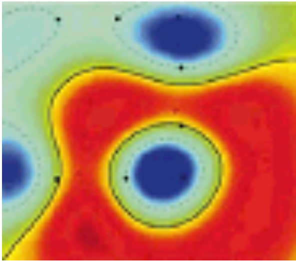
WiSe 2024/25

Machine Learning 1/1-X

Technische Universität Berlin



Lecture 5

Support Vector Machines

Support Vector Machines

Outline

- Lagrange Duality
- KKT optimality conditions
- Large margin classifiers
- Hard-margin SVM (Primal / Dual)
- Soft-margin SVM (Primal)
- Kernel SVM
- SVM and Hinge Loss
- SVM beyond Classification
- Applications

Lagrange Duality (1)

- We consider optimization problem in **canonical** form:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f_0(\boldsymbol{x}) \\ \text{subject to} \quad & f_i(\boldsymbol{x}) \leqslant 0, \quad i = 1, \ldots, m \\ & h_i(\boldsymbol{x}) = 0, \quad i = 1, \ldots, p \end{aligned}$$

- The **Lagrange function** $\mathcal{L} \colon \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as a weighted sum of the objective and constraint functions:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{i=1}^{p} \mu_i h_i(\boldsymbol{x}),$$

where $\boldsymbol{x}$ is called **primal** and $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ the **dual** variables.

Lagrange Duality (2)

- The (Lagrange) **dual function** $g\colon \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as:

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \text{domain}(f_0)} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

- The (convex!) optimization problem

$$\begin{aligned} \underset{(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\text{maximize}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \succeq \mathbf{0} \end{aligned}$$

  is called the (Lagrange) **dual problem**.

- The inequality $d^* \leqslant p^*$ always holds, where p^*, d^* are the optimal values of the primal and dual problem, respectively.

Lagrange Duality (3)

- The (Lagrange) **dual function** $g\colon \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as:

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \text{domain}(f_0)} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

- The (convex!) optimization problem

$$\begin{aligned} \underset{(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\text{maximize}} \quad & g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \succeq \mathbf{0} \end{aligned}$$

  is called the (Lagrange) **dual problem**.

- The inequality $d^* \leqslant p^*$ always holds, where p^*, d^* are the optimal values of the primal and dual problem, respectively. We refer to the difference $p^* - d^*$ as duality gap. In the case $p^* = d^*$ we talk about **strong duality**.

Karush–Kuhn–Tucker (KKT) Conditions

Theorem: Optimality Conditions

- For any optimization problem with differentiable objective and constraint functions for which strong duality obtains, any pair of primal and dual optimal $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ must satisfy KKT-conditions:

$$
\begin{aligned}
&\nabla_x \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = 0, && \text{(stationarity)} \\
&f_i(\boldsymbol{x}^*) \leqslant 0, && \text{(primal feasibility)} \\
&h_i(\boldsymbol{x}^*) = 0, && \text{(primal feasibility)} \\
&\lambda_i^* \geqslant 0, && \text{(dual feasibility)} \\
&\lambda_i^* \cdot f_i(\boldsymbol{x}^*) = 0 && \text{(complementary slackness)}
\end{aligned}
$$

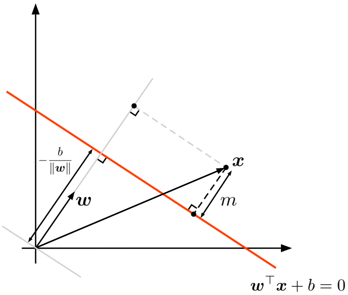
- For any convex problem, the KKT-conditions are sufficient for $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ to be optimal with zero duality gap.

