mathbf{x}^{(1)}$.

We want $\mathbf{x}^{(1)T}(\mathbf{w} + \Delta\mathbf{w}) = (\mathbf{w} + \Delta\mathbf{w})^T \mathbf{x}^{(1)} > 0$, so that the adjusted weight defines a new decision boundary in the feature space that correctly classifies the feature point $\mathbf{x}^{(1)}$.

Let the adjustment $\Delta \mathbf{w} = c\mathbf{x}^{(1)}$.

In the feature space, we have

$$\left(\mathbf{w} + c\mathbf{x}^{(1)}\right)^T \mathbf{x}^{(1)} = \mathbf{w}^T \mathbf{x}^{(1)} + c\mathbf{x}^{(1)T} \mathbf{x}^{(1)}.$$

To correctly classify $\mathbf{x}^{(1)}$, we need

$$\mathbf{w}^T \mathbf{x}^{(1)} + c\mathbf{x}^{(1)T} \mathbf{x}^{(1)} > 0,$$

so that

$$c > \frac{-\mathbf{w}^T \mathbf{x}^{(1)}}{\left\|\mathbf{x}^{(1)}\right\|^2}.$$

Recall that $\mathbf{w}^T \mathbf{x}^{(1)} < 0$.

Suppose $\mathbf{x}^{(k)}$ is misclassified.

Suppose $t^{(k)} = 0$.

Let the adjustment $\Delta\mathbf{w} = -|c|\mathbf{x}^{(k)}$.

In the feature space, we have

$$\left(\mathbf{w} - |c|\mathbf{x}^{(k)}\right)^T \mathbf{x}^{(k)} = \mathbf{w}^T\mathbf{x}^{(k)} - |c|\mathbf{x}^{(k)T}\mathbf{x}^{(k)}.$$

To correctly classify $\mathbf{x}^{(k)}$, we need

$$\mathbf{w}^T\mathbf{x}^{(k)} - |c|\mathbf{x}^{(k)T}\mathbf{x}^{(k)} < 0,$$

so that

$$|c| > \frac{\mathbf{w}^T\mathbf{x}^{(k)}}{\left\|\mathbf{x}^{(k)}\right\|^2}.$$

Suppose $t^{(k)} = 1$.

Let the adjustment $\Delta\mathbf{w} = c\mathbf{x}^{(k)}$.

In the feature space, we have

$$\left(\mathbf{w} + c\mathbf{x}^{(k)}\right)^T \mathbf{x}^{(k)} = \mathbf{w}^T\mathbf{x}^{(k)} + c\mathbf{x}^{(k)T}\mathbf{x}^{(k)}.$$

To correctly classify $\mathbf{x}^{(k)}$, we need

$$\mathbf{w}^T\mathbf{x}^{(k)} + c\mathbf{x}^{(k)T}\mathbf{x}^{(k)} > 0,$$

so that

$$c > \frac{-\mathbf{w}^T\mathbf{x}^{(k)}}{\left\|\mathbf{x}^{(k)}\right\|^2}.$$

Linear Classifier

Widrow Hoff algorithm

Magnitude of change in weight is a function of the inner product between the weight and the input

Perceptron algorithm

Magnitude of change in weight is constant

Least Squares Problem and Solution

Let $\left\{\left(\mathbf{x}^{(i)}, t^{(i)}\right): i = 0, \cdots, N_1 - 1, N_1, \cdots, N - 1\right\}$ be a training set so that the first N_1 samples are from Class 1 and the next N_0 are from Class 0.

Define a matrix of data in which the ith row is $\mathbf{x}^{(i)T}$: $\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(0)T} \\ \mathbf{x}^{(1)T} \\ \vdots \\ \mathbf{x}^{(N-1)T} \end{bmatrix}$.

Let the matrix of augmented data vectors be written explicitly as $\widetilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \mathbf{1}_N \end{bmatrix}$, where $\mathbf{1}_N$ is a column vector of length N in which all entries are 1.

Put all the target values in a vector $\mathbf{t} = \begin{bmatrix} t^{(0)} \\ t^{(1)} \\ \vdots \\ t^{(N-1)} \end{bmatrix}$. Note that $\mathbf{t}$ can be written as $\begin{bmatrix} \mathbf{1}_{N_1} \\ -\mathbf{1}_{N_0} \end{bmatrix}$.

