University of Stuttgart | ANALYTIC COMPUTING

Machine Learning Exercise (SS 21)

Assignment 6: Support Vector Machines (SVM) (Solution)

Dr. Decky Aspandi
decky.aspandi-latif@ipvs.uni-stuttgart.de

Akram Hosseini
Akram.Hosseini@ipvs.uni-stuttgart.de

Daniel Frank
daniel.frank@ipvs.uni-stuttgart.de

This assignment sheet consists of 6 pages with the following 4 tasks:

- Task 1: Support Vector Machines (SVM) Concepts (30 Points) 2
- Task 2: Perceptron (40 Points) 3
- Task 3: Polynomial Kernel (15 Points) 5
- Task 4: Gaussian Kernel (15 Points) 6

Submit your solution in ILIAS as a single PDF file.[1] Make sure to list full names of all participants, matriculation number, study program and B.Sc. or M.Sc on the first page. Optionally, you can *additionally* upload source files (e.g. PPTX files). If you have any questions, feel free to ask them in the exercise forum in ILIAS.

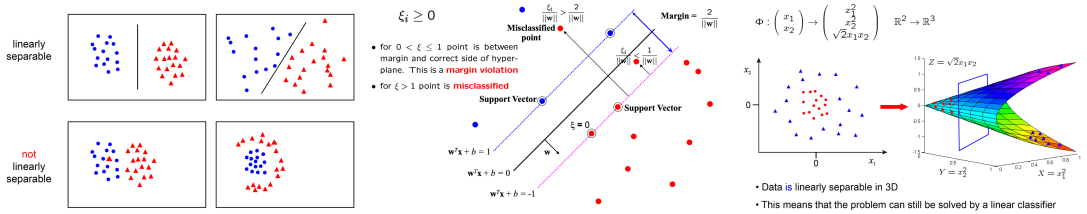
Submission is open until Tuesday, 15th June 2021, 23:59.

[1] Your drawing software probably allows to export as PDF. An alternative option is to use a PDF printer. If you create multiple PDF files, use a merging tool (like pdfarranger) to combine the PDFs into a single file.

Task 1: Support Vector Machines (SVM) Concepts (30 Points)

Explain the following terms and how they are related to SVM in your own words and with (visual) examples:

1. **Task (10 Points)**: Linear separability.
   Solution: A dataset is linearly separable if a linear function (also called discriminant) $\hat{f}(x)$ exists (linear in features) that separates the (two) classes of the data set, i.e. all data points of the one class are on one side (e.g. $\hat{f}(x) < 0$) of the line/plane/hyperplane spanned by $\hat{f}(x) = 0$ while all the data points belonging to the other class are on the other side (e.g. $\hat{f}(x) > 0$) of the line/plane/hyperplane.
2. **Task (10 Points)**: Slack variables.
   Solution: Slack variables ξ_i soften the objective up by allowing small violations of constraints (e.g. missclassifications or margin violations where a point is on the correct side of the hyperplane but between the margin and the hyperplane). This allows to find solutions for not linearly separable data sets.
3. **Task (10 Points)**: Kernel functions.
   Solution: Kernel functions measure the similarity of two data points x and x' (basically comparing them) by expressing how correlated their respective function outputs y and y' are. They help us to work directly on the data points without using the feature space first. Each kernel corresponds to a specific feature choice (although not always that obvious) and vice versa as follows: $k(x, x') = \phi(x)^T \phi(x')$.



(a) Linear separability (Zisserman 2015)

(b) Slack variables (Zisserman 2015)

(c) Kernel functions (Zisserman 2015)

Figure 1 Solution for Task 1.

Task 2: Perceptron (40 Points)

1. **Task (10 Points)**: Define the classification function for the perceptron classifier.

   Solution: $\hat{f}(x) = \begin{cases} 1 & \text{if } w^\top x + b > 0 \\ -1 & \text{if } w^\top x + b < 0 \end{cases}$

2. **Task (20 Points)**: The dataset for the OR function is given by:

$$X = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

$$y = \begin{bmatrix} -1 & 1 & 1 & 1 \end{bmatrix}^T \; op$$

   Given the initial weights of $w = \begin{bmatrix} 1 & -1 & 0.5 \end{bmatrix}$, where w_3 is the bias. Perform the perceptron algorithm (slide 10) with $\alpha = 0.6$ until all data points are correctly classified. Show your computations for each training step. (Note: In the case of $w \cdot x = 0$ output 1.).

