Exercise 06

Support Vector Machines

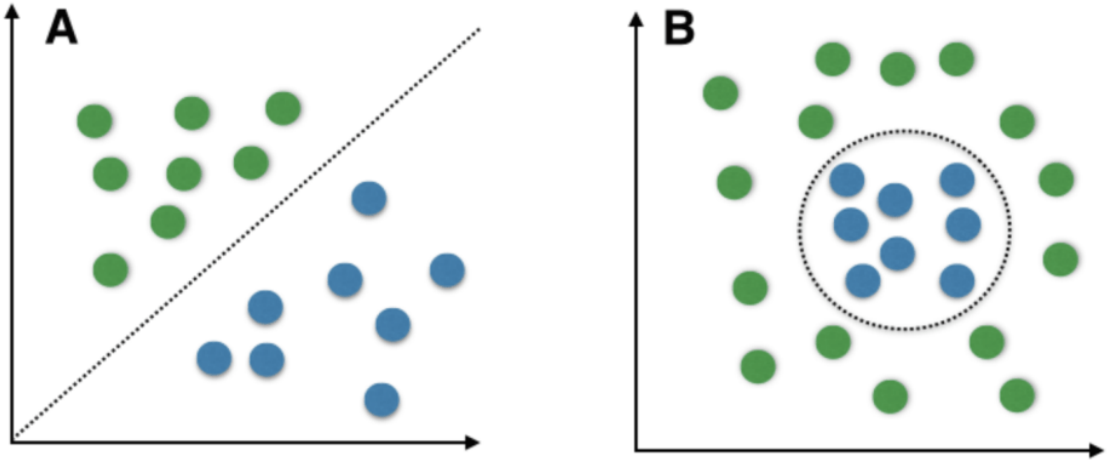
Alexander Tiessen 2965198

Patrick Schneefuss 2951267

Jonas Lammert 3149269

We have not yet made a video for any task of a previous assignment, so we would like present a task from this assignment 06.

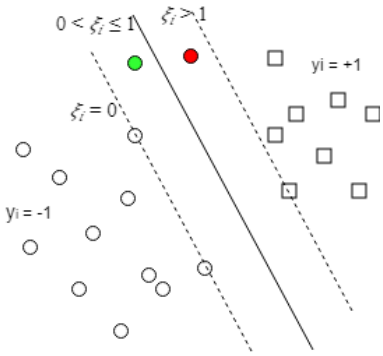
Two sets in the euclidean space are linearly separable if a hyperplane exists which separates these sets. The SVM requires the problem to be linearly separable in order to create a hyperplane which can be used for classification. If a given problem is not linear separable the SVM uses the so called “kernel trick”. Here, the data is transformed in a higher dimension in a way that the problem is linear separable. Afterwards, the Euclidean Space is transformed back into the original dimensionality, including the hyperplane which is probably not a plane (not linear) anymore. There are many different kernel functions for the mentioned space transformation. The selection of an appropriate function is dependent on the problem.



linear separable not linear separable

https://i0.wp.com/www.tarekatwan.com/wp-content/uploads/2017/12/linear_sep.png?resize=1000%2C409

Slack variables are tool to allow for outliers in the trainings data. This way, there are points within the margin created by the support vectors and the separating plane (green and red dot).



(https://nianlonggu.github.io/img/2019-06-07-SVM/svm-slack-variable.svg)

Task 2 Perceptron

1) The classification function for the perceptron function is given by

$\hat{f}(x) = w^T x + b$

with x being training data $\in \mathbb{R}^d$ and w a weightvector assigning a weight to each variable of x

We can rewrite $\hat{f}(x)$ to

$\hat{f}(x) = w^T x$

by prepending the bias b to our weightvector (or appending) and adding a 1 at the respective position in x

For binary classification ($y \in \{-1, 1\}$) data is then classified as:

$$y_i = \begin{cases} 1 & \text{if } \hat{f}(x_i) > 0 \\ -1 & \text{if } \hat{f}(x_i) < 0 \end{cases}$$