Then the weight vector that defines a linear classifier can be found from $\widetilde{\mathbf{X}} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \mathbf{t}$.

The least squares solution that minimizes $\left\| \widetilde{\mathbf{X}} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} - \mathbf{t} \right\|^2$ is $\begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \left(\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^T \mathbf{t}$.

Least Squares Problem

The least squares solution finds a weight vector and the bias from $\tilde{\mathbf{X}}\begin{bmatrix}\mathbf{w}\\ w_p\end{bmatrix} = \mathbf{t}$.

We can expand the terms and write $\begin{bmatrix}\mathbf{X} & \mathbf{1}_N\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ w_p\end{bmatrix} = \begin{bmatrix}\mathbf{1}_{N_1}\\ -\mathbf{1}_{N_0}\end{bmatrix}$.

If we put the Class 1 vectors as rows of $\mathbf{X}_1$ and the Class 0 vectors as rows of $\mathbf{X}_0$,

we can write $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_0\end{bmatrix}$, so that the least squares problem becomes finding the weight vector and bias from $\begin{bmatrix}\mathbf{X}_1 & \mathbf{1}_{N_1}\\ \mathbf{X}_0 & \mathbf{1}_{N_0}\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ w_p\end{bmatrix} = \begin{bmatrix}\mathbf{1}_{N_1}\\ -\mathbf{1}_{N_0}\end{bmatrix}$.

The least squares solution of course does not change and remains (in expanded terms)

$$\begin{bmatrix}\mathbf{w}\\ w_p\end{bmatrix} = \left(\begin{bmatrix}\mathbf{X}_1 & \mathbf{1}_{N_1}\\ \mathbf{X}_0 & \mathbf{1}_{N_0}\end{bmatrix}^T\begin{bmatrix}\mathbf{X}_1 & \mathbf{1}_{N_1}\\ \mathbf{X}_0 & \mathbf{1}_{N_0}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{X}_1 & \mathbf{1}_{N_1}\\ \mathbf{X}_0 & \mathbf{1}_{N_0}\end{bmatrix}^T\begin{bmatrix}\mathbf{1}_{N_1}\\ -\mathbf{1}_{N_0}\end{bmatrix}.$$

Least Squares with Rectified Training Data

The least squares solution finds a weight vector and the bias from $\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ \mathbf{X}_0 & \mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{N_1} \\ -\mathbf{1}_{N_0} \end{bmatrix}$.

If we rectify the data set , the least squares problem becomes finding the weight vector and bias from $\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{N_1} \\ \mathbf{1}_{N_0} \end{bmatrix}$.

The least squares solution is

$$\begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \left(\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{1}_{N_1} \\ \mathbf{1}_{N_0} \end{bmatrix}.$$

It can be shown that rectifying the training vectors does not change the solution by verifying

$$\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ \mathbf{X}_0 & \mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ \mathbf{X}_0 & \mathbf{1}_{N_0} \end{bmatrix} \text{ and}$$

$$\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{1}_{N_1} \\ \mathbf{1}_{N_0} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ \mathbf{X}_0 & \mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{1}_{N_1} \\ -\mathbf{1}_{N_0} \end{bmatrix}.$$

A Different Least Squares Problem

Suppose we modify the targets so that the target vector is now $\mathbf{t} = \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix}$;

i.e., targets for the class with fewer samples have higher values.

The least squares problem becomes finding the weight vector and bias from $\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix}$.

The least squares solution is

$$\begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \left(\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix}.$$

Revisit the Least Squares Solution

The least squares problem is to find the weight vector and bias from

$$\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix};$$

i.e., the solution satisfies

$$\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix}.$$

Revisit the Least Squares Solution

$$\text{Since} \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T = \begin{bmatrix} \mathbf{X}_1^T & -\mathbf{X}_0^T \\ \mathbf{1}_{N_1}^T & -\mathbf{1}_{N_0}^T \end{bmatrix},$$

$$\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T & -\mathbf{X}_0^T \\ \mathbf{1}_{N_1}^T & -\mathbf{1}_{N_0}^T \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & \mathbf{X}_1^T\mathbf{1}_{N_1} + \mathbf{X}_0^T\mathbf{1}_{N_0} \\ \mathbf{1}_{N_1}^T\mathbf{X}_1 + \mathbf{1}_{N_0}^T\mathbf{X}_0 & \mathbf{1}_{N_1}^T\mathbf{1}_{N_1} + \mathbf{1}_{N_0}^T\mathbf{1}_{N_0} \end{bmatrix}.$$

