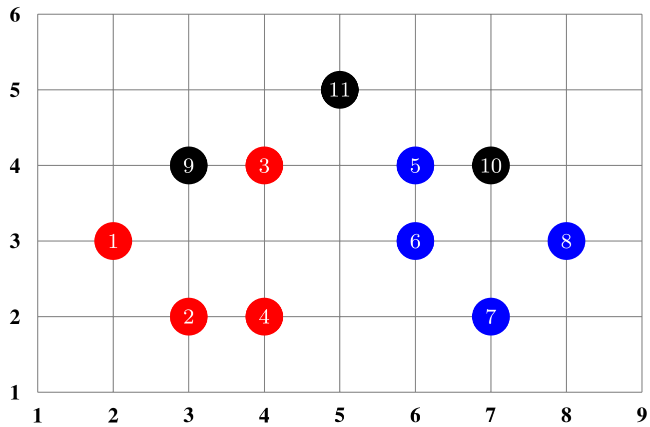
Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Max Berrendorf, Julian Busch

Knowledge Discovery and Data Mining I

WS 2018/19

Exercise 11: SVM, Kernel Trick, Linear Separability

Exercise 11-1 Support Vector Machines



Consider the following training data:

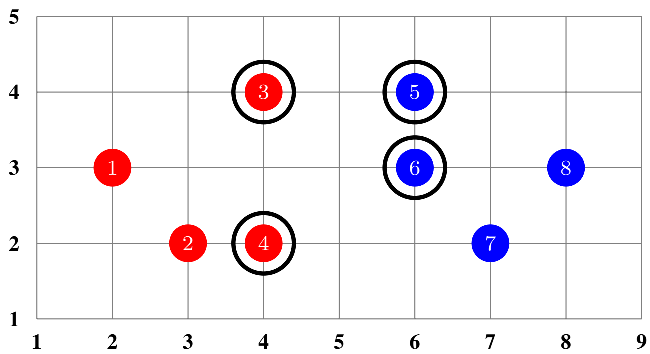
$$x_1 = (2,3), x_2 = (3,2), x_3 = (4,4), x_4 = (4,2)$$

$$x_5 = (6,4), x_6 = (6,3), x_7 = (7,2), x_8 = (8,3)$$

Let $y_A = -1, y_B = +1$ be the class indicators for both classes

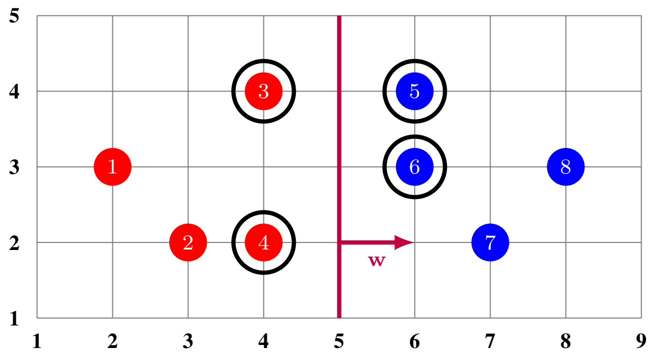
$$A = \{x_1, x_2, x_3, x_4\}, B = \{x_5, x_6, x_7, x_8\}.$$

(a) Just using the above-standing plot, specify which of the points should be identified as support vectors.



The points $\{x_3, x_4, x_5, x_6\}$ are chosen as support vectors.

(b) Draw the maximum margin line which separates the classes (you don't have to do any computations here). Write down the normalized normal vector $\mathbf{w} \in \mathbb{R}^2$ of the separating line and the offset parameter $b \in \mathbb{R}$.



We obtain $\mathbf{w} = (1, 0)^T$, and $b = -5$.

(c) Consider the decision rule: $H(x) = \langle \mathbf{w}, x \rangle + b$. Explain how this equation classifies points on either side of a line. Determine the class for the points $x_9 = (3, 4)$, $x_{10} = (7, 4)$ and $x_{11} = (5, 5)$.

