Exercises for the course

Machine Learning 1

Winter semester 2024/25

Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 6

Exercise 1: Dual formulation of the Soft-Margin SVM $(5+20+10+5$ P)

The primal program for the linear soft-margin SVM is

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

subject to

$$\forall_{i=1}^{N}: \; y_i \cdot (\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0$$

where $\|.\|$ denotes the Euclidean norm, ϕ is a feature map, $\boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}$ are the parameter to optimize, and $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ are the labeled data points regarded as fixed constants. Once the hard-margin SVM has been learned, prediction for any data point $\boldsymbol{x} \in \mathbb{R}^d$ is given by the function

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

(a) *State* the conditions on the data under which a solution to this program can be found from the Lagrange dual formulation *(Hint: verify the Slater's conditions)*.

(b) *Derive* the Lagrange dual and show that it reduces to a constrained quadratic optimization problem. State both the objective function and the constraints of this optimization problem.

(c) *Describe* how the solution $(\boldsymbol{w}, b)$ of the primal program can be obtained from a solution of the dual program.

(d) *Write* a kernelized version of the dual program and of the learned decision function.

Exercise 2: SVMs and Quadratic Programming (10 P)

We consider the `CVXOPT` Python software for convex optimization. The method `cvxopt.solvers.qp` solves quadratic optimization problems given in the matrix form:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^\top P \boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} \\ \text{subject to} \quad & G\boldsymbol{x} \preceq \boldsymbol{h} \\ \text{and} \quad & A\boldsymbol{x} = \boldsymbol{b}. \end{aligned}$$

Here, $\preceq$ denotes the element-wise inequality: $(\boldsymbol{h} \preceq \boldsymbol{h}') \Leftrightarrow (\forall_i : h_i \leq h'_i)$. Note that the meaning of the variables $\boldsymbol{x}$ and $\boldsymbol{b}$ is different from that of the same variables in the previous exercise.

(a) *Express* the matrices and vectors $P, \boldsymbol{q}, G, \boldsymbol{h}, A, \boldsymbol{b}$ in terms of the variables of Exercise 1, such that this quadratic minimization problem corresponds to the kernel dual SVM derived above.

Exercise 3: Programming (50 P)

Download the programming files on ISIS and follow the instructions.

Exercise 1: Dual formulation of the Soft-Margin SVM ($5 + 20 + 10 + 5$ P)

The primal program for the linear soft-margin SVM is

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

subject to

$$\forall_{i=1}^{N}: \; y_i \cdot (\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0$$

where $\|.\|$ denotes the Euclidean norm, ϕ is a feature map, $\boldsymbol{w} \in \mathbb{R}^d, b \in \mathbb{R}$ are the parameter to optimize, and $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$ are the labeled data points regarded as fixed constants. Once the hard-margin SVM has been learned, prediction for any data point $\boldsymbol{x} \in \mathbb{R}^d$ is given by the function

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

(a) *State* the conditions on the data under which a solution to this program can be found from the Lagrange dual formulation *(Hint: verify the Slater's conditions)*.

According to the Slater's Theorem, if the problem is convex and Slater's condition is satisfied, (i.e. there exists w^* such that $\forall_{i=1}^{N}: y_i \cdot (w^{*T}\phi(x_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$), then the strong duality holds.

Verify the Slater's condition:

We could simply assume all training data seperated correctly, i.e. $\xi_i = 0, \forall_{i=1}^{N}$

$\Rightarrow \forall_{i=1}^{N}: y_i \cdot (w^T\phi(x_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ always holds. if we choose $\xi_i = 0, \forall_{i=1}^{N}$

(a) Soft-margin SVM is given by a convex optimization problem: the objective is convex and the inequality constraints are linear (therefore also convex). Furthermore, the Slater's Theorem guarantees that if there is a feasible point $(\boldsymbol{w}, b, \boldsymbol{\xi})$ which strictly satisfies the inequality constraints, then strong duality holds. Here, for any $(\boldsymbol{w}, b)$ we can always choose sufficiently large values for the slack variables $\boldsymbol{\xi}$ such that all inequality constraints are strictly satisfied. Therefore, strong duality (in contrast to the hard-margin) holds always for the soft-margin formulation.

