universität freiburg

Assignment 04

Machine Learning WS23/24

Assignment 04
Solution

Support Vector Machines

Support Vector Machines

Explain the *kernel trick* and why we use it in Support Vector Machines.

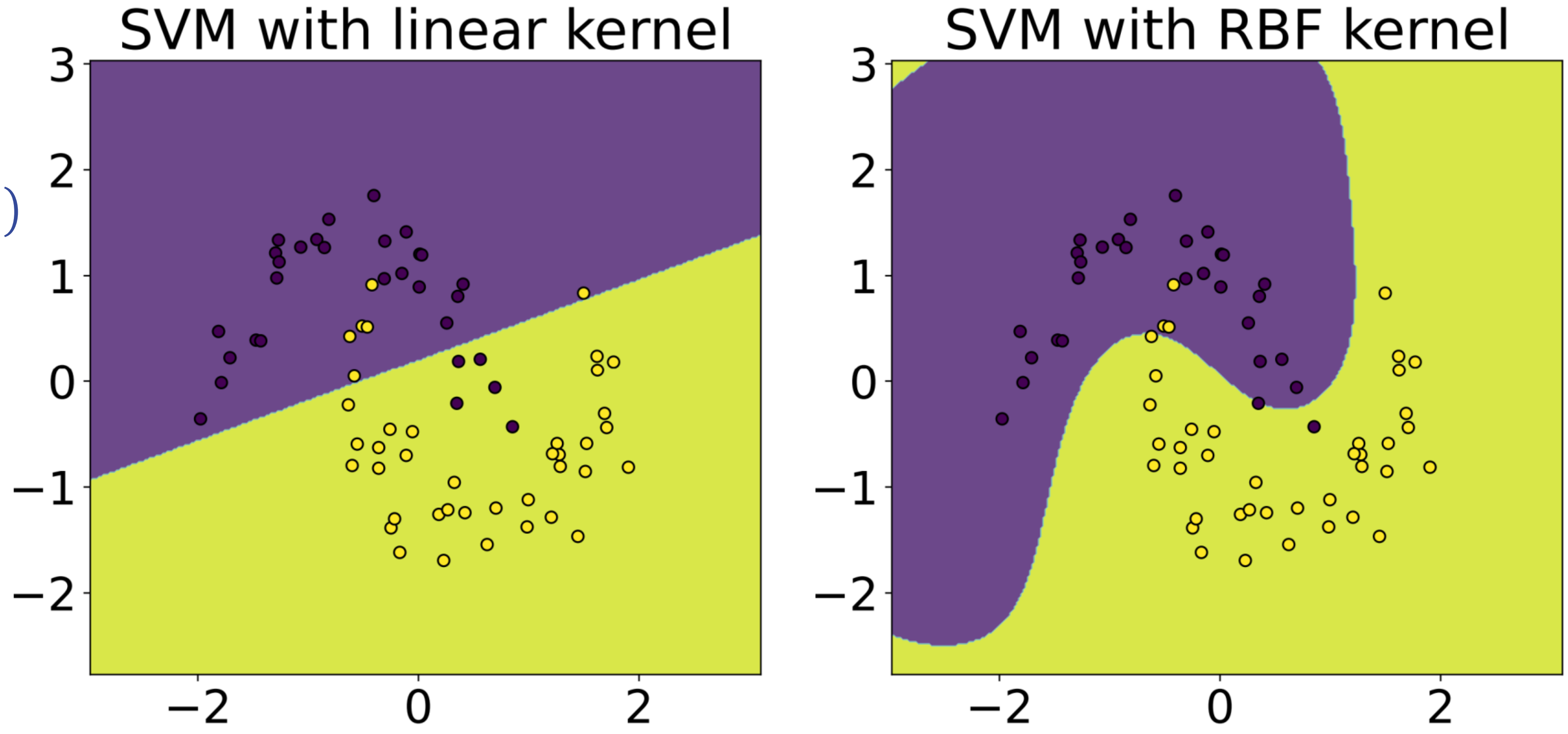
- The kernel trick works by applying a kernel function to the data which, implicitly, finds a higher-dimensional space without having to compute the coordinates of the data in that space.
- We use it to enable the SVM to find a linear separating hyperplane in a higher-dimensional space for data that is not linearly separable.
- Common kernel functions include the polynomial kernel, the radial basis function (RBF) or Gaussian kernel, the sigmoid kernel, and others. Each kernel function has different properties and suits different types of data and problems.

Support Vector Machines

Explain the *kernel trick* and why we use it in Support Vector Machines.