The solution therefore satisfies

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & \mathbf{X}_1^T\mathbf{1}_{N_1} + \mathbf{X}_0^T\mathbf{1}_{N_0} \\ \mathbf{1}_{N_1}^T\mathbf{X}_1 + \mathbf{1}_{N_0}^T\mathbf{X}_0 & \mathbf{1}_{N_1}^T\mathbf{1}_{N_1} + \mathbf{1}_{N_0}^T\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix}$$

Revisit the Least Squares Solution

Examine the elements in $\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & \mathbf{X}_1^T\mathbf{1}_{N_1} + \mathbf{X}_0^T\mathbf{1}_{N_0} \\ \mathbf{1}_{N_1}^T\mathbf{X}_1 + \mathbf{1}_{N_0}^T\mathbf{X}_0 & \mathbf{1}_{N_1}^T\mathbf{1}_{N_1} + \mathbf{1}_{N_0}^T\mathbf{1}_{N_0} \end{bmatrix}$.

Let $\mathbf{m}_1 = \frac{1}{N_1}\sum_{i=0}^{N_1-1}\mathbf{x}^{(i)}$ and $\mathbf{m}_0 = \frac{1}{N_0}\sum_{i=N_1}^{N_1+N_0-1}\mathbf{x}^{(i)}$.

Then $\mathbf{X}_1^T\mathbf{1}_{N_1} = N_1\mathbf{m}_1$ and $\mathbf{X}_0^T\mathbf{1}_{N_0} = N_0\mathbf{m}_0$ so that the top right element is $N_1\mathbf{m}_1 + N_0\mathbf{m}_0$.

The lower left element is $\mathbf{1}_{N_1}^T\mathbf{X}_1 + \mathbf{1}_{N_0}^T\mathbf{X}_0 = N_1\mathbf{m}_1^T + N_0\mathbf{m}_0^T$.

The lower right element is $\mathbf{1}_{N_1}^T\mathbf{1}_{N_1} + \mathbf{1}_{N_0}^T\mathbf{1}_{N_0} = N_1 + N_0 = N$.

Revisit the Least Squares Solution

The solution satisfies

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & \mathbf{X}_1^T\mathbf{1}_{N_1} + \mathbf{X}_0^T\mathbf{1}_{N_0} \\ \mathbf{1}_{N_1}^T\mathbf{X}_1 + \mathbf{1}_{N_0}^T\mathbf{X}_0 & \mathbf{1}_{N_1}^T\mathbf{1}_{N_1} + \mathbf{1}_{N_0}^T\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix}$$

Since $\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T = \begin{bmatrix} \mathbf{X}_1^T & -\mathbf{X}_0^T \\ \mathbf{1}_{N_1}^T & -\mathbf{1}_{N_0}^T \end{bmatrix}$,

$$\begin{bmatrix} \mathbf{X}_1 & \mathbf{1}_{N_1} \\ -\mathbf{X}_0 & -\mathbf{1}_{N_0} \end{bmatrix}^T \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^T & -\mathbf{X}_0^T \\ \mathbf{1}_{N_1}^T & -\mathbf{1}_{N_0}^T \end{bmatrix} \begin{bmatrix} \frac{N}{N_1}\mathbf{1}_{N_1} \\ \frac{N}{N_0}\mathbf{1}_{N_0} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{N}{N_1}\mathbf{X}_1^T\mathbf{1}_{N_1} - \frac{N}{N_0}\mathbf{X}_0^T\mathbf{1}_{N_0} \\ \frac{N}{N_1}\mathbf{1}_{N_1}^T\mathbf{1}_{N_1} - \frac{N}{N_0}\mathbf{1}_{N_0}^T\mathbf{1}_{N_0} \end{bmatrix} = \begin{bmatrix} N\mathbf{m}_1 - N\mathbf{m}_0 \\ N - N \end{bmatrix} = \begin{bmatrix} N(\mathbf{m}_1 - \mathbf{m}_0) \\ 0 \end{bmatrix}.$$