(b) *Derive* the Lagrange dual and show that it reduces to a constrained quadratic optimization problem. State both the objective function and the constraints of this optimization problem.

Optimization Problem in Canonical Form:

$$\underset{w,b,\xi}{\text{minimize}} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to} \quad 1 - \xi_i - y_i(w^T\phi(x_i) + b) \leq 0 \qquad i = 1, \dots, n$$

$$-\xi_i \leq 0 \qquad i = 1, \dots, n$$

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i(w^T\phi(x_i) + b)\right) + \sum_{i=1}^{n}\beta_i(-\xi_i)$$

$$g(\alpha, \beta) = \inf_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i(w^T\phi(x_i) + b)\right) + \sum_{i=1}^{n}\beta_i(-\xi_i)$$

$$= \inf_{w,b,\xi} \left(\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i y_i(w^T\phi(x_i) + b)\right) + \left(-\sum_{i=1}^{n}\alpha_i y_i b\right) + \left(C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i\xi_i\right) + \sum_{i=1}^{n}\alpha_i$$

$$= \inf_{w}\left\{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i y_i w^T\phi(x_i)\right\} + \inf_{b}\left\{-b\sum_{i=1}^{n}\alpha_i y_i\right\} + \inf_{\xi}\left\{\sum_{i=1}^{n}\xi_i(C - \alpha_i - \beta_i)\right\} + \sum_{i=1}^{n}\alpha_i$$

Note that the minimization over b and $\boldsymbol{\xi}$ is completely unrestricted. Therefore, the only way for the infimum to be bigger that $-\infty$ if the constrains $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $C - \alpha_i - \beta_i = 0$ are satisfied. This is in agreement with the results below. To find the minimizing arguments $(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*)$ we set the gradient of the corresponding terms to zero as follows:

$$\nabla_w L = w - \sum_{i=1}^{n} \alpha_i y_i \phi(x_i) = 0 \qquad \Rightarrow w^* = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i)$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \qquad \overset{\beta_i \geq 0}{\Longrightarrow} \quad 0 \leq \alpha_i \leq C$$

$$g(\alpha, \beta) = \inf_w \left\{ \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i y_i w^T \phi(x_i) \right\} + \inf_b \left\{ -b \sum_{i=1}^{n} \alpha_i y_i \right\} + \inf_\xi \left\{ \sum_{i=1}^{n} \xi_i (C - \alpha_i - \beta_i) \right\} + \sum_{i=1}^{n} \alpha_i$$

$$= \frac{1}{2}\|w^*\|^2 - \sum_{i=1}^{n} \alpha_i y_i w^{*T} \phi(x_i) + \inf_b \left\{ -b \sum_{i=1}^{n} \alpha_i y_i \right\} + \inf_\xi \left\{ \sum_{i=1}^{n} \xi_i (C - \alpha_i - \beta_i) \right\} + \sum_{i=1}^{n} \alpha_i$$

$$= \begin{cases} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) + \sum_{i=1}^{n} \alpha_i & \text{if } \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ and } C - \alpha_i - \beta_i = 0 \\ -\infty & \text{otherwise} \end{cases}$$

where we used

$$\frac{1}{2}\|w^*\|^2 - \sum_{i=1}^{n} \alpha_i y_i w^{*T} \phi(x_i) = \frac{1}{2} \left\langle \sum_{i=1}^{n} \alpha_i y_i \phi(x_i), \sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \right\rangle - \sum_{i=1}^{n} \alpha_i y_i \left\langle \sum_{i=1}^{n} \alpha_i y_i \phi(x_i), \phi(x_i) \right\rangle$$

$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i y_i \alpha_j y_j \phi(x_i)^T \phi(x_j)$$

$$= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j)$$

dual problem: $\quad \max_{\alpha_1, \dots, \alpha_n} \; -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j) + \sum_{i=1}^{n} \alpha_i$

subject to $\quad \forall i: 0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$

Problematical!!!!!!!
See the SOLUTION!!