---

2) $w_{init} = [1 \ -1 \ 0{,}5]$ with 0,5 being b

$\hat{f}(x) = (1 \ -1 \ 0{,}5) \cdot x$

$\hat{f}(x_1) = (1 \ -1 \ 0{,}5) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 0{,}5$; $0{,}5 \cdot (-1) = -0{,}5$ $(\hat{f}(x_i) y_i)$

→ wrongly classified

$w_{new} = w_{old} - 0{,}6 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot \text{sign}(\hat{f}(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}))$

$= \begin{pmatrix} 1 \\ -1 \\ 0{,}5 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0{,}6 \end{pmatrix} \cdot 1 = \begin{pmatrix} 1 \\ -1 \\ -0{,}1 \end{pmatrix}$

$\hat{f}(x_1) = (1 \ -1 \ -0{,}1) \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = -0{,}1$; $-0{,}1 \cdot (-1) = 0{,}1$ ✓

$\hat{f}(x_2) = (1 \ -1 \ -0{,}1) \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = -1{,}1$; $-1{,}1 \cdot 1 = -1{,}1$

→ wrongly classified!

$w_{new} = \begin{pmatrix} 1 \\ -1 \\ -0{,}1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0{,}6 \\ 0{,}6 \end{pmatrix} \cdot (-1) = \begin{pmatrix} 1 \\ -0{,}4 \\ 0{,}5 \end{pmatrix}$

$\hat{f}(x_1) = \begin{pmatrix} 1 & -0,4 & 0,5 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 0,5 \; ; \; 0,5 \cdot (-1) = -0,5$

→ wrongly classfied

$w_{new} = \begin{pmatrix} 1 \\ -0,4 \\ 0,5 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0,6 \end{pmatrix} \cdot 1 = \begin{pmatrix} 1 \\ -0,4 \\ -0,1 \end{pmatrix}$

$\hat{f}(x_1) = \begin{pmatrix} 1 & -0,4 & -0,1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = -0,1 \; ; \; -0,1 \cdot (-1) = 0,1$ ✓

$\hat{f}(x_2) = \begin{pmatrix} 1 & -0,4 & -0,1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = -0,5 \; ; \; -0,5 \cdot 1 = -0,5$

→ wrongly classfied

$w_{new} = \begin{pmatrix} 1 \\ -0,4 \\ -0,1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0,6 \\ 0,6 \end{pmatrix} \cdot (-1) = \begin{pmatrix} 1 \\ 0,2 \\ 0,5 \end{pmatrix}$

$\hat{f}(x_1) = \begin{pmatrix} 1 & 0,2 & 0,5 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = 0,5 \; ; \; 0,5 \cdot (-1) = -0,5$

→ wrongly classfied

$w_{new} = \begin{pmatrix} 1 \\ 0,2 \\ 0,5 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0,6 \end{pmatrix} \cdot 1 = \begin{pmatrix} 1 \\ 0,2 \\ -0,1 \end{pmatrix}$

$\hat{f}(x_1) = \begin{pmatrix} 1 & 0,2 & -0,1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = -0,1 \; ; \; -0,1 \cdot (-1) = 0,1$ ✓

$\hat{f}(x_2) = \begin{pmatrix} 1 & 0,2 & -0,1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} = 0,1 \; ; \; 0,1 \cdot 1 = 0,1$ ✓

$\hat{f}(x_3) = \begin{pmatrix} 1 & 0,2 & -0,1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} = 0,9 \; ; \; 0,9 \cdot 1 = 0,9$ ✓

$\hat{f}(x_4) = \begin{pmatrix} 1 & 0,2 & -0,1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 1,1 \; ; \; 1,1 \cdot 1 = 1,1$ ✓

⇒ no more wrong classifications

$w_{final} = \begin{bmatrix} 1 & 0,2 & -0,1 \end{bmatrix}$

3) Assume there is a linear perceptron that classifies each x_i of the XOR function correctly. That classifier is of form