Hard-Margin SVM (Derivation)

- Given data $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ with $y_i \in \{-1, 1\}$, we want to maximize the separation margin of the linear classifier $y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$:

$$\underset{\boldsymbol{w}, b}{\text{maximize}} \ \frac{1}{\|\boldsymbol{w}\|} \min_{i=1,\ldots,n} y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b)$$

- Observation: rescaling $\boldsymbol{w} \mapsto k\boldsymbol{w}$ and $\boldsymbol{b} \mapsto k\boldsymbol{b}$, $k \neq 0$ results in the same objective value. We can use this fact to set $\min_{i=1,\ldots,n} y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) = 1$.



$$\boxed{m = \frac{\boldsymbol{w}^\top \boldsymbol{x}}{\|\boldsymbol{w}\|} - \left(-\frac{b}{\|\boldsymbol{w}\|}\right) = \frac{\boldsymbol{w}^\top \boldsymbol{x} + b}{\|\boldsymbol{w}\|}}$$

Hard-Margin SVM (Derivation)

- This gives the following optimization problem

$$\underset{\boldsymbol{w},b}{\text{maximize}}\ \frac{1}{\|\boldsymbol{w}\|} \quad \text{subject to} \quad \min_{i=1,\ldots,n} y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) = 1$$

  or equivalently

$$\underset{\boldsymbol{w},b}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{subject to} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x_i} + b) \geqslant 1, i = 1, \ldots, n$$

- Replacing $\boldsymbol{x}$ by (non-linear) features $\phi(\boldsymbol{x})$ gives the (primal) **hard-margin** formulation of the Support Vector Machine:

$$\underset{\boldsymbol{w},b}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{subject to} \quad y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geqslant 1, i = 1, \ldots, n$$

Hard-Margin SVM (Primal Problem)

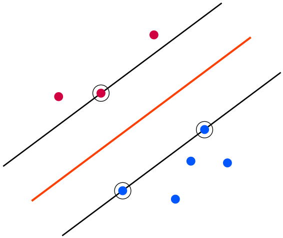
1. The classifier with largest margin between the positive and negative data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\dots,n}$ can be obtained by solving a convex optimization problem:

$$\boxed{\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ \text{subject to} & y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geqslant 1, \quad i = 1, \dots, n \end{array}}$$

2. The decision function $f \colon \mathbb{R}^d \to \{1, -1\}$ is given by

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \phi(\mathbf{x}) + b)$$

and it can be used to classify new data.



3. Data points $(\boldsymbol{x}_i, y_i)$ where a corresponding constraint is active, i. e., $y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) = 1$ are called **support vectors**.

Deriving the Dual of the Hard-Margin SVM

- Consider the hard-margin formulation of SVM

$$\underset{\boldsymbol{w},b}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{subject to} \quad y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geqslant 1, \quad i = 1, \ldots, n$$

- Write the Lagrangian

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}) + b))$$

- Compute the dual function $g(\boldsymbol{\alpha}) = \inf_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha})$:

$$g(\boldsymbol{\alpha}) = \begin{cases} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle, & \text{if } \sum_{i=1}^n \alpha_i y_i = 0 \\ -\infty & \text{else} \end{cases}$$

where we used the fact that $\mathcal{L}$ is strictly convex and

$$\nabla_{\boldsymbol{w}} \mathcal{L} = \mathbf{0} \ \Rightarrow \ \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \phi(\boldsymbol{x}_i) \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial b} = 0 \ \Rightarrow \ \sum_{i=1}^{n} \alpha_i y_i = 0$$

Hard-Margin SVM (Dual Problem)

- The dual problem has the following form:

$$\boxed{\begin{aligned} \underset{\boldsymbol{\alpha}\succeq 0}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j)\rangle \\ \text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}}$$

- Due to the relationship $\boldsymbol{w} = \sum \alpha_i y_i \phi(\boldsymbol{x}_i)$ the decision function is given as

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x})\rangle + b$$

- How do we find the bias b? Note that for each support vector $\boldsymbol{x}_i \in S$, it holds $y_i \cdot f(\boldsymbol{x}_i) = 1$, where S denotes the set of support vectors. Here, it is enough to use one arbitrary support vector to compute b. However, the following provides numerically more stable solution:

$$b = \frac{1}{|S|}\sum_{i\in S}(y_i - \sum_{j\in S}\alpha_j y_j \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle)$$