We have the following decision rule:

$$H(x) = sign\left(\left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, x \right\rangle - 5\right)$$

and hence,

$$H\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}\right) = sign\left(\left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \end{pmatrix} \right\rangle - 5\right) = sign(3 - 5) = sign(-2) = -1,$$

i.e. point x_9 is classified as belonging to class A (red).

$$H\left(\begin{pmatrix} 7 \\ 4 \end{pmatrix}\right) = sign\left(\left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 7 \\ 4 \end{pmatrix} \right\rangle - 5\right) = sign(7 - 5) = sign(2) = 1,$$

i.e. point x_{10} is classified as belonging to class B (blue).

$$H\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}\right) = sign\left(\left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 \\ 5 \end{pmatrix} \right\rangle - 5\right) = sign(5 - 5) = sign(0) = 0,$$

i.e. point x_{11} lies exactly on the decision boundary.

Exercise 11-2 Kernel Trick

Consider the polynomial kernel function

$$K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}, (x, y) \mapsto (x^T y + \gamma)^p, \text{ with } p = 2, \gamma = 1.$$

Furthermore let

$$\phi : \mathbb{R}^2 \to \mathbb{R}^6, x \mapsto (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2).$$

Show that $K(x, y) = \langle \phi(x), \phi(y) \rangle$.

$$\begin{aligned}
K(x, y) &= \langle \phi(x), \phi(y) \rangle \\
(x^T y + 1)^2 &= \left\langle (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2), (1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, y_2^2, \sqrt{2}y_1 y_2) \right\rangle \\
(x_1 y_1 + x_2 y_2 + 1)^2 &= 1 + 2x_1 y_1 + 2x_2 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1 x_2 y_1 y_2 \\
x_1^2 y_1^2 + 2x_1 y_1 x_2 y_2 + 2x_1 y_1 + x_2^2 y_2^2 + 2x_2 y_2 + 1 &= 1 + 2x_1 y_1 + 2x_2 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1 x_2 y_1 y_2
\end{aligned}$$

Exercise 11-3 Mercer Kernels

As known from the lecture, a Mercer kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ needs to fulfil

(1) Symmetry, i.e., $\kappa(x, y) = \kappa(y, x)$

(2) Positive semi-definiteness, i.e. the kernel matrix $\kappa(X) := (\kappa(x_i, x_j))_{ij} \in \mathbb{R}^n$ is positive semi-definite for all $X = \{x_1, \dots, x_n\} \subseteq \mathcal{X}$.

Show that the following functions are Mercer kernels for $x, y \in \mathcal{X} = \mathbb{R}^d$.

(a) $\kappa_1(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}.$

Obviously, κ_1 is symmetric. Furthermore, we have $\kappa_1(X) = I_n$ for all $X \subseteq \mathcal{X}$ with $|X| = n$. Thus, for arbitrary $c \in \mathbb{R}^n$ it holds

$$c^T \kappa_1(X) c = c^T (I_n) c = c^T c = \|c\|_2^2 \geq 0$$

Hence, κ_1 is a Mercer kernel.

(b) $\kappa_2(x, y) = x^T y.$

Due to $x^T y = y^T x$ for $x, y \in \mathbb{R}^d$, κ_2 is symmetric. Let $\mathfrak{X} \in \mathbb{R}^{d \times n}$ with $\mathfrak{X}_{ij} = (x_j)_i$. Then, for arbitrary $c \in \mathbb{R}^n$ it holds

$$c^T \kappa_2(X) c = c^T (\mathfrak{X}^T \mathfrak{X}) c = (c^T \mathfrak{X}^T)(\mathfrak{X} c) = (\mathfrak{X} c)^T (\mathfrak{X} c) = \|\mathfrak{X} c\|_2^2 \geq 0$$

Therefore, κ_2 is Mercer kernel.