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i \cdot ((w \cdot \phi(x_i) + b) - 1 \right)$$

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{1}{2} \cdot 2 \cdot w - \sum_{i=1}^{N} \alpha_i y_i \phi(x_i) = 0 \qquad \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i \phi(x_i)$$

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i = 0$$

The dual function is:

$$\max \quad W(\alpha) = \mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i \cdot ((w \cdot \phi(x_i) + b) - 1 \right)$$

$$= \frac{1}{2} w^T \cdot w - \sum_{i=1}^{N} \alpha_i \left(y_i \left(\sum_{j=1}^{N} \alpha_j y_j \phi(x_j) \phi(x_i) + b \right) \right) + \sum_{i=1}^{N} \alpha_i$$

$$= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j) - \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j) + \alpha_i y_i b \right) + \sum_{i=1}^{N} \alpha_i$$

$$= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j) + \underbrace{\sum_{i=1}^{N} \alpha_i y_i \cdot b}_{=0} + \sum_{i=1}^{N} \alpha_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \phi(x_i)\phi(x_j) + \sum_{i=1}^{N}\alpha_i$$

Subject to $\quad C \geq \alpha_i \geq 0,\ i = 1,\cdots,N$ and $\sum_{i=1}^{N}\alpha_i y_i = 0$

(c) *Describe* how the solution $(\boldsymbol{w}, b)$ of the primal program can be obtained from a solution of the dual program.

from the dual problem we obtain all α_i, $\forall i_{i=1}^{N}$. Therefore the w can be computed by $w = \sum_{i=1}^{N}\alpha_i y_i \phi(x_i)$

(from 1.b)

If $\alpha_i > 0$, then the corresponding data point x_i is a support vector, which implies $y_i[(w\cdot\phi(x_i)+b] = 1$, i.e the support vector x_i is in the margin

Therefore we could use all the support vector x_i, where $0 < \alpha_i < C$, to calculate b by $y_i[(w\cdot\phi(x_i)+b] = 1$

$\Rightarrow$ the solution (w, b) is found

(c) From the previous solution in (b) we know that

$$\boldsymbol{w}^* = \sum_{i=1}^{n}\alpha_i^* y_i \boldsymbol{\phi}(\boldsymbol{x}_i).$$

To find b^* we use the KKT condition (complementary slackness) "$\lambda_i \cdot f_i(\boldsymbol{x}) = 0$". Note that the data points $\boldsymbol{x}_i$ with $\alpha_i = 0$ do not contribute to the decision boundary. All other points with $\alpha_i > 0$ constitute the support vectors. Points with $\alpha_i = C$ lie inside the margin (or even on the wrong side of the decision boundary). Consider a support vector with $0 < \alpha_i < C$. Such support vectors lie exactly on the margin boundary! This follows from the complementary slackness:

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

$$\beta_i(-\xi_i) = 0 \quad \overset{\beta_i = C - \alpha_i > 0}{\Longrightarrow} \quad \xi_i = 0,$$

which together implies

$$0 < \alpha_i < C \quad \Longrightarrow \quad b = y_i - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_i) = y_i - \sum_{j=1}^{n}\alpha_j y_j \boldsymbol{\phi}(\boldsymbol{x}_j)^\top \boldsymbol{\phi}(\boldsymbol{x}_i).$$

(d) *Write* a kernelized version of the dual program and of the learned decision function.

dual fuction (kernelized)

from 1.b: $\quad \max$

$$W(\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \phi(x_i)\phi(x_j) + \sum_{i=1}^{N}\alpha_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(x_i, x_j) + \sum_{i=1}^{N}\alpha_i$$

