Exercises Week 7

Machine Learning/Advanced Machine Learning
IT University of Copenhagen

Fall 2019

Theoretical Exercises 7.1: from the Book

As stated on learnit, solve the following exercises from the book:

- (13.16.10)
- (14.10.1)
- (14.10.7)

Theoretical Exercises 7.2. SVM

The XOR problem consists of four points from two classes, which are not linearly separable, as follows:

- class 1: $\boldsymbol{x}_1$, $\boldsymbol{x}_2$,
- class 2: $\boldsymbol{x}_3$, $\boldsymbol{x}_4$,

given the four points:

$$\boldsymbol{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{x}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \qquad \boldsymbol{x}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{x}_4 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{1}$$

with labels:

$$r_1 = +1, \qquad r_2 = +1 \qquad r_3 = -1, \qquad r_4 = -1. \tag{2}$$

The goal of this exercise is to compute the discriminant:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \tag{3}$$

which enables a linear classification in a higher dimension, enabled by the basis function $\boldsymbol{\phi}$.

(a) Draw the points and highlight to which class which point belongs.

(b) Since the points are not linearly separable in 2D, they should be transferred to a higher dimension, such that they become. Use the following basis function to transfer each of the four 2D points to 6D:

$$\boldsymbol{\phi} : \mathbb{R}^2 \to \mathbb{R}^6 \tag{4}$$

$$\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{\phi}(x_1, x_2) = \left(1,\ \sqrt{2}\ x_1,\ \sqrt{2}\ x_2,\ \sqrt{2}\ x_1 x_2,\ x_1^2,\ x_2^2\right)^{\mathrm{T}}, \tag{5}$$

i.e. calculate $\boldsymbol{z}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$, $i = 1, \ldots, 4$.

(c) Use the known values to complete Eq. (13.26):

$$L_d(\boldsymbol{\alpha}) = L_d(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_i \alpha_j r_i r_j \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{z}_j \tag{6}$$

$$= \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_i \alpha_j r_i r_j \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}_j) \tag{7}$$

(d) Compute the derivative of $L_d(\boldsymbol{\alpha})$ with respect to α_i, i.e. the four components of the gradient:

$$\nabla L_d(\boldsymbol{\alpha}) = \begin{pmatrix} \frac{\partial}{\partial \alpha_1} L_d(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \\ \frac{\partial}{\partial \alpha_2} L_d(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \\ \frac{\partial}{\partial \alpha_3} L_d(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \\ \frac{\partial}{\partial \alpha_4} L_d(\alpha_1, \alpha_2, \alpha_3, \alpha_4) \end{pmatrix} = \dots \tag{8}$$

(e) Derive the equation system from $\nabla L_d(\boldsymbol{\alpha}) = 0$ and solve for $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^{\mathrm{T}}$.

(f) Which of the four training points are support vectors? How do the values of α_i answer this question?

(g) Now that all four values of $\boldsymbol{\alpha}$ have been computed, employ Eq. (13.24) to compute $\boldsymbol{w}$:

$$\boldsymbol{w} = \sum_{i=1}^{4} \alpha_i r_i \boldsymbol{z}_i = \sum_{i=1}^{4} \alpha_i r_i \boldsymbol{\phi}(\boldsymbol{x}_i) \tag{9}$$

Please note: $\boldsymbol{w}, \boldsymbol{z}_i \in \mathbb{R}^6$.

(h) Give the discriminant function g based on the original input space

$$g(\boldsymbol{x}) = g(x_1, x_2) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) = \dots \tag{10}$$

(i) Apply the discriminant function and compute the resulting values for the training input samples $g(\boldsymbol{x}_i)$, $i = 1, \dots, 4$. How are they classified? Are they correctly classified?

Programming Exercise 7.3.: kernel SVM

The goal is to implement and evaluate different kernels for SVMs for one dataset. For this programming exercise the notebook `exercise_svm.ipynb` is provided, which should be used and adapted.

(a) Implement:

- a linear kernel
- a radial basis function kernel
- a polynomial kernel

(b) Which of these performs best on the data, in terms of speed and quality? Do not forget to set the random seed to receive reproducible results.

(c) Test different values of c and d for the polynomial kernel. Which of them work best?

(d) Test different values of γ for the RBF kernel. Which of them works best?

(e) Change the part of the code which generates the data such that it becomes linearly separable.

(f) Re-evaluate the three kernels. Do you get the same result?