Revisit the Least Squares Solution

The solution satisfies

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & \mathbf{X}_1^T\mathbf{1}_{N_1} + \mathbf{X}_0^T\mathbf{1}_{N_0} \\ \mathbf{1}_{N_1}^T\mathbf{X}_1 + \mathbf{1}_{N_0}^T\mathbf{X}_0 & \mathbf{1}_{N_1}^T\mathbf{1}_{N_1} + \mathbf{1}_{N_0}^T\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} N(\mathbf{m}_1 - \mathbf{m}_0) \\ 0 \end{bmatrix},$$

or

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & N_1\mathbf{m}_1 + N_0\mathbf{m}_0 \\ N_1\mathbf{m}_1^T + N_0\mathbf{m}_0^T & N \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} N(\mathbf{m}_1 - \mathbf{m}_0) \\ 0 \end{bmatrix}.$$

Revisit the Least Squares Solution

The solution satisfies

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & \mathbf{X}_1^T\mathbf{1}_{N_1} + \mathbf{X}_0^T\mathbf{1}_{N_0} \\ \mathbf{1}_{N_1}^T\mathbf{X}_1 + \mathbf{1}_{N_0}^T\mathbf{X}_0 & \mathbf{1}_{N_1}^T\mathbf{1}_{N_1} + \mathbf{1}_{N_0}^T\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} N(\mathbf{m}_1 - \mathbf{m}_0) \\ 0 \end{bmatrix},$$

or

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & N_1\mathbf{m}_1 + N_0\mathbf{m}_0 \\ N_1\mathbf{m}_1^T + N_0\mathbf{m}_0^T & N \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} N(\mathbf{m}_1 - \mathbf{m}_0) \\ 0 \end{bmatrix}.$$

Equating the two bottom elements, we have

$$\left(N_1\mathbf{m}_1^T + N_0\mathbf{m}_0^T\right)\mathbf{w} + Nw_p = 0$$

so that

$$w_p = -\frac{1}{N}\left(N_1\mathbf{m}_1^T + N_0\mathbf{m}_0^T\right)\mathbf{w}.$$

Revisit the Least Squares Solution

The solution satisfies

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & \mathbf{X}_1^T\mathbf{1}_{N_1} + \mathbf{X}_0^T\mathbf{1}_{N_0} \\ \mathbf{1}_{N_1}^T\mathbf{X}_1 + \mathbf{1}_{N_0}^T\mathbf{X}_0 & \mathbf{1}_{N_1}^T\mathbf{1}_{N_1} + \mathbf{1}_{N_0}^T\mathbf{1}_{N_0} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} N(\mathbf{m}_1 - \mathbf{m}_0) \\ 0 \end{bmatrix},$$

or

$$\begin{bmatrix} \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 & N_1\mathbf{m}_1 + N_0\mathbf{m}_0 \\ N_1\mathbf{m}_1^T + N_0\mathbf{m}_0^T & N \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ w_p \end{bmatrix} = \begin{bmatrix} N(\mathbf{m}_1 - \mathbf{m}_0) \\ 0 \end{bmatrix}.$$

Equating the two top elements, we have

$$(\mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0)\mathbf{w} + (N_1\mathbf{m}_1 + N_0\mathbf{m}_0)w_p = N(\mathbf{m}_1 - \mathbf{m}_0),$$

to solve for $\mathbf{w}$.

Sample Covariance Matrix

For $k = 0$ and 1, define $\mathbf{S}_k = \sum_i \left(\mathbf{x}^{(i)} - \mathbf{m}_k\right)\left(\mathbf{x}^{(i)} - \mathbf{m}_k\right)^T$,

which simplifies to $\mathbf{S}_k = \sum_i \mathbf{x}^{(i)}\mathbf{x}^{(i)^T} - N_k \mathbf{m}_k \mathbf{m}_k^T$,

where the summation is taken over the vectors in Class k.

For either class, $\sum_i \mathbf{x}^{(i)}\mathbf{x}^{(i)^T}$ is a $(p \times p)$ matrix and can be

shown to equal $\mathbf{X}_k^T \mathbf{X}_k$.