   Solution:

   $\odot$ denotes the point-wise multiplication. Append a one for every data point to include the bias:

$$X = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

   1. iteration:

   $\hat{y} = \hat{f}(X) = Xw = \begin{bmatrix} 0.5 & -0.5 & 1.5 & 0.5 \end{bmatrix}^\top$, $\hat{y} \odot y = \begin{bmatrix} -0.5 & -0.5 & 1.5 & 0.5 \end{bmatrix}^\top$

   Fix first data point: $w_{new} = w_{old} - 0.6\, x_1\, \text{sign}(\hat{f}(x_1)) = \begin{bmatrix} 1 & -1 & -0.1 \end{bmatrix}^\top$

   2. iteration:

   $\hat{y} = \hat{f}(X) = Xw = \begin{bmatrix} -0.1 & -1.1 & 0.9 & -0.1 \end{bmatrix}^\top$, $\hat{y} \odot y = \begin{bmatrix} 0.1 & -1.1 & 0.9 & -0.1 \end{bmatrix}^\top$

   Fix second data point: $w_{new} = w_{old} - 0.6\, x_2\, \text{sign}(\hat{f}(x_2)) = \begin{bmatrix} 1 & -0.4 & 0.5 \end{bmatrix}^\top$

   3. iteration:

   $\hat{y} = \hat{f}(X) = Xw = \begin{bmatrix} 0.5 & 0.1 & 1.5 & 1.1 \end{bmatrix}^\top$, $\hat{y} \odot y = \begin{bmatrix} -0.5 & 0.1 & 1.5 & 1.1 \end{bmatrix}^\top$

   Fix first data point: $w_{new} = w_{old} - 0.6\, x_1\, \text{sign}(\hat{f}(x_1)) = \begin{bmatrix} 1 & -0.4 & -0.1 \end{bmatrix}^\top$

   4. iteration:

   $\hat{y} = \hat{f}(X) = Xw = \begin{bmatrix} -0.1 & -0.5 & 0.9 & 0.5 \end{bmatrix}^\top$, $\hat{y} \odot y = \begin{bmatrix} 0.1 & -0.5 & 0.9 & 0.5 \end{bmatrix}^\top$

   Fix second data point: $w_{new} = w_{old} - 0.6\, x_2\, \text{sign}(\hat{f}(x_2)) = \begin{bmatrix} 1 & 0.2 & 0.5 \end{bmatrix}^\top$

   5. iteration:

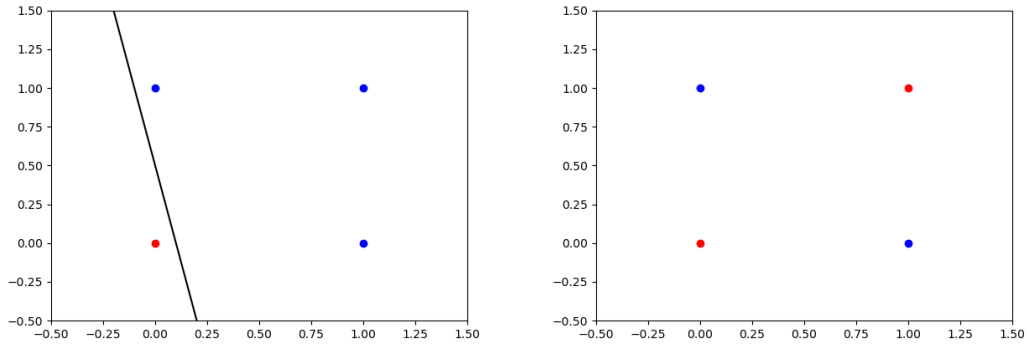
$\hat{y} = \hat{f}(X) = Xw = \begin{bmatrix} 0.5 & 0.7 & 1.5 & 1.7 \end{bmatrix}^\top$, $\hat{y} \odot y = \begin{bmatrix} -0.5 & 0.7 & 1.5 & 1.7 \end{bmatrix}^\top$

Fix first data point: $w_{new} = w_{old} - 0.6\ x_1\ \text{sign}(\hat{f}(x_1)) = \begin{bmatrix} 1 & 0.2 & -0.1 \end{bmatrix}^\top$

6. iteration:

$\hat{y} = \hat{f}(X) = Xw = \begin{bmatrix} -0.1 & 0.1 & 0.9 & 1.1 \end{bmatrix}^\top$, $\hat{y} \odot y = \begin{bmatrix} 0.1 & 0.1 & 0.9 & 1.1 \end{bmatrix}^\top$

We are done since $\forall x_i : \hat{y}_i \odot y_i > 0$. Final $w = \begin{bmatrix} 1 & 0.2 & -0.1 \end{bmatrix}^\top$.



