Exercises for the course
Machine Learning 2
Summer semester 2024

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 9

Exercise 1: Structured Prediction for Classification (20 P)

While structured output learning is typically used for predicting complex output signals such as sequences or trees, the same framework can also be used to address the more standard problem of classification. Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ be two data points and $y, y' \in \{1, \ldots, C\}$ their respective classes. Consider the structured output kernel

$$k_{\text{struct}}((\boldsymbol{x}, y), (\boldsymbol{x}', y')) = k(\boldsymbol{x}, \boldsymbol{x}') \cdot I(y = y'),$$

where $k(\boldsymbol{x}, \boldsymbol{x}')$ is a positive semi-definite kernel with associated feature map $\phi(\boldsymbol{x})$, and where $I(\cdot)$ is an indicator function that is 1 when the argument is true and 0 otherwise.

(a) *Show* that the kernel $k_{\text{struct}}((\boldsymbol{x}, y), (\boldsymbol{x}', y'))$ is positive semi-definite, that is, show that

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j \, k_{\text{struct}}((\boldsymbol{x}_i, y_i), (\boldsymbol{x}_j, y_j)) \geq 0$$

for all input/output pairs $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ and choice of real numbers $c_1, \ldots, c_N$.

(b) *Find* a feature map $\phi_{\text{struct}}(\boldsymbol{x}, y)$ associated this kernel, i.e. satisfying

$$\left\langle \phi_{\text{struct}}(\boldsymbol{x}, y), \phi_{\text{struct}}(\boldsymbol{x}', y') \right\rangle = k_{\text{struct}}((\boldsymbol{x}, y), (\boldsymbol{x}', y'))$$

for all pairs $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}', y')$.

Exercise 2: Dual Formulation of Structured Output Learning (20 P)

In structured output learning, we look for a linear model in joint feature space that produces a large margin between the correct prediction and all other possible predictions. The primal formulation of this problem can be expressed as:

$$\min_{\boldsymbol{w}, \xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{n=1}^{N} \xi_n$$

subject to the constraints:

$$\forall_{n=1}^{N} \forall_{\boldsymbol{y} \neq \boldsymbol{y}_n} : \ \boldsymbol{w}^\top \Psi_{n,\boldsymbol{y}} \geq 1 - \xi_n \qquad \text{and} \qquad \forall_{n=1}^{N} : \ \xi_n \geq 0$$

where we have used the shortcut notation $\Psi_{n,\boldsymbol{y}} = \phi(\boldsymbol{x}, \boldsymbol{y}_n) - \phi(\boldsymbol{x}, \boldsymbol{y})$.

(a) *Show* that the associated dual optimization problem is given by:

$$\max_{\boldsymbol{\alpha}} \ \sum_{n=1}^{N} \sum_{\boldsymbol{y} \neq \boldsymbol{y}_n} \alpha_{n,\boldsymbol{y}} \ - \ \frac{1}{2} \sum_{n,n'} \sum_{\substack{\boldsymbol{y} \neq \boldsymbol{y}_n \\ \boldsymbol{y}' \neq \boldsymbol{y}_{n'}}} \alpha_{n,\boldsymbol{y}} \alpha_{n',\boldsymbol{y}'} \langle \Psi_{n,\boldsymbol{y}}, \Psi_{n',\boldsymbol{y}'} \rangle$$

subject to the constraints:

$$\forall_{n=1}^{N} \forall_{\boldsymbol{y} \neq \boldsymbol{y}_n} : \ 0 \leq \alpha_{n,\boldsymbol{y}} \qquad \text{and} \qquad \forall_{n=1}^{N} : \ \sum_{\boldsymbol{y} \neq \boldsymbol{y}_n} \alpha_{n,\boldsymbol{y}} \leq C$$

(b) Assuming $k((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}'))$ is the kernel that induces the feature map $\phi(\boldsymbol{x}, \boldsymbol{y})$, *express* the dot product $\langle \Psi_{n,\boldsymbol{y}}, \Psi_{n',\boldsymbol{y}'} \rangle$ in terms of this kernel.

Exercise 3: Prediction of Output Sequences (20 P)

Consider output sequences to predict to be of the type $\boldsymbol{y} \in \{-1, 1\}^L$, and the feature map:

$$\phi(\boldsymbol{x}, \boldsymbol{y}) = \left[\boldsymbol{x} \odot \boldsymbol{y} \ , \ 2 \cdot (\boldsymbol{y}_{1\ldots L-1} \odot \boldsymbol{y}_{2\ldots L})\right]$$

where $\odot$ denotes the element-wise product between two vectors. The structured output model looks for the output $\boldsymbol{y}$ that maximizes the matching function, i.e.

$$\max_{\boldsymbol{y}} \ \boldsymbol{w}^\top \phi(\boldsymbol{x}, \boldsymbol{y})$$

with $\boldsymbol{w} \in \mathbb{R}^{2L-1}$. In the following, we assume that $L = 3$, that the current parameter is $\boldsymbol{w} = (1, 1, 1, 1, 1)$ and that we receive the input $\boldsymbol{x} = (1, -1, 1)$.

(a) *Show* that the problem of maximizing the matching function simplifies to:

$$\max_{y_1 \in \{-1,1\}} \left\{ y_1 + \max_{y_2 \in \{-1,1\}} \left\{ 2y_1y_2 - y_2 + \max_{y_3 \in \{-1,1\}} \left\{ 2y_2y_3 + y_3 \right\} \right\} \right\}$$

(b) *Find* using the Viterbi procedure the best output (y_1, y_2, y_3), that is, solve $\max_{y_3}\{\}$ for every y_2, then solve $\max_{y_2}\{\}$ for every y_1, and then solve $\max_{y_1}\{\}$. While doing so, keep track of the values in the sequence that have produced the respective maximums so that the optimal sequence can be reconstructed.

Exercise 4: Programming (40 P)

Download the programming files on ISIS and follow the instructions.