RBF kernel:

$$k(x, x') = exp\left(-\gamma \|x - x'\|^2\right)$$



Support Vector Machines

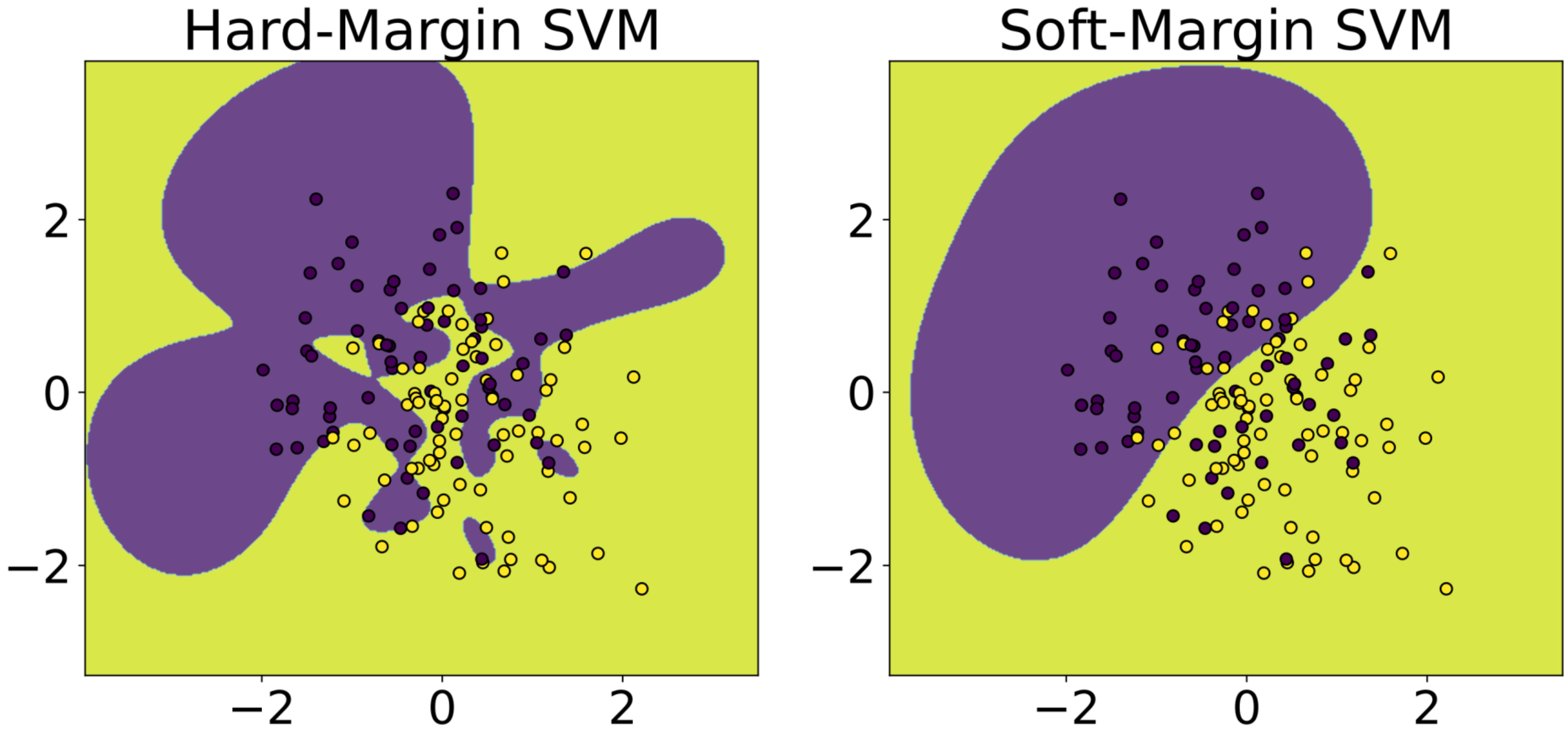
What is the difference between hard- and soft-margin SVM?

	Hard-margin SVM	Soft-Margin SVM
Assumption	Assumes that the data is linearly separable and that there exists a hyperplane that can perfectly separate the classes without any errors.	Assumes that the data may not be perfectly linearly separable, and there may be some degree of overlap or noise in the classes.
Outliers	Very sensitive to outliers because it tries to find a hyperplane that perfectly classifies all training samples.	Less sensitive to outliers and noise because it can tolerate some misclassifications.
Use Case	It is typically used when the data is known to be separable and there is confidence that there are no outliers or noise in the data.	It is used in most real-world scenarios where data is rarely perfectly separable and may contain noise or outliers.

Support Vector Machines

What is the difference between hard- and soft-margin SVM?