Subject to $\quad C \geq \alpha_i \geq 0,\ i = 1,\cdots,N$ and $\sum_{i=1}^{N}\alpha_i y_i = 0$

decision function: $\quad f(x) = \text{sign}(w^{T}\phi(x) + b)$

$$= \text{sign}\cdot\left(\sum_{i=1}^{N}\alpha_i y_i \phi(x_i)\cdot\phi(x) + b\right)$$

$$= \text{sign}\left(\sum_{i=1}^{N}\alpha_i y_i k(x_i, x) + b\right)$$

$$= \text{sign}\left(\sum_{i\in\#SVs}\alpha_i y_i k(x_i, x) + b\right)$$

Exercise 2: SVMs and Quadratic Programming (10 P)

We consider the `CVXOPT` Python software for convex optimization. The method `cvxopt.solvers.qp` solves quadratic optimization problems given in the matrix form:

$$\begin{aligned}\min_{\boldsymbol{x}} \quad & \frac{1}{2}\boldsymbol{x}^\top P\boldsymbol{x} + \boldsymbol{q}^\top \boldsymbol{x} \\ \text{subject to} \quad & G\boldsymbol{x} \preceq \boldsymbol{h} \\ \text{and} \quad & A\boldsymbol{x} = \boldsymbol{b}.\end{aligned}$$

Here, $\preceq$ denotes the element-wise inequality: $(\boldsymbol{h} \preceq \boldsymbol{h}') \Leftrightarrow (\forall_i : h_i \le h'_i)$. Note that the meaning of the variables $\boldsymbol{x}$ and $\boldsymbol{b}$ is different from that of the same variables in the previous exercise.

(a) *Express* the matrices and vectors $P, \boldsymbol{q}, G, \boldsymbol{h}, A, \boldsymbol{b}$ in terms of the variables of Exercise 1, such that this quadratic minimization problem corresponds to the kernel dual SVM derived above.

from 1.b:

$$\max_{\alpha} \quad W(\alpha) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(x_i,x_j) + \sum_{i=1}^{N}\alpha_i$$

subject to $\quad C \geq \alpha_i \geq 0, \; i = 1,\dots,N$

and $\sum_{i=1}^{N}\alpha_i y_i = 0$

$$\Rightarrow \quad \min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(x_i,x_j) - \sum_{i=1}^{N}\alpha_i$$

$$= \frac{1}{2}\cdot \underbrace{\vec{\alpha}^T}_{x^T} \underbrace{\left(\langle \vec{y}\,\vec{y}^T, K\rangle\right)}_{P} \underbrace{\vec{\alpha}}_{x} + \underbrace{(\vec{-1})^T}_{q^T} \cdot \underbrace{\vec{\alpha}}_{x}$$

$$q = (\vec{-1}) = \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$$

subject to $\quad \vec{0} \preceq 1\cdot\vec{\alpha} \preceq \vec{C}$

$$P = \operatorname{diag}(y_1,\dots,y_n)\, K \operatorname{diag}(y_1,\dots,y_n)$$

and $\quad \underbrace{\vec{y}^T}_{A} \cdot \underbrace{\vec{\alpha}}_{x} = \underbrace{0}_{b}$

$$\vec{0} \preceq 1\cdot\vec{\alpha} \preceq \vec{C} \quad \Leftrightarrow \quad \begin{aligned} -1\cdot\vec{\alpha} &\preceq \vec{0} \\ 1\cdot\vec{\alpha} &\preceq \vec{C} \end{aligned}$$

$$\Leftrightarrow \quad \underbrace{\begin{bmatrix} -1 & & \\ & \ddots & \\ & & -1 \\ 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix}}_{G} \vec{\alpha} \preceq \underbrace{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ C \\ \vdots \\ C \end{bmatrix}}_{h}$$

$$G = \begin{bmatrix} -I \\ I \end{bmatrix}, h = \begin{bmatrix} \mathbf{0} \\ C\cdot\mathbf{1} \end{bmatrix},$$

where $I \in \mathbb{R}^{n\times n}$ denotes the identity matrix, that is, $G \in \mathbb{R}^{2n\times n}$ and $h \in \mathbb{R}^{2n}$. The equality constraint $\sum_{i=1}^{n}\alpha_i y_i = 0$ can be represented as

$$A = \boldsymbol{y}^\top, b = 0,$$

where $\boldsymbol{y} = (y_1, ..., y_n)$.