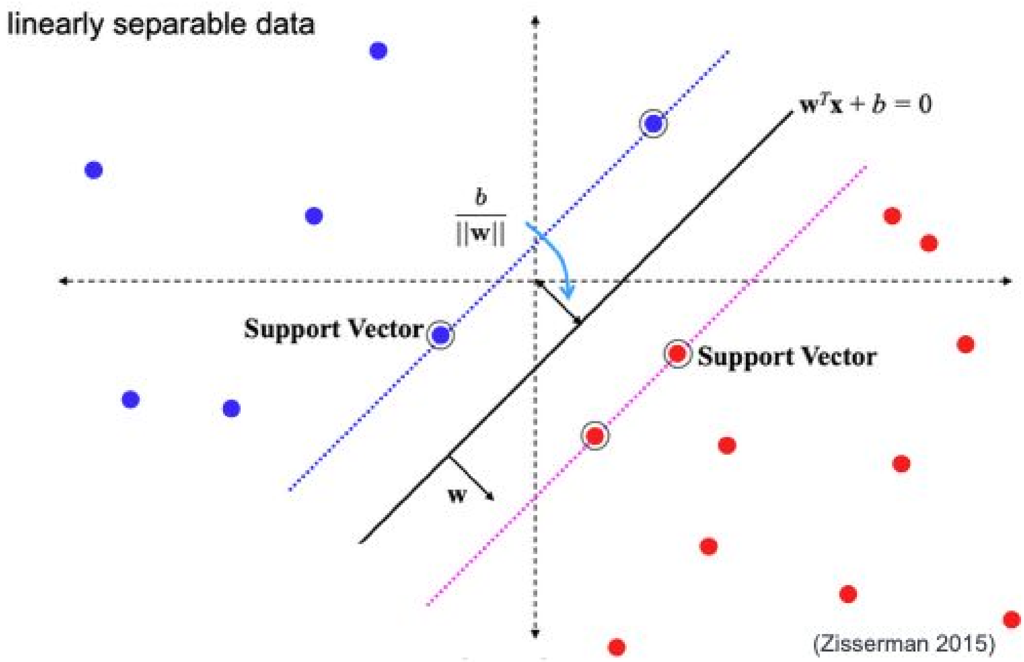
**Assignment 6**
**Support Vector Machines**

Kuang Yu Li, st169971@stud.uni-stuttgart.de, 3440829
Ya Jen Hsu, st169013@stud.uni-stuttgart.de, 3449448
Gabriella Ilena, st169935@stud.uni-stuttgart.de, 3440942

## 1 Concepts

1. Linear separability:
   The idea of SVM is that we try to find a linear boundary that has a maximum margin solution in a classification problem. The term "linear separability" refers to the capability that a certain data set can be separated into different classes with a linear boundary. The support vectors are the data that are closest to the linear boundary.




2. Slack variables:
   Slack variables are introduced to solve the optimization problem of SVM. With slack variables, the optimization constraint can now tolerate both misclassification and margin violation.

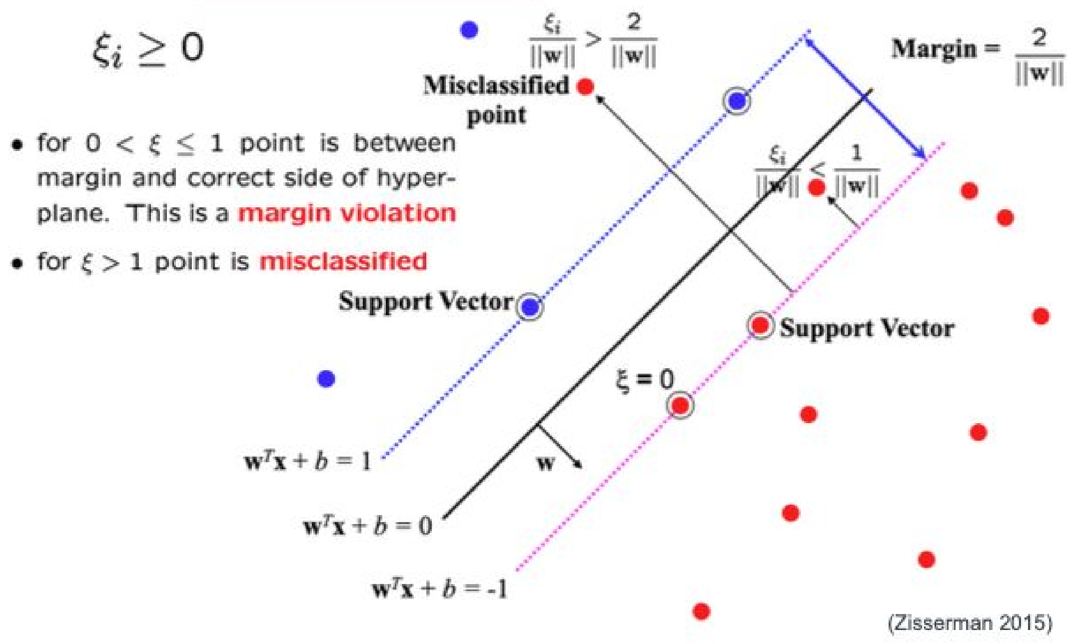
   The constraint becomes:

$$y_i\left( w \cdot x_i + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

   and optimization prblem becomes:

$$\min_{w, \xi_i} \ \|w\|^2 + C\sum_{i=1}^{N} \xi_i$$




3. Kernel functions

   Given a mapping function $\phi: X \to V$, we call the function $K: X \to \mathbb{R}$ defined by $K(x, x') = \langle \phi(x), \phi(x') \rangle_v$, *where* $\langle , \rangle_v$ *denotes an inner product in* $V$, called kernel function. By using kernel functions, we are able to compare the original data with the one transformed in feature space. It is a mathematical way to quantize the similarity.

Reference: Lecture slide chap-6

## 2 Perceptron

1. *Define the classification function for the perceptron classifier.*

   $f(x) = w^T x + b$

   at convergence $w = \sum_{i=1}^{N} \alpha_i x_i$

2. Initial $w = [1 \ -1 \ 0.5]$, $\alpha = 0.6$
   1. for $x_1 = [0 \ 0 \ 1]$, $f(x_1) = 0.5$, $f(x_1)y_1 = 0.5 \cdot (-1) = -0.5 < 0$
      ➔ do $w = w - \alpha x_1 sign(f(x_1)) = [1 \ -1 \ -0.1]$
   2. for $x_2 = [0 \ 1 \ 1]$, $f(x_2) = -1.1$, $f(x_2)y_2 = -1.1 \cdot 1 = -1.1 < 0$
      ➔ do $w = w - \alpha x_2 sign(f(x_2)) = [1 \ -0.4 \ 0.5]$
   3. for $x_3 = [1 \ 0 \ 1]$, $f(x_3) = 1.5$, $f(x_3)y_3 = 1.5 \cdot 1 = 1.5 > 0$
      ➔ do nothing
   4. for $x_4 = [1 \ 1 \ 1]$, $f(x_4) = 1.1$, $f(x_4)y_4 = 1.1 \cdot 1 = 1.1 > 0$
      ➔ do nothing

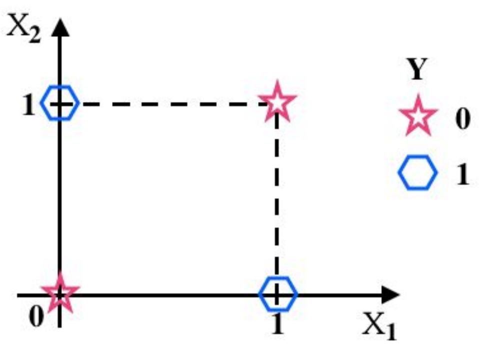
5. for $x_1 = [0\ \ 0\ \ 1]$, $\quad f(x_1) = 0.5$, $\quad f(x_1)y_1 = 0.5 \cdot (-1) =-0.5 < 0$
   ➔ do $w = w - \alpha x_1 sign(f(x_1)) = [1\ \ -0.4\ \ -0.1]$
6. for $x_2 = [0\ \ 1\ \ 1]$, $\quad f(x_2) =-0.5$, $\quad f(x_2)y_2 =-0.5 \cdot 1 =-0.5 < 0$
   ➔ do $w = w - \alpha x_1 sign(f(x_1)) = [1\ \ 0.2\ \ 0.5]$
7. for $x_3 = [1\ \ 0\ \ 1]$, $\quad f(x_3) = 1.5$, $\quad f(x_3)y_3 = 1.5 \cdot 1 = 1.5 > 0$
   ➔ do nothing
8. for $x_4 = [1\ \ 1\ \ 1]$, $\quad f(x_4) = 1.7$, $\quad f(x_4)y_4 = 1.7 \cdot 1 = 1.7 > 0$
   ➔ do nothing
9. for $x_1 = [0\ \ 0\ \ 1]$, $\quad f(x_1) = 0.5$, $\quad f(x_1)y_1 = 0.5 \cdot (-1) =-0.5 < 0$
   ➔ do $w = w - \alpha x_1 sign(f(x_1)) = [1\ \ 0.2\ \ -0.1]$
10. for $x_2 = [0\ \ 1\ \ 1]$, $\quad f(x_2) = 0.1$, $\quad f(x_2)y_2 = 0.1 \cdot 1 = 0.1 > 0$
    ➔ do nothing
11. for $x_3 = [1\ \ 0\ \ 1]$, $\quad f(x_3) = 0.9$, $\quad f(x_3)y_3 = 0.9 \cdot 1 = 0.9 > 0$
    ➔ do nothing
12. for $x_4 = [1\ \ 1\ \ 1]$, $\quad f(x_4) = 1.1$, $\quad f(x_4)y_4 = 1.1 \cdot 1 = 1.1 > 0$
    ➔ do nothing
13. for $x_1 = [0\ \ 0\ \ 1]$, $\quad f(x_1) =-0.1$, $\quad f(x_1)y_1 =-0.1 \cdot (-1) = 0.1 > 0$
    ➔ do nothing

=> convergence

3. *Prove that the XOR function cannot be represented by a (linear) perceptron.*

The data set for the XOR function is given by:

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |




The 0 and 1 points cannot be separated by a linear line, or effectively, there does not exist a linear line that can separate the 0 and 1 points. Therefore, these cases are not linearly separable and thus cannot be represented by a linear perceptron.

## 3 Polynomial Kernel

Given the second-order polynomial kernel for a two-dimensional feature vector $x_i = [x_{i1}, x_{i2}]^T$:

$$\phi(x_i) = \left[x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2\right]$$
$$\phi(x_j) = \left[x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2\right]$$

The scalar product of the kernel can then be calculated as follows:

$$\begin{aligned}
\left\langle \phi(\underline{x_i}), \phi(\underline{x_j}) \right\rangle &= \left[x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2\right]^T \cdot \left[x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2\right] \\
&= x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{j1}x_{i2}x_{j2} + x_{i2}^2 x_{j2}^2 \\
&= \left(x_{i1}x_{j1} + x_{i2}x_{j2}\right)^2 \\
&= \left\langle \underline{x_i}, \underline{x_j} \right\rangle^2
\end{aligned}$$

Hence, the scalar product of the kernel is just the scalar product of the two-dimensional feature vectors, so a mapping from two-dimensional to three-dimensional vector is not necessary.

## 4 Gaussian Kernel

The Gaussian or the radial basis function (RBF) kernel is given as: $e^{\frac{\|x-x'\|^2}{2\sigma^2}}$. We can rewrite the equation of this kernel as:

$$\begin{aligned}
exp\left(\frac{\|x-x'\|^2}{2\sigma^2}\right) &= exp\left(\frac{\left\langle x-x', x-x' \right\rangle}{2\sigma^2}\right) \\
&= exp\left(\frac{\left\langle (x, x-x') - (x', x-x') \right\rangle}{2\sigma^2}\right) \quad \rightarrow \text{using the distributive over addition property of scalar product} \\
&= exp\left(\frac{\left\langle x, x \right\rangle - \left\langle x, x' \right\rangle - \left\langle x', x \right\rangle + \left\langle x', x' \right\rangle}{2\sigma^2}\right) \\
&= exp\left(\frac{\left\langle x, x \right\rangle - \left\langle x, x' \right\rangle - \left\langle x', x \right\rangle + \left\langle x', x' \right\rangle}{2\sigma^2}\right) \\
&= exp\left(\frac{\|x\|^2 + \|x'\|^2 - 2\left\langle x, x' \right\rangle}{2\sigma^2}\right)
\end{aligned}$$

$$= exp\left(\frac{\|x\|^2+\|x'\|^2}{2\sigma^2}\right)exp\left(\frac{-2\langle x,x'\rangle}{2\sigma^2}\right)$$

$$= const \cdot exp\left(\frac{-\langle x,x'\rangle}{\sigma^2}\right)$$

If we take the Taylor expansion of the equation above,

$$= \ const \cdot \sum_{n=0}^{\infty} \frac{\left(\frac{-\langle x,x'\rangle}{2\sigma^2}\right)^n}{n!}$$

$$= \ const \cdot \sum_{n=0}^{\infty} \left(\frac{-1}{2\sigma^2}\right)^n \frac{(\langle x,x'\rangle)^n}{n!}$$

we can see that the Gaussian kernel consists of an infinite sum of the polynomial kernel.