Test Accuracy	
Hard-Margin SVM	0.58
Soft-Margin SVM	0.73



Support Vector Machines

The SVM problem can be considered as optimizing the balance between the average hinge loss over the examples and the regularization term that keeps the parameters small (increasing the margin). This balance is set by the regularization term $\lambda > 0$. Here, we focus on the case without the offset parameter w_0(setting it to zero), hence the training objective will become:

$$\frac{1}{n}\sum_{i=1}^{n} L(y_{i,} w^T x_i) + \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2$$

where the hinge loss is given by $L(y_{i,} w^T x_i) = max\left(0, 1 - y_i(w^T x_i)\right)$

Support Vector Machines

The SVM problem: $\frac{1}{n}\sum_{i=1}^{n} L(y_i, w^T x_i) + \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2$

The hinge loss: $L(y_i, w^T x_i) = max\left(0, 1 - y_i(w^T x_i)\right)$

1. **Compute the gradient of the training objective w.r.t.** w_j

$$\nabla_{w_j} L = \nabla_{w_j}\left(\frac{1}{n}\sum_{i=1}^{n} max\left(0, 1 - y_i(w^T x_i)\right) + \frac{\lambda}{2}\sum_{j=1}^{m} w_j^2\right)$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} \nabla_{w_j} max\left(0, 1 - y_i(w^T x_i)\right) + \frac{\lambda}{2}\sum_{j=1}^{m} \nabla_{w_j}(w_j^2)\right) = \left(\frac{1}{n}\sum_{i=1}^{n} \nabla_{w_j} max\left(0, 1 - y_i(w^T x_i)\right) + \frac{\lambda}{2}(2w_j)\right)$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} \nabla_{w_j} max\left(0, 1 - y_i(w^T x_i)\right) + \lambda w_j\right) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} -y_i x_i & if \quad y_i(w^T x_i) < 1 \\ 0 & otherwise \end{cases} + \lambda w_j$$

Support Vector Machines

1. **Describe the pseudocode for the gradient descent algorithm.**

```
Initialize w
Choose learning rate eta
Choose regularization parameter lambda
Repeat until convergence {
    Compute gradient of hinge loss for each example i:
        if y_i * (w^T * x_i) < 1:
            grad_hinge_loss = -y_i * x_i
        else:
            grad_hinge_loss = 0

    Compute gradient of regularization term:
        grad_regularization = lambda * w

    Combine gradients:
        grad_total = (1/n) * sum(grad_hinge_loss) + grad_regularization

    Update weights:
        w = w - eta * grad_total
}
```

Support Vector Machines

$$\nabla_{w_j} L_T = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} -y_i x_i & \text{if} \quad y_i\left(w^T x_i\right) < 1 \\ 0 & \text{otherwise} \end{cases} + \lambda w_j$$

2. Use gradient descent to update the weights based on the following example:

$$\lambda = 0.5, \eta = 0.01, y = 1, x = \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix}, w = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}$$

We first compute the gradient of the hinge loss using the condition:

$y\left(w^T x\right) = 1 \cdot \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}^T \begin{pmatrix} 1.0 \\ 0.0 \end{pmatrix} = 1 \cdot 1 = 1$, so the gradient of the hinge loss is 0.

The gradient of the regularization term is λw_j, so for w_1 and w_2 it is 0.5.

Now, we can update the weights:

$$w^{(1)} = w - \eta \cdot \nabla_w = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} - 0.01 \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix} = \begin{pmatrix} 0.995 \\ 0.995 \end{pmatrix}$$

Linear Separability

Linear Separability

Given the following dataset:

1. **Create a sketch of the data. Is it linearly separable? If so, draw a separating hyperplane.**