Hard-Margin SVM (Dual Problem)

- On the previous slide we saw how to compute bias in the dual formulation:

$$b = \frac{1}{|S|} \sum_{i \in S} (y_i - \sum_{j \in S} \alpha_j y_j \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle)$$

  The remaining question here is how to find S.
- Based on the complementary slackness in the KKT-conditions

$$\alpha_i \cdot (1 - y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b)) = 0$$

  we conclude

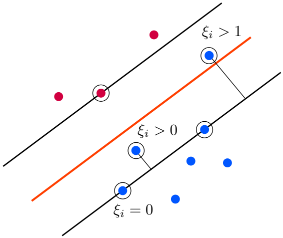
$$\boldsymbol{x}_i \text{ is a support vector } \Leftrightarrow \alpha_i > 0.$$

Soft-Margin SVM (Primal Problem)

1. If the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\ldots,n}$ is not separable (e.g. due to noise), we introduce *slack variables* $(\xi_i)_i$ that allows for data points to violate the margin constraints at the cost of additional penalty. We refer to this formulation as **soft-margin** SVM:

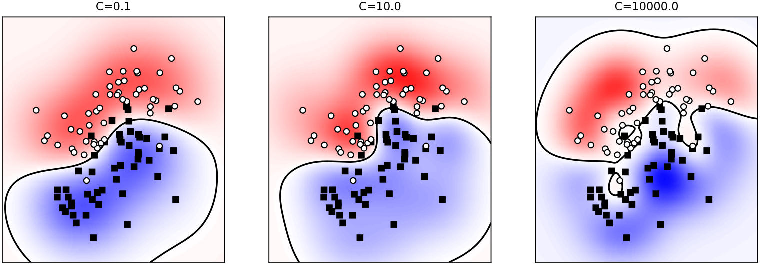
$$\begin{aligned} \underset{\boldsymbol{w},b,\boldsymbol{\xi}}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) + b) \geqslant 1 - \xi_i, \quad && i = 1, \ldots, n \\ & \xi_i \geqslant 0, && i = 1, \ldots, n \end{aligned}$$

2. Here, $C \in (0, \infty)$ is a regularization constant controlling the trade-off between the margin size and the constraint violation. For $C \to \infty$ we recover the hard-margin formulation.
3. Data points $(\boldsymbol{x}_i, y_i)$ for which either $\xi_i > 0$ or $y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) + b) = 1$ holds are called **support vectors**.



Effect of the parameter C

The larger the parameter C the more the decision boundary is forced to correctly classify every data point. For $C \to \infty$ we recover the hard-margin formulation. For $C \to 0$ the robustness of "correctly" classified points increases.



Kernel Functions

Definition (Kernel function)

A kernel is a function $\kappa \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ satisfies

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle$$

where $\phi \colon \mathcal{X} \to \mathcal{F}$ is a mapping from some $\mathcal{X}$ to a Hilbert space $(\mathcal{F}, \langle \cdot, \cdot \rangle)$.

Definition (Finitely positive semi-definite functions)

A function $\kappa \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfies the finitely positive semi-definite property if it is symmetric and for which the matrices formed by restriction to any finite subset of the space $\mathcal{X}$ are positive semi-definite.

Theorem (Kernel matrices)

The kernel functions satisfy the finitely positive semi-definite property. That is, the corresponding kernel matrices are positive semi-definite.

Examples of Kernels

Observation: In the SVM dual form, we never need to access the feature map $\phi(\cdot)$ explicitly. Instead, we can always express computations in terms of the kernel function.

Examples of commonly used kernels satisfying the Mercer property are:

Linear	$k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$	
Polynomial	$k(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + \beta)^\gamma$	$\beta \in \mathbb{R}_{\geq 0}, \gamma \in \mathbb{N}$
Gaussian	$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$	$\gamma \in \mathbb{R}_{>0}$

Note: The feature map associated to the Gaussian kernel is infinite-dimensional. However, in the dual form, we never need to access it for training and prediction, and we only need the kernel function.