EPFL | MGT-418 : Convex Optimization | Project 2

Learning the Kernel Function (graded)

(This project is due on **November 24, 2021, at 23:59**. You may form teams of up to two people. Each team should upload a single zip-file containing their report and code to Moodle. Make sure to clearly state the team members in your report.)

Description

Given training samples $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, m$, consisting of inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ and outputs $y_i \in \{-1, 1\}$,[1] we aim to predict the outputs of test samples based on their inputs by using a linear classifier. As the training samples may not be linearly separable, we use a feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^D$ that lifts the inputs to a higher-dimensional space $\mathbb{R}^D$, and we solve the soft-margin support vector machine problem in the lifted space. Specifically, we solve the problem

$$\begin{array}{ll} \underset{\boldsymbol{w}\in\mathbb{R}^D, \boldsymbol{s}\in\mathbb{R}^m, b\in\mathbb{R}}{\text{minimize}} & \sum_{i=1}^m s_i + \frac{\rho}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subject to} & y_i(\boldsymbol{w}^\top \Phi(\boldsymbol{x}_i) + b) + s_i \geq 1 \quad \forall i = 1, \ldots, m \\ & s_i \geq 0 \qquad\qquad\qquad\quad \forall i = 1, \ldots, m \end{array} \tag{1}$$

for some $\rho > 0$. The corresponding dual problem (see lecture slide 183) can be expressed as

$$\begin{array}{ll} \underset{\boldsymbol{\lambda}\in\mathbb{R}^m}{\text{maximize}} & \sum_{i=1}^m \lambda_i - \frac{1}{2\rho} \sum_{i,j=1}^m \lambda_i \lambda_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{subject to} & \sum_{i=1}^m y_i \lambda_i = 0 \\ & 0 \leq \lambda_i \leq 1 \quad \forall i = 1, \ldots, m, \end{array} \tag{2}$$

where $k(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\boldsymbol{x})^\top \Phi(\boldsymbol{x}')$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ denotes the kernel function. In contrast to the primal problem (1), the dual problem (2) can be solved without knowledge of the feature map Φ. Instead, it suffices to know the kernel function. Gaussian, polynomial, exponential, and tangential kernel functions are commonly used, but it is not always clear which function works best for a given problem. In practice, kernel functions are often chosen through cross-validation, that is, by solving problem (2) repeatedly for multiple kernels and regularization parameters. In this project, instead, we formulate a single optimization problem that determines both the optimal classifier and the optimal kernel function. To this end, we first choose L candidate kernel functions $\hat{k}^1, \ldots, \hat{k}^L$, and then we form the candidate kernel matrices $\hat{K}^1, \ldots, \hat{K}^L$ corresponding to the training samples, that is, we set $\hat{K}^l_{ij} = \hat{k}^l(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for all $i, j = 1, \ldots, m$ and $l = 1, \ldots, L$. Next, we define the set

$$\mathcal{K} = \left\{ K \in \mathbb{S}^m_+ : K = \sum_{l=1}^L \mu_l \hat{K}^l,\ K \succeq 0,\ \mathrm{tr}(K) = c,\ \mu_l \geq 0 \ \ \forall l = 1, \ldots, L \right\}$$

of conic combinations of the candidate kernel matrices $\hat{K}^1, \ldots, \hat{K}^L$, whose trace equals $c > 0$. We now express the optimal value of problem (2) in vectorized form as a function ω of the kernel matrix K,

$$\omega(K) = \left\{ \begin{array}{ll} \underset{\boldsymbol{\lambda}\in\mathbb{R}^m}{\max} & \boldsymbol{\lambda}^\top \mathbf{1} - \frac{1}{2\rho} \boldsymbol{\lambda}^\top G(K) \boldsymbol{\lambda} \\ \text{s.t.} & \boldsymbol{\lambda}^\top \boldsymbol{y} = 0 \\ & 0 \leq \lambda_i \leq 1 \quad \forall i = 1, \ldots, m, \end{array} \right. \tag{3}$$

where $G(K) \in \mathbb{S}^m$ is the matrix with entries $G_{ij}(K) = K_{ij} y_i y_j$ for all $i, j = 1, \ldots, m$. We may thus determine both the optimal classifier and the optimal kernel by solving the optimization problem

$$\underset{K\in\mathcal{K}}{\text{minimize}} \quad \omega(K). \tag{4}$$

[1] We assume that there is at least one sample with output -1 and at least one sample with output 1.