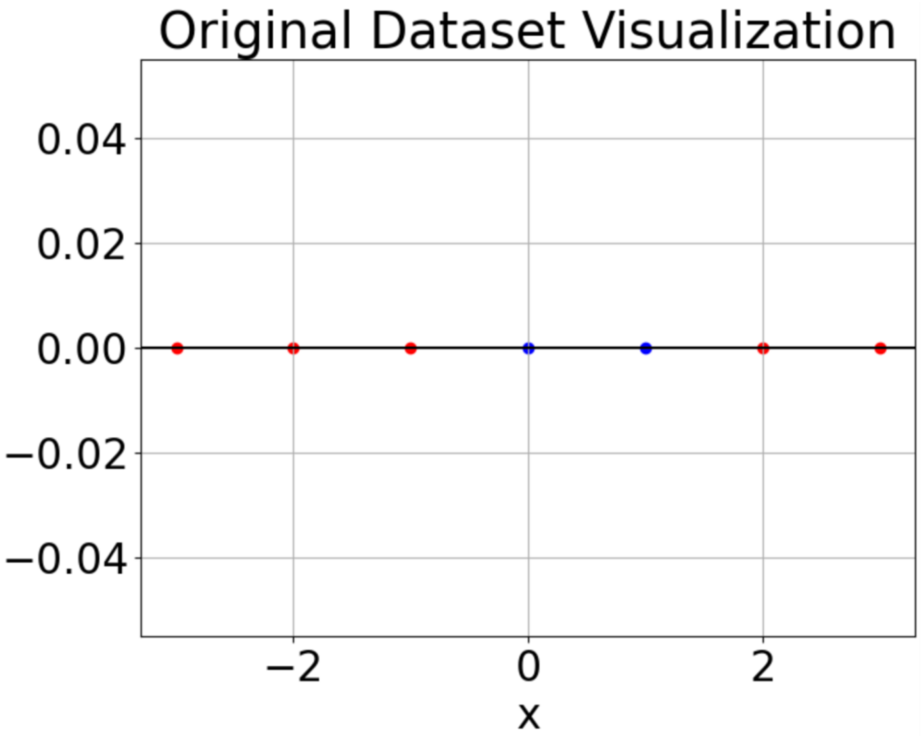
x	y
-3	-1
-2	-1
-1	-1
0	1
1	1
2	-1
3	-1

Linear Separability

Given the following dataset:

1. **Create a sketch of the data. Is it linearly separable? If so, draw a separating hyperplane.**

x	y
-3	-1
-2	-1
-1	-1
0	1
1	1
2	-1
3	-1



Linear Separability

Given the following dataset:

2. Apply the mapping $g: \mathbb{R} \rightarrow \mathbb{R}^2$ defined by: $g(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$ and create a plot of it.

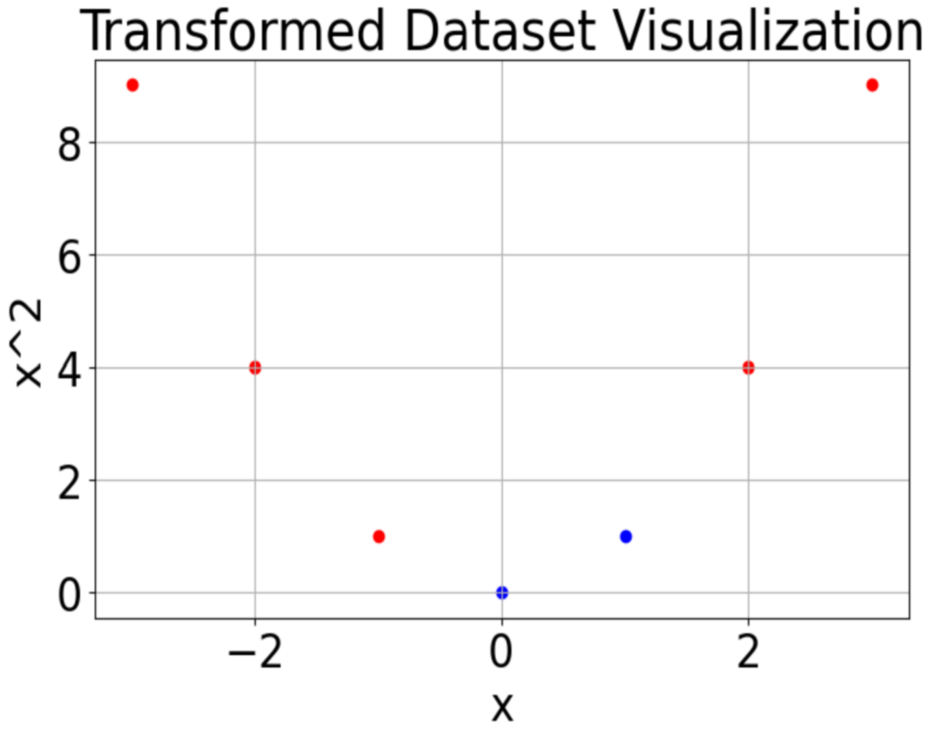
x	x^2	y
-3	9	-1
-2	4	-1
-1	1	-1
0	0	1
1	1	1
2	4	-1
3	9	-1

Linear Separability

Given the following dataset:

2. **Apply the mapping $g: \mathbb{R} \rightarrow \mathbb{R}^2$ defined by: $g(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$ and create a plot of it.**

x	x^2	y
-3	9	-1
-2	4	-1
-1	1	-1
0	0	1
1	1	1
2	4	-1
3	9	-1

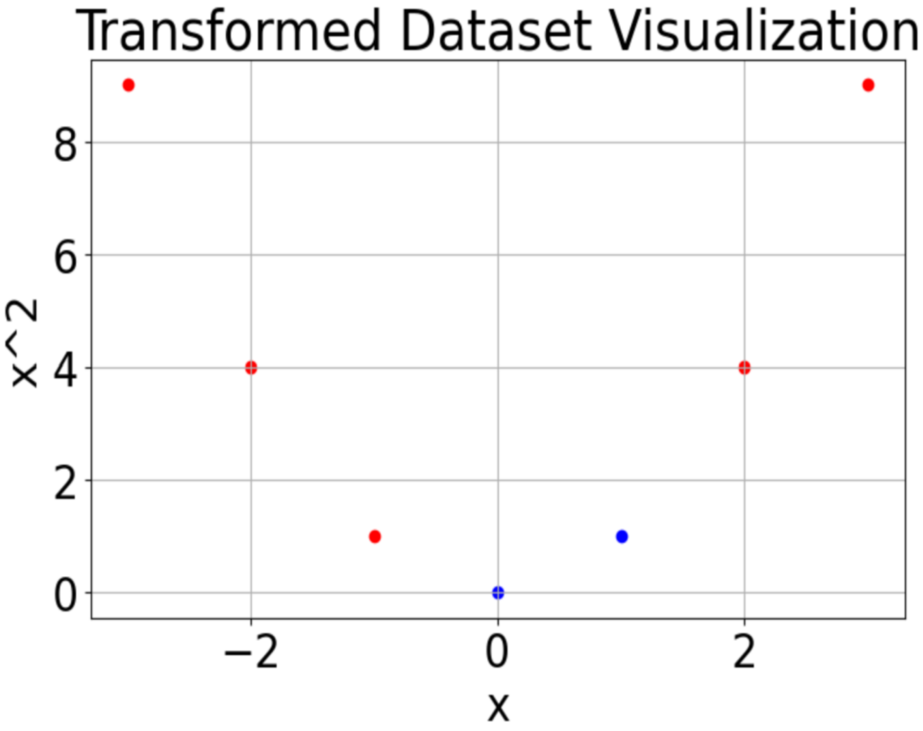


Linear Separability

Given the following dataset:

2. Is the transformed data set linearly separable? If yes, find a separating hyperplane $\langle w, x \rangle + b = 0$ and plot it.

x	x^2	y
-3	9	-1
-2	4	-1
-1	1	-1
0	0	1
1	1	1
2	4	-1
3	9	-1



Linear Separability

Given the following dataset:

2. Is the transformed data set linearly separable? If yes, find a separating hyperplane $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \boldsymbol{b} = \boldsymbol{0}$ and plot it.

We solve the Lagrangian $\boldsymbol{L(w, b, \alpha) = -\sum_{i=1}^{n} \alpha_i \left(y_i\left(w^T x_i + b\right) - 1\right) + \frac{1}{2}\|w\|^2}$ for the support vectors.

- $\nabla_w L = -\sum_{i=1}^{n} \alpha_i y_i \, x_i \; + w = -\sum_{i=3}^{6} \alpha_i y_i \, x_i \; + w = -\alpha_3 y_3 \, x_3 - \alpha_4 y_4 \, x_4 - \alpha_5 y_5 \, x_5 - \alpha_6 y_6 \, x_6 + w$

$$= \alpha_3 x_3 - \alpha_4 x_4 - \alpha_5 x_5 + \alpha_6 x_6 + w = \alpha_3 \begin{pmatrix} -1 \\ 1 \end{pmatrix} - \alpha_4 \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \alpha_5 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \alpha_6 \begin{pmatrix} 2 \\ 4 \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

$$= \begin{pmatrix} -\alpha_3 \\ \alpha_3 \end{pmatrix} - \begin{pmatrix} \alpha_5 \\ \alpha_5 \end{pmatrix} + \begin{pmatrix} 2\alpha_6 \\ 4\alpha_6 \end{pmatrix} + \begin{pmatrix} \boldsymbol{w_1} \\ \boldsymbol{w_2} \end{pmatrix} = \begin{cases} -\alpha_3 - \alpha_5 + 2\alpha_6 + w_1 \\ \alpha_3 - \alpha_5 + 4\alpha_6 + w_2 \end{cases}$$

- $\nabla_b L = -\sum_{i=1}^{n} \alpha_i y_i \; = -\sum_{i=3}^{6} \alpha_i y_i = -\alpha_3 \, y_3 - \alpha_4 y_4 - \alpha_5 y_5 - \alpha_6 y_6 = \alpha_3 - \alpha_4 - \alpha_5 + \alpha_6$

	x	x^2	y
1	-3	9	-1
2	-2	4	-1
3	-1	1	-1
4	0	0	1
5	1	1	1
6	2	4	-1
	3	9	-1

Linear Separability

Given the following dataset:

2. Is the transformed data set linearly separable? If yes, find a separating hyperplane $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \boldsymbol{b} = \mathbf{0}$ and plot it.