The within - class scatter matrix (or the sample covariance matrix)

can then be written as

$$S = S_1 + S_0 = \mathbf{X}_1^T\mathbf{X}_1 - N_1\mathbf{m}_1\mathbf{m}_1^T + \mathbf{X}_0^T\mathbf{X}_0 - N_0\mathbf{m}_0\mathbf{m}_0^T$$

$$= \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 - N_1\mathbf{m}_1\mathbf{m}_1^T - N_0\mathbf{m}_0\mathbf{m}_0^T.$$

We want to solve for $\mathbf{w}$ from

$$\left(\mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0\right)\mathbf{w} + \left(N_1\mathbf{m}_1 + N_0\mathbf{m}_0\right)w_p = N\left(\mathbf{m}_1 - \mathbf{m}_0\right).$$

We know that the within - class scatter matrix is

$$S = \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 - N_1\mathbf{m}_1\mathbf{m}_1^T - N_0\mathbf{m}_0\mathbf{m}_0^T.$$

We further know that the bias is

$$w_p = -\frac{1}{N}\left(N_1\mathbf{m}_1^T + N_0\mathbf{m}_0^T\right)\mathbf{w}.$$

The left hand side of the equation is :

$$\left(\mathbf{S} + N_1\mathbf{m}_1\mathbf{m}_1^T - N_0\mathbf{m}_0\mathbf{m}_0^T\right)\mathbf{w} - \frac{1}{N}\left(N_1\mathbf{m}_1 + N_0\mathbf{m}_0\right)\left(N_1\mathbf{m}_1^T + N_0\mathbf{m}_0^T\right)\mathbf{w}$$

$$= \left\{\mathbf{S} + N_1\mathbf{m}_1\mathbf{m}_1^T - N_0\mathbf{m}_0\mathbf{m}_0^T\right\}\mathbf{w} - \left\{\frac{N_1^2}{N}\mathbf{m}_1\mathbf{m}_1^T + \frac{N_1N_0}{N}\mathbf{m}_1\mathbf{m}_0^T + \frac{N_1N_0}{N}\mathbf{m}_0\mathbf{m}_1^T + \frac{N_0^2}{N}\mathbf{m}_0\mathbf{m}_0^T\right\}\mathbf{w}$$

$$= \left\{\mathbf{S} + N_1\mathbf{m}_1\mathbf{m}_1^T - N_0\mathbf{m}_0\mathbf{m}_0^T - \frac{N_1^2}{N}\mathbf{m}_1\mathbf{m}_1^T - \frac{N_1N_0}{N}\mathbf{m}_1\mathbf{m}_0^T - \frac{N_1N_0}{N}\mathbf{m}_0\mathbf{m}_1^T - \frac{N_0^2}{N}\mathbf{m}_0\mathbf{m}_0^T\right\}\mathbf{w}.$$

Consider the coefficients of $\mathbf{m}_k\mathbf{m}_k^T, k = 0$ and 1 in

$$\left\{\mathbf{S} + N_1\mathbf{m}_1\mathbf{m}_1^T - N_0\mathbf{m}_0\mathbf{m}_0^T - \frac{N_1^2}{N}\mathbf{m}_1\mathbf{m}_1^T - \frac{N_1N_0}{N}\mathbf{m}_1\mathbf{m}_0^T - \frac{N_1N_0}{N}\mathbf{m}_0\mathbf{m}_1^T - \frac{N_0^2}{N}\mathbf{m}_0\mathbf{m}_0^T\right\}\mathbf{w}.$$

The coefficient of $\mathbf{m}_1\mathbf{m}_1^T = N_1 - \frac{N_1^2}{N} = \frac{NN_1 - N_1^2}{N} = \frac{N_1(N - N_1)}{N} = \frac{N_1N_0}{N}$.

Similarly, the coefficient of $\mathbf{m}_0\mathbf{m}_0^T$ is $\frac{N_1N_0}{N}$.

The term therefore becomes

$$\left\{\mathbf{S} + \frac{N_1N_0}{N}\mathbf{m}_1\mathbf{m}_1^T + \frac{N_1N_0}{N}\mathbf{m}_0\mathbf{m}_0^T - \frac{N_1N_0}{N}\mathbf{m}_1\mathbf{m}_0^T - \frac{N_1N_0}{N}\mathbf{m}_0\mathbf{m}_1^T\right\}\mathbf{w}$$

$$= \left\{\mathbf{S} + \frac{N_1N_0}{N}\left(\mathbf{m}_1\mathbf{m}_1^T + \mathbf{m}_0\mathbf{m}_0^T - \mathbf{m}_1\mathbf{m}_0^T - \mathbf{m}_0\mathbf{m}_1^T\right)\right\}\mathbf{w}$$

$$= \left\{\mathbf{S} + \frac{N_1N_0}{N}(\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^T\right\}\mathbf{w} = \mathbf{S}\mathbf{w} + \frac{N_1N_0}{N}(\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^T\mathbf{w}.$$

