***** Advanced Machine Learning *****
Support Vector Machines (SVMs)

M. Sc. Daniel Wehner

SAP SE / DHBW Mannheim

Summer term 2020

DHBW
Duale Hochschule
Baden-Württemberg

Agenda for this Unit

Introduction

What is a Support Vector Machine?

- A **support vector machine** is a **binary classifier**. The classes have to be denoted by $\{-1, +1\}$. The classes -1 and $+1$ are denoted by $\ominus$ and $\oplus$, respectively.
- The original algorithm was proposed by *Vapnik* and *Chervonenkis* already in 1963. Several extensions were made in the 90s, including **non-linear SVMs** as well as **soft-margin SVMs**.
- Multi-class classification can be performed using the well-known techniques:
  - One-vs-Rest (OVR)
  - One-vs-One (OVO)
- An SVM finds the best separating hyperplane and therefore has built-in **generalization guarantees**. Question: *What is the best hyperplane?*
- An SVM is no physical machine, rather it is a mathematical construct (cf. Turing machine).

Discriminant Functions

- The simplest discriminant function has a linear form:

$$\widehat{h}(\boldsymbol{x}) = \boldsymbol{w}^\intercal \boldsymbol{x} + b = \sum_{j=1}^{m} w_j x_j + b = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + b \tag{1}$$

- The parameters to be optimized are $\boldsymbol{\theta} = \{\boldsymbol{w}, b\}$. $\boldsymbol{w}$ is called the weight vector, b is called the bias.
- An arbitrary input vector $\boldsymbol{x}$ is assigned to class $\oplus$, if $\widehat{h}(\boldsymbol{x}) \geqslant 0$, class $\ominus$ otherwise.

$$h(\boldsymbol{x}) = \text{sign}(\widehat{h}(\boldsymbol{x})) \tag{2}$$

- The **decision boundary** is defined by the relation: $\widehat{h}(\boldsymbol{x}) = 0$.
- The boundary is a $(D-1)$-dimensional hyperplane within the D-dimensional input space.

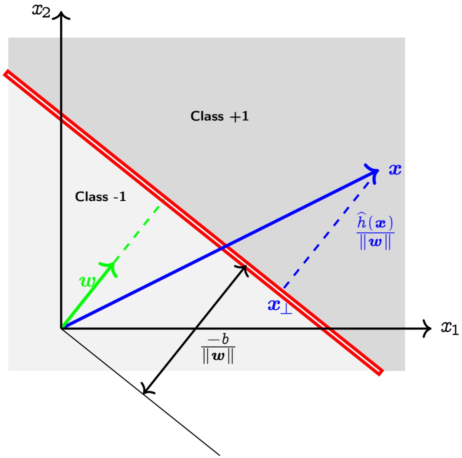

Figure 1: A linear discriminant function

- Consider two points, $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$, which both lie on the decision surface.
- Since $\widehat{h}(\boldsymbol{x}_a) = \widehat{h}(\boldsymbol{x}_b) = 0$, we have $\boldsymbol{w}^\intercal(\boldsymbol{x}_a - \boldsymbol{x}_b) = 0$. Hence, $\boldsymbol{w}$ is orthogonal to every vector lying within the decision surface.
- Thus, $\boldsymbol{w}$ determines the orientation of the decision surface.
-

Linear Separability

- Consider n input vectors $\boldsymbol{X} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}\}$.
- Each input vector $\boldsymbol{x}^{(i)}$ is labeled with $y^{(i)}$, where $y^{(i)} \in \{-1, +1\}$.
- A data set is linearly separable in feature space, if $\exists(\boldsymbol{w}, b)$ such that:

$$\widehat{h}(\boldsymbol{x}^{(i)}) = \boldsymbol{w}^\intercal \boldsymbol{x}^{(i)} + b > 0 \qquad \forall \boldsymbol{x}^{(i)} \text{ with } y^{(i)} = +1 \tag{3}$$

$$\widehat{h}(\boldsymbol{x}^{(i)}) = \boldsymbol{w}^\intercal \boldsymbol{x}^{(i)} + b < 0 \qquad \text{otherwise } (y^{(i)} = -1) \tag{4}$$

- This can be written as:

$$y^{(i)} \widehat{h}(\boldsymbol{x}^{(i)}) > 0 \qquad \forall i \tag{5}$$

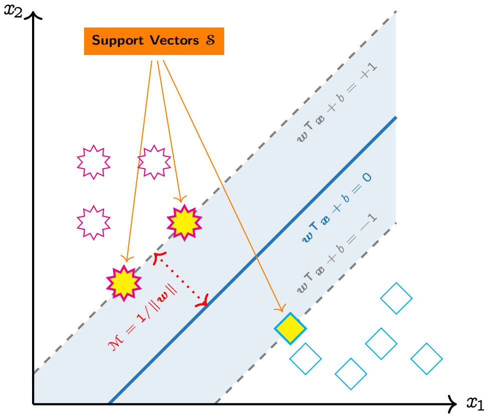

Figure 2: Maximum margin classifiers find the best separating hyperplane

Thank you very much for the attention!

Topic: ***** Advanced Machine Learning ***** Support Vector Machines (SVMs)
Term: Summer term 2020

Contact:
M. Sc. Daniel Wehner
SAP SE / DHBW Mannheim
daniel.wehner@sap.com

Do you have any questions?