(a) Resulting discriminant function for OR. (Task 2.2)

(b) No linear separability for XOR. (Task 2.3)

Figure 2 Solution for Task 2.

3. **Task (10 Points)**:Prove that the XOR function cannot be represented by a (linear) perceptron.

   Solution: As can be seen in Figure 2b, this data set is not linearly separable.

Task 3: Polynomial Kernel (15 Points)

Task (15 Points): The second-order polynomial kernel for a two-dimensional vector $x_i = \begin{bmatrix} x_{i1} & x_{i2} \end{bmatrix}^T$ *op* is defined as:

$$\phi(x_i) = \begin{bmatrix} x_{i1}^2 \\ \sqrt{2}x_{i1}x_{i2} \\ x_{i2}^2 \end{bmatrix}$$

Show that the mapping of the two-dimensional vector to three dimensions is not necessary for calculating the scalar product $\langle \phi(x_i), \phi(x_j) \rangle$. This implies that the mapping to higher dimension is implicitly performed when calculating the results of kernel functions. (Hint: Transform the equation such that it only uses the scalar product of two-dimensional vectors.).

Solution:

$$\phi(x_i)^T \phi(x_j) = \begin{bmatrix} x_{i1}^2 \\ \sqrt{2}x_{i1}x_{i2} \\ x_{i2}^2 \end{bmatrix}^T \begin{bmatrix} x_{j1}^2 \\ \sqrt{2}x_{j1}x_{j2} \\ x_{j2}^2 \end{bmatrix} = x_{i1}^2 x_{j1}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + x_{i2}^2 x_{j2}^2$$
$$= (x_{i1}x_{j1} + x_{i2}x_{j2})^2 = (x_i^T x_j)^2$$

As you can see, we can directly use the two data points without mapping them to three dimensions beforehand.

Task 4: Gaussian Kernel (15 Points)

Only for all students other than B.Sc. Data Science.

Task (15 Points): Slide 69th in the theory section mentions that the Gaussian kernel, also called Radial Basis Function (RBF), projects to an infinite dimensional feature space. Give an intuition on why this is the case and prove it. (Note: Use the Taylor expansion over e^x to show that the Gaussian kernel is an infinite sum over the polynomial kernels.).

Solution: Definition of Gaussian kernel: $k(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}$

$$\begin{aligned} k(x, x') &= e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \\ &= e^{-\frac{(x-x')^T(x-x')}{2\sigma^2}} \\ &= e^{-\frac{x^Tx-2x^Tx'+x'^Tx'}{2\sigma^2}} \\ &= e^{-\frac{\|x\|^2+\|x'\|^2}{2\sigma^2}+\frac{2x^Tx'}{2\sigma^2}} \\ &= e^{-\frac{\|x\|^2+\|x'\|^2}{2\sigma^2}}\ e^{\frac{x^Tx'}{\sigma^2}} \end{aligned}$$

We fix the first part as a constant ($c := e^{-\frac{\|x\|^2+\|x'\|^2}{2\sigma^2}}$) while we use the Taylor expansion over e^x ($e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$) for the second part:

$$\begin{aligned} k(x, x') &= c\ e^{\frac{x^Tx'}{\sigma^2}} \\ &= c \sum_{n=0}^{\infty} \frac{(\frac{x^Tx'}{\sigma^2})^n}{n!} \\ &= c \sum_{n=0}^{\infty} \frac{(x^Tx')^n}{\sigma^{2n}n!} \end{aligned}$$

The numerator $(x^Tx')^n$ is the polynomial kernel. Therefore, the Gaussian kernel consists of an infinite sum over polynomial kernels of x and x', leading to an infinite dimensional feature space.