(c) $\kappa_3(x, y) = \alpha x^T y + \beta$ for $\alpha, \beta \in \mathbb{R}$ with $\alpha, \beta \geq 0$

First, we notice $\kappa_3(x, y) = \alpha \kappa_2(x, y) + \beta$. As κ_2 is symmetric, the same holds for κ_3. Moreover,

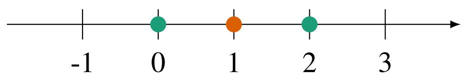
$$c^T \kappa_3(X) c = c^T (\alpha \kappa_2(X) + \beta) c = \alpha \underbrace{c^T \kappa_2(X) c}_{:= \gamma \geq 0} + \beta c^T c = \alpha\gamma + \beta \|c\|_2^2 \geq 0$$

Exercise 11-4 Linear Separability

In the following exercise, provide minimal subsets $\{x_1, \dots, x_m\} = X \subseteq \mathcal{X} = \mathbb{R}^d$ together with class labels $y_1, \dots, y_m \in \{-1, 1\}$ for the given dimensionality $d \in \mathbb{N}$ that are not linear separable. Prove both, the minimality (i.e. every $X' \subseteq \mathcal{X}$ with $|X'| < |X|$ is linearly separable), as well as the non-separability of X.

(a) $d = 1$

Consider $X = \{x_1, x_2, x_3\} = \{1, 2, 3\}$, and $y_1 = y_3 = 1$, $y_2 = -1$ as depicted below:



In $\mathbb{R}^1$, a hyperplane consists of a single threshold point τ and a linear separation can be achieved using a decision function

$$\begin{aligned} H(x) &= \text{sign}(x - \tau) \\ &= \begin{cases} -1 & x < \tau \\ 1 & x \geq \tau \end{cases} \end{aligned}$$

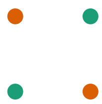
For the sake of contradiction, assume that the classes are linearly separable. Then, $x_1 < x_2$, and $y_1 \neq y_2$ implies that there is a separation between x_1 and x_2, i.e. $x_1 < \tau \leq x_2$. Hence, $y = 1$. But then, $x_2 < x_3$ and $\tau \leq x_2$ implies that $H(x_2) = H(x_3)$. This contradicts $y_2 \neq y_3$. Thus, the classes are not linearly separable.

Moreover, there is no smaller such set. Consider the case $m = 2$ and let $X' = \{x_1, x_2\}$. If $y_1 = y_2$, there are no classes to separate and we are finished. Hence, let $y_1 \neq y_2$. However, choosing $\tau = \frac{1}{2}(x_1 + x_2)$, and $y = y_1$ yields a linear classifier with perfect prediction, i.e. X' is linearly separable. Since linear separability of all sets of size m implies linear separability of all sets of size $m - 1$, X is minimal.

(b) $d = 2$

We can re-use the example from above, and just append a constant dimension to every data point.

However, if we forbid that the data is situated in a 1-dimensional subspace, we need one more point. Consider $X = \{x_1, \ldots, x_4\}$ with $x_1 = (-1, -1)$, $x_2 = (-1, 1)$, $x_3 = (1, -1)$, $x_4 = (1, 1)$, and $y_1 = y_4 = 1$, and $y_2 = y_3 = -1$, as depicted below:

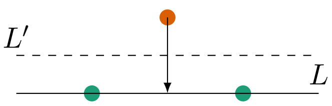


Assume, there exists a linear split by $\mathbf{w} = (w_0, w_1, w_2) \in \mathbb{R}^3$. Then, is must hold that

$$
\begin{aligned}
(w^T \tilde{x}_i) y_i &> 0 \quad \text{for all } x_i && (1)\\
\implies \sum_i (w^T \tilde{x}_i) y_i &> 0 && (2)\\
\iff w^T \tilde{X}^T y &> 0 && (3)\\
\iff \begin{pmatrix} w_0 & w_1 & w_2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} &> 0 && (4)\\
\iff \begin{pmatrix} w_0 & w_1 & w_2 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} &> 0 && (5)\\
\iff 0 &> 0 && (6)
\end{aligned}
$$

Obviously, the last line is not true and hence, such parameter vector does not exist.

Assume there is a $X' = \{x_1, \ldots, x_3\}$ that is not linearly separable, and spans over 2 dimensions. If all y_i are the same, nothing remains to be shown. Hence, without loss of generality, assume $y_1 = -1$, $y_2 = y_3 = 1$.



Then, there exists a line that separates x_1 from x_2 and x_3: Point x_1 has a non-zero distance to the line L through x_2 and x_3 (otherwise, the three points would lie on one line, and thus not span a 2-dimensional space). Hence, we can use a line L' parallel L and between L and x_1 as separating hyperplane (cf. image).