Questions

1. **(10 points)** Explain why every kernel matrix induced by any feature map Φ is positive semidefinite.
2. **(25 points)** Show that the minimization problem in (4) is equivalent to the convex maximization problem

$$\begin{array}{lll} \underset{\boldsymbol{\lambda}\in\mathbb{R}^m, z\in\mathbb{R}}{\text{maximize}} & \boldsymbol{\lambda}^T\mathbf{1} - cz & \\ \text{subject to} & z\hat{r}_l \geq \frac{1}{2\rho}\boldsymbol{\lambda}^\top G(\hat{K}^l)\boldsymbol{\lambda} & l = 1,\ldots,L \\ & \boldsymbol{\lambda}^\top \boldsymbol{y} = 0 & \\ & 0 \leq \lambda_i \leq 1 & i = 1,\ldots,m, \end{array} \tag{5}$$

   where the vector $\hat{\boldsymbol{r}} \in \mathbb{R}^L$ has entries $\hat{r}_l = \operatorname{tr}(\hat{K}^l)$ for all $l = 1,\ldots,L$. *Hint: You may want to use the following minimax theorem. If some function $f : U \times W \to \mathbb{R}$ is convex in $u \in U$ and concave in $v \in V$, and if U and V are convex, closed, and bounded sets, then*

$$\min_{u\in U}\max_{v\in V} f(u,v) = \max_{v\in V}\min_{u\in U} f(u,v).$$

3. **(20 points)** When solving problem (5) numerically, most solvers will not only determine the optimal primal decisions $\boldsymbol{\lambda}^\star$ and $z^\star$ but also the optimal dual decisions. For $l = 1,\ldots,L$, let $\mu_l^\star$ denote the optimal dual variable of the constraint $z\hat{r}_l \geq \frac{1}{2\rho}\boldsymbol{\lambda}^\top G(\hat{K}^l)\boldsymbol{\lambda}$. We will show that $K^\star = \sum_{l=1}^L \mu_l^\star \hat{K}^l$ solves problem (4) by first reformulating problem (5) as

$$\begin{array}{lll} \underset{\boldsymbol{\lambda}\in\mathbb{R}^m_+, \boldsymbol{\lambda}^\top\boldsymbol{y}=0, \boldsymbol{\lambda}\leq\mathbf{1}}{\max} & \max\limits_{z\in\mathbb{R}} & \boldsymbol{\lambda}^T\mathbf{1} - cz \\ & \text{s.t.} & z\hat{r}_l \geq \frac{1}{2\rho}\boldsymbol{\lambda}^\top G(\hat{K}^l)\boldsymbol{\lambda} \quad l = 1,\ldots,L. \end{array} \tag{6}$$

   (a) **(10 points)** Derive the Lagrangian dual of the inner maximization problem in (6). Explain why strong duality holds. Denote the dual variables by $\boldsymbol{\mu} \in \mathbb{R}^L$.

   (b) **(10 points)** Reformulate the max-min problem derived in (a) as a min-max problem by using the minimax theorem from (2) and denote its solution by $\boldsymbol{\mu}^\star$. Show that $K^\star = \sum_{l=1}^L \mu_l^\star \hat{K}^l$ solves (4).

4. **Radar Classification (30 points):** A lab in Goose Bay, Labrador, analyzes the ionosphere through radar signals. All signals are classified as either suitable or unsuitable for further analysis. We are given 351 classified signals with 33 features and asked to find a linear classifier for automatic signal classification.

   (a) **(5 points)** Read the data file `ionosphere.data` into memory by using the scripts `read_data.py` or `read_data.m`. Use the code skeletons `main.ipnyb` or `main.m` to randomly select 80% of the data for training.

   (b) **(15 points)** Define $\hat{k}^1(\boldsymbol{x},\boldsymbol{x}') = (1 + \boldsymbol{x}^\top\boldsymbol{x}')^p$ as the polynomial, $\hat{k}^2(\boldsymbol{x},\boldsymbol{x}') = \exp(-(\boldsymbol{x}-\boldsymbol{x}')^\top(\boldsymbol{x}-\boldsymbol{x}')/(2\sigma))$ as the Gaussian, and $\hat{k}^3(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{x}^\top\boldsymbol{x}'$ as the linear kernel function, and construct the kernel matrices $\hat{K}^l$, $l = 1,2,3$, for all training samples. Solve the QCQP in (5) for $\rho = 2$, $p = 2$, $\sigma = 0.5$ and $c = \sum_{l=1}^3 \operatorname{tr}(\hat{K}^l)$ with CVXPY and MOSEK in Python or with YALMIP and GUROBI in MATLAB, and record the optimal dual variables $\mu_1^\star, \mu_2^\star$, and $\mu_3^\star$. Use the code skeletons `kernel_learning` (in `main.ipynb`) or `kernel_learning.m`.

   (c) **(10 points)** Let $k^\star(\boldsymbol{x},\boldsymbol{x}') = \mu_1^\star\hat{k}^1(\boldsymbol{x},\boldsymbol{x}') + \mu_2^\star\hat{k}^2(\boldsymbol{x},\boldsymbol{x}') + \mu_3^\star\hat{k}^3(\boldsymbol{x},\boldsymbol{x}')$ for all $