Exercises for the course

Machine Learning 1

Winter semester 2021/22

Abteilung 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 8

Exercise 1: Dual formulation of the Soft-Margin SVM ($\mathbf{5+20+10+5}$ 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

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

(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.

Sheet 8

1

a) The solution can be found from the Lagrange dual formulation if strong duality holds for this convex optimization problem. Strong duality can be confirmed by verifying Slater's conditions:

I <u>Slater condition 1:</u>

All convex inequality constraints are fulfilled for at least one point $\tilde{x} \in \mathbb{R}^d$:

$$\forall_{i=1}^{N} \quad y_i(\underline{w}^T \phi(\underline{\tilde{x}}) + b) \geq 1 - \xi_i$$

II <u>Slater condition 2:</u>

All affine equality and inequality constraints are fulfilled for at least one point $\tilde{x}$:

$$\forall_{i=1}^{N} \quad \xi_i \geq 0$$

II is easy to verify because we can simply choose all ξ_i such that $\xi_i \geq 0$ for all i.

I Similarly, $y_i(\underline{w}^T \phi(\underline{\tilde{x}}) + b) \geq 1 - \xi_i$ always holds if we choose all ξ_i very large: $\forall_{i=1}^{N} \quad \xi_i >> 1$.

1) The primal is given by $\min_{w,\beta,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$

We can add the constraints by making them zero:

$\forall_{i=1}^{N} \; y_i(\underline{w}^T \phi(\underline{x}_i) + b) \geq 1 - \xi_i$

$$\Leftrightarrow \quad 0 \geq 1 - \xi_i - y_i(\underline{w}^T \phi(\underline{x}_i) + b)$$

$$\forall_{i=1}^{N} \; \xi_i \geq 0 \quad \Leftrightarrow \quad 0 \geq -\xi_i$$

Therefore, the dual is obtained maximizing the Lagrange multipliers λ, λ' for both constraints on the primal:

$$\text{Dual} = \max_{\lambda,\lambda' \geq 0} \text{Primal} = \max_{\lambda,\lambda' \geq 0} \min_{w,\beta,\xi} \underbrace{J(w,\beta,\xi)}_{\text{objective}}$$

$$= \max_{\lambda,\lambda' \geq 0} \min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \lambda_i \left(1 - \xi_i - y_i(\underline{w}^T \phi(\underline{x}_i) + b)\right)$$

$$- \sum_{i=1}^{N} \lambda'_i \xi_i$$

We minimize the objective by taking the partial derivatives:

$$\frac{\partial J}{\partial w} = w - \sum_{i=1}^{N} \lambda_i y_i \phi(x_i) \overset{!}{=} 0 \quad \Leftrightarrow w = \sum_{i=1}^{N} \lambda_i y_i \phi(x_i)$$

$$\frac{\partial J}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i \overset{!}{=} 0$$

$$\frac{\partial J}{\partial \xi} = C - \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda'_i$$

$$\Leftrightarrow \frac{\partial J}{\partial \xi_i} = C - \lambda_i - \lambda'_i \quad \rightarrow \quad \underset{\text{(box constraint)}}{0 \leq \lambda_i \leq C}$$

Now we obtain the dual in the simplified form

$$\text{dual} = \max_{\lambda} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \phi(x_i)^T \phi(x_j) + \sum_{i=1}^{N} \lambda_i$$

with the constraints $\sum_{i=1}^{N} \lambda_i y_i = 0$ and $\forall_{i=1}^{N}\ 0 \le \lambda_i \le C$

c) From b) we have $w = \sum_{i=1}^{N} \lambda_i y_i \phi(\underline{x}_i)$ and we know

$\lambda_i (1 - \xi_i - y_i (w^T \phi(x_i) + b)) = 0$ and

$(C - \lambda_i)(-\xi_i) = 0$

Since $\lambda_i \neq C \Rightarrow \xi_i = 0$
and $\lambda_i \neq 0 \Rightarrow 1 - w^T \phi(x_i) - b = 0$ } if we define $y_i := 1$

Using $y_i := 1$ and the constraint of b) $0 < \lambda_i < C$ we obtain

$b = 1 - w^T \phi(\underline{x}_i)$

d) we are given the function $f(x)$ and insert w, b, and the kernel:

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

$$= \text{sign}\left(\sum_{i=1}^{N} \lambda_i y_i k(x_i, x) + 1 - \sum_{i=1}^{N} \lambda_i y_i k(x_i, x_{sv})\right)$$

2

a) In 1.b) we found

$$\max_{\lambda} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \phi(x_i)^T \phi(x_j) + \sum_{i=1}^{N} \lambda_i$$

$$\text{with } \sum_{i=1}^{N} \lambda_i y_i = 0 \quad \text{and} \quad 0 \leq \lambda_i \leq c$$

We can rewrite it such that

$$\min_{\lambda} \frac{1}{2} \sum \lambda_i y_i y_j k(x_i, x_j) \lambda_j - \sum_{i=1}^{N} \lambda_i$$

$$= \min_{\lambda} \frac{1}{2} \underline{\lambda}^T P_{ij} \underline{\lambda} - \mathbb{1} \cdot \underline{\lambda} \iff \min_{x} \frac{1}{2} \underline{x}^T \underline{\underline{P}} \underline{x} + \underline{q}^T \underline{x}$$

Therefore $\boxed{\underline{x} \mathrel{\hat{=}} \underline{\lambda}, \quad \underline{P} \mathrel{\hat{=}} y_i y_j k(x_i, x_j), \quad \underline{q}^T \mathrel{\hat{=}} -\mathbb{1}}$

The constraint $\sum_{i=1}^{N} \lambda_i y_i = 0$ simplifies to

$$\underline{\lambda}\, \underline{y}^T = \underline{Y}^T \underline{\lambda} = 0 \iff A\underline{x} = \underline{b}$$

Therefore $\boxed{A \mathrel{\hat{=}} Y^T, \quad \underline{b} \mathrel{\hat{=}} \underline{0}}$.

The constraint $0 \leq \lambda_i \leq c$ corresponds to

$$\left.\begin{matrix} -\lambda_i \leq 0 \\ \lambda_i \leq c \end{matrix}\right\} = \begin{pmatrix} -\mathbb{1} \\ \mathbb{1} \end{pmatrix} \lambda \preceq \begin{pmatrix} 0 \\ c \end{pmatrix} \mathbb{1}$$

Therefore, $\boxed{G \mathrel{\hat{=}} \begin{pmatrix} -\mathbb{1} \\ \mathbb{1} \end{pmatrix}, \quad h \mathrel{\hat{=}} \begin{pmatrix} 0 \\ c \end{pmatrix} \mathbb{1}}$