We solve the Lagrangian $\boldsymbol{L(w, b, \alpha)} = -\sum_{i=1}^{n} \boldsymbol{\alpha_i (y_i (w^T x_i + b) - 1)} + \frac{1}{2} \|\boldsymbol{w}\|^2$ for the support vectors.

- $\nabla_w L = \begin{cases} -\alpha_3 - \alpha_5 + 2\alpha_6 + w_1 \\ \alpha_3 - \alpha_5 + 4\alpha_6 + w_2 \end{cases}$
- $\nabla_b L = \alpha_3 - \alpha_4 - \alpha_5 + \alpha_6$
- $\nabla_{\alpha_i} L = -y_i\, (w^T x_i + b) + 1$
- $\nabla_{\alpha_3} L = -y_3 (w^T x_3 + b) + 1 = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}^T \begin{pmatrix} -1 \\ 1 \end{pmatrix} + b + 1 = -w_1 + w_2 + b + 1$
- $\nabla_{\alpha_4} L = -y_4 (w^T x_4 + b) + 1 = -\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}^T \begin{pmatrix} 0 \\ 0 \end{pmatrix} - b + 1 = -b + 1$
- $\nabla_{\alpha_5} L = -y_5 (w^T x_5 + b) + 1 = -\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}^T \begin{pmatrix} 1 \\ 1 \end{pmatrix} - b + 1 = -w_1 - w_2 - b + 1$
- $\nabla_{\alpha_6} L = -y_6 (w^T x_6 + b) + 1 = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}^T \begin{pmatrix} 2 \\ 4 \end{pmatrix} + b + 1 = 2w_1 + 4w_2 + b + 1$

	x	x²	y
1	-3	9	-1
2	-2	4	-1
3	-1	1	-1
4	0	0	1
5	1	1	1
6	2	4	-1
	3	9	-1

Linear Separability

Given the following dataset:

2. Is the transformed data set linearly separable? If yes, find a separating hyperplane $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \boldsymbol{b} = \mathbf{0}$ and plot it.

We solve the Lagrangian $\boldsymbol{L}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} \boldsymbol{\alpha}_i (\boldsymbol{y}_i (\boldsymbol{w}^T \boldsymbol{x}_i + \boldsymbol{b}) - \mathbf{1}) + \frac{1}{2} \|\boldsymbol{w}\|^2$ for the support vectors.

- $\nabla_w L = \begin{cases} -\alpha_3 - \alpha_5 + 2\alpha_6 + w_1 \\ \alpha_3 - \alpha_5 + 4\alpha_6 + w_2 \end{cases}$
- $\nabla_b L = \alpha_3 - \alpha_4 - \alpha_5 + \alpha_6$
- $\nabla_{\alpha_3} L = -w_1 + w_2 + b + 1$
- $\nabla_{\alpha_4} L = -b + 1$
- $\nabla_{\alpha_5} L = -w_1 - w_2 - b + 1$
- $\nabla_{\alpha_6} L = 2w_1 + 4w_2 + b + 1$

	x	x^2	y
1	-3	9	-1
2	-2	4	-1
3	-1	1	-1
4	0	0	1
5	1	1	1
6	2	4	-1
	3	9	-1

Linear Separability

Given the following dataset:

2. Is the transformed data set linearly separable? If yes, find a separating hyperplane $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \boldsymbol{b} = \boldsymbol{0}$ and plot it.

We solve the Lagrangian $\boldsymbol{L}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} \boldsymbol{\alpha}_i (\boldsymbol{y}_i (\boldsymbol{w}^T \boldsymbol{x}_i + \boldsymbol{b}) - \boldsymbol{1}) + \frac{1}{2} ||\boldsymbol{w}||^2$ for the support vectors.

- $\nabla_w L = \begin{cases} -\alpha_3 - \alpha_5 + 2\alpha_6 + w_1 \\ \alpha_3 - \alpha_5 + 4\alpha_6 + w_2 \end{cases} = 0$
- $\nabla_b L = \alpha_3 - \alpha_4 - \alpha_5 + \alpha_6 = 0$
- $\nabla_{\alpha_3} L = -w_1 + w_2 + b + 1 = 0$
- $\nabla_{\alpha_4} L = -b + 1 = 0$
- $\nabla_{\alpha_5} L = -w_1 - w_2 - b + 1 = 0$
- $\nabla_{\alpha_6} L = 2w_1 + 4w_2 + b + 1 = 0$

	x	x^2	y
1	-3	9	-1
2	-2	4	-1
3	-1	1	-1
4	0	0	1
5	1	1	1
6	2	4	-1
	3	9	-1

Linear Separability

Given the following dataset:

2. Is the transformed data set linearly separable? If yes, find a separating hyperplane $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \boldsymbol{b} = \boldsymbol{0}$ and plot it.

We solve the Lagrangian $\boldsymbol{L}(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\alpha}) = -\sum_{i=1}^{n} \boldsymbol{\alpha_i}(\boldsymbol{y_i}(\boldsymbol{w^T x_i} + \boldsymbol{b}) - \boldsymbol{1}) + \frac{\boldsymbol{1}}{\boldsymbol{2}}\|\boldsymbol{w}\|^{\boldsymbol{2}}$ for the support vectors.

- $\nabla_w L = \begin{cases} -\alpha_3 - \alpha_5 + 2\alpha_6 + w_1 \\ \alpha_3 - \alpha_5 + 4\alpha_6 + w_2 \end{cases} = 0$
- $\nabla_b L = \alpha_3 - \alpha_4 - \alpha_5 + \alpha_6 = 0$
- $\nabla_{\alpha_3} L = -w_1 + w_2 + b + 1 = 0$
- $\nabla_{\alpha_4} L = -b + 1 = 0$
- $\nabla_{\alpha_5} L = -w_1 - w_2 - b + 1 = 0$
- $\nabla_{\alpha_6} L = 2w_1 + 4w_2 + b + 1 = 0$

From $\nabla_{\alpha_4} L = 0 \rightarrow -b + 1 = 0 \rightarrow \boldsymbol{b = 1}$

$$\nabla_{\alpha_3} L = -w_1 + w_2 + 2 = 0$$
$$\nabla_{\alpha_5} L = -w_1 - w_2 = 0$$
$$\nabla_{\alpha_6} L = 2w_1 + 4w_2 + 2 = 0$$

From $\nabla_{\alpha_5} L = 0 \rightarrow -w_1 - w_2 = 0 \rightarrow w_1 = -w_2$

From $\nabla_{\alpha_3} L = 0 \rightarrow -w_1 + w_2 + 2 = 0 \rightarrow -(-w_2) + w_2 = -2 \rightarrow$
$2w_2 = -2 \rightarrow \boldsymbol{w_2 = -1} \rightarrow \boldsymbol{w_1 = 1}$

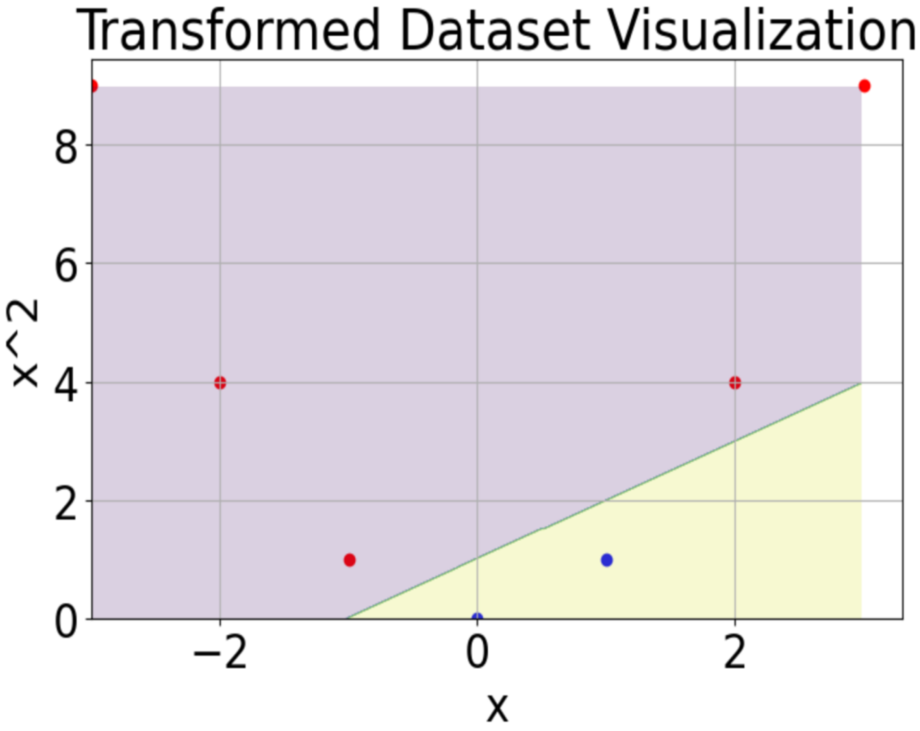
Linear Separability

Given the following dataset:

2. Is the transformed data set linearly separable? If yes, find a separating hyperplane $\langle w, x \rangle + b = 0$ and plot it.

$\boldsymbol{w_1 = 1,} \quad \boldsymbol{w_2 = -1,} \quad \boldsymbol{b = 1,} \quad \boldsymbol{\hat{y} = sign(\langle w, x \rangle + b)}$

	x	x^2	y
1	-3	9	-1
2	-2	4	-1
3	-1	1	-1
4	0	0	1
5	1	1	1
6	2	4	-1
	3	9	-1



Linear Separability

Given the following dataset:

3. Use the computed hyperplane from and compute its prediction for $x = \frac{1+\sqrt{5}}{2}$. Does it belong to the positive or negative class and why?

$$\boldsymbol{w_1 = 1, \qquad w_2 = -1, \qquad b = 1, \qquad \hat{y} = sign(\langle w, x\rangle + b)}$$

$$x = \begin{pmatrix} \frac{1+\sqrt{5}}{2} \\ \left(\frac{1+\sqrt{5}}{2}\right)^2 \end{pmatrix} = \begin{pmatrix} \frac{1+\sqrt{5}}{2} \\ \frac{6+2\sqrt{5}}{4} \end{pmatrix}$$

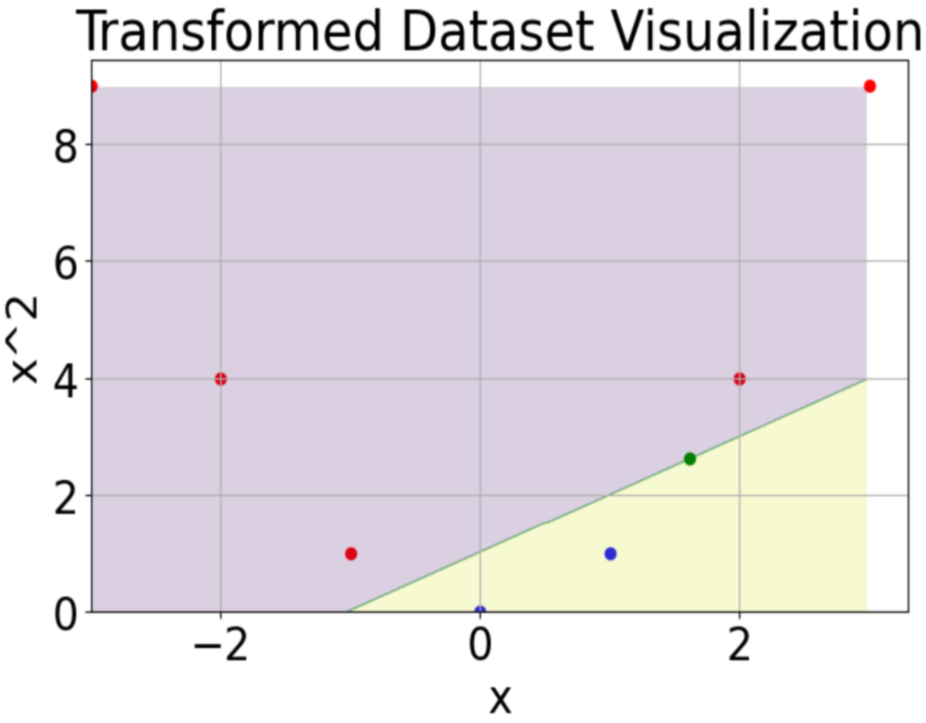
$$\hat{y} = sign(\langle w, x\rangle + b) = sign\left(\left\langle \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} \frac{1+\sqrt{5}}{2} \\ \frac{6+2\sqrt{5}}{4} \end{pmatrix} \right\rangle + 1\right)$$

$$= sign\left(\frac{1+\sqrt{5}}{2} - \frac{6+2\sqrt{5}}{4} + 1\right) = sign\left(\frac{2+2\sqrt{5}-6-2\sqrt{5}}{4} + 1\right) = sign\left(-\frac{4}{4} + 1\right) = sign(0)$$

Linear Separability

Given the following dataset:

3. Use the computed hyperplane from and compute its prediction for $x = \frac{1+\sqrt{5}}{2}$. Does it belong to the positive or negative class and why?



The point belongs on the decision hyperplane. As the *sign* function does not traditionally return a value for zero, these are common ways to handle it:

- Assign to a Default Class
- Report Uncertainty or Ambiguity
- Use a Tie-Breaking Rule