Recall that we want to solve for $\mathbf{w}$ from

$$\left(\mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0\right)\mathbf{w} + \left(N_1\mathbf{m}_1 + N_0\mathbf{m}_0\right)w_p = N\left(\mathbf{m}_1 - \mathbf{m}_0\right).$$

After manipulating the terms on the left hand side, we have

$$\mathbf{S}\mathbf{w} + \frac{N_1 N_0}{N}\left(\mathbf{m}_1 - \mathbf{m}_0\right)\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T\mathbf{w} = N\left(\mathbf{m}_1 - \mathbf{m}_0\right),$$

so that

$$\frac{1}{N}\mathbf{S}\mathbf{w} + \frac{N_1 N_0}{N}\left(\mathbf{m}_1 - \mathbf{m}_0\right)\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T\mathbf{w} = \left(\mathbf{m}_1 - \mathbf{m}_0\right).$$

A plausible solution is to factor out $\mathbf{w}$ and inverting the left matrix to obtain

$$\mathbf{w} = \left(\frac{1}{N}\mathbf{S} + \frac{N_1 N_0}{N^2}\left(\mathbf{m}_1 - \mathbf{m}_0\right)\left(\mathbf{m}_1 - \mathbf{m}_0\right)^T\right)^{-1}\left(\mathbf{m}_1 - \mathbf{m}_0\right).$$

This is almost never done in practice.

In practice, we solve for the direction of $\mathbf{w}$ from

$$\frac{1}{N}\mathbf{S}\mathbf{w} + \frac{N_1 N_0}{N^2}(\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_0).$$

Observe that the term $\frac{N_1 N_0}{N^2}(\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{w}$ is a vector in the direction of $(\mathbf{m}_1 - \mathbf{m}_0)$ by noting that $(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{w}$ is a scalar.

Let $\gamma_1 = \frac{N_1 N_0}{N^2}$ and $\gamma_2 = (\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{w}$. Next, let $(1-\alpha) = \gamma_1 \gamma_2$.

Then $\frac{N_1 N_0}{N^2}(\mathbf{m}_1 - \mathbf{m}_0)(\mathbf{m}_1 - \mathbf{m}_0)^T \mathbf{w} = (1-\alpha)(\mathbf{m}_1 - \mathbf{m}_0)$.

The solution $\mathbf{w}$ satisfies

$$\frac{1}{N}\mathbf{S}\mathbf{w} + (1-\alpha)(\mathbf{m}_1 - \mathbf{m}_0) = (\mathbf{m}_1 - \mathbf{m}_0), \text{ or, } \frac{1}{N}\mathbf{S}\mathbf{w} = \alpha(\mathbf{m}_1 - \mathbf{m}_0)$$

so that $\mathbf{w} = N\alpha\mathbf{S}^{-1}(\mathbf{m}_1 - \mathbf{m}_0)$.

Of course α is an unknown scalar because it was defined in terms of $\mathbf{w}$.

Therefore, in general, we have the solution

$\mathbf{w} \propto \mathbf{S}^{-1}(\mathbf{m}_1 - \mathbf{m}_0)$.

The classifier defined by $\mathbf{w} \propto \mathbf{S}^{-1}(\mathbf{m}_1 - \mathbf{m}_0)$ is the Fisher linear discriminant.

Fisher Linear Discriminant

The classifier defined by $\mathbf{w} \propto \mathbf{S}^{-1}\left(\mathbf{m}_1 - \mathbf{m}_0\right)$ is the Fisher linear discriminant.

The within-class scatter matrix is:

$$\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_0 = \mathbf{X}_1^T\mathbf{X}_1 - N_1\mathbf{m}_1\mathbf{m}_1^T + \mathbf{X}_0^T\mathbf{X}_0 - N_0\mathbf{m}_0\mathbf{m}_0^T$$

$$= \mathbf{X}_1^T\mathbf{X}_1 + \mathbf{X}_0^T\mathbf{X}_0 - N_1\mathbf{m}_1\mathbf{m}_1^T - N_0\mathbf{m}_0\mathbf{m}_0^T,$$

where $\mathbf{S}_k = \sum_i \left(\mathbf{x}^{(i)} - \mathbf{m}_k\right)\left(\mathbf{x}^{(i)} - \mathbf{m}_k\right)^T$, for $k = 0$ and 1.