$\hat{f}(x_i) = w^T x_i$ (Let w_3 be bias b)

This yields the following 4 inequalities (one for each datapoint x_1 to x_4). (Remember $y = [-1\ 1\ 1\ -1]^T$)

$$w^T x_1 = w_3 \cdot 1 < 0 \quad (1)$$

$$w^T x_2 = w_2 + w_3 > 0 \quad (2)$$

$$w^T x_3 = w_1 + w_3 > 0 \quad (3)$$

$$w^T x_4 = w_1 + w_2 + w_3 < 0 \quad (4)$$

Now obviously (1) gives us $w_3 < 0$, plugging this in (2) and (3) gives $w_2 > 0$ and $w_1 > 0$ and also $w_2 > (-w_3)$ and $w_1 > (-w_3)$.

This means, that $w_1 + w_2 + w_3 < 0$ can not hold!

Such classifier for the XOR function can't exist!

Task 3 Polynomial Kernel

$$\langle \phi(x_i), \phi(x_j) \rangle = x_{i1}^2 x_{j1}^2 + \sqrt{2} x_{i1} x_{i2} \sqrt{2} x_{j1} x_{j2} + x_{i2}^2 x_{j2}^2$$

$$= x_{i1}^2 x_{j1}^2 + 2 x_{i1} x_{i2} x_{j1} x_{j2} + x_{i2}^2 x_{j2}^2$$

$$= (x_{i1} x_{j1})^2 + 2 \cdot (x_{i1} x_{j1})(x_{i2} x_{j2}) + (x_{i2} x_{j2})^2$$

$$= (x_{i1} x_{j1} + x_{i2} x_{j2})^2$$

$$= \langle x_i, x_j \rangle^2$$

Task 4

Basically, RBF Kernel-value can be interpreted as similarity value. Similarity will be $\in [0,1]$ where 1 denotes maximum similarity. Having m datapoints, there are m similarity values for this point, one for each other datapoint plus the similarity to itsself. Since we can have an infinite number of points, the resulting feature space can also be infinite.

Proof:

For simplicity, we choose σ such that $2\sigma^2 = 1$

then: $k(x, x') = e^{-\|x - x'\|^2}$

Let x and x' be $\in \mathbb{R}^2$ with $x = (x_1, x_2)$ and $x' = (x_1', x_2')$

$$k(x, x') = e^{-((x_1 - x_1')^2 + (x_2 - x_2')^2)}$$

$$= \exp\left(-\left(x_1^2 - 2x_1x_1' + x_1'^2 + x_2^2 - 2x_2x_2' + x_2'^2\right)\right)$$

$$= \exp\left(-x_1^2 + 2x_1x_1' - x_1'^2 - x_2^2 + 2x_2x_2' - x_2'^2\right)$$

$$= \exp\left(-\|x\|^2\right) \cdot \exp\left(-\|x'\|^2\right) \cdot \exp\left(2x_1x_1' + 2x_2x_2'\right)$$

$$= \exp\left(-\|x\|^2\right) \cdot \exp\left(-\|x'\|^2\right) \cdot \exp\left(2x^Tx'\right)$$

Now $2x^Tx'$ already looks like a polynomial kernel of degree 1 we use taylor expansion on $e^{2x^Tx'}$

$$= \exp\left(-\|x\|^2\right) \exp\left(-\|x'\|^2\right) \cdot \sum_{n=0}^{\infty} \frac{(2x^Tx)^n}{n!}$$

$\sum_{n=0}^{\infty} \frac{(2x^T x)^n}{n!}$ means that we are adding

infinitely many polynomial kernels with increasing (up to infinite) degree.
When adding two polynomial kernels, the resulting feature space has the sum of the ~~feature spaces~~ added kernels' feature spaces as dimension, thus the RBF kernel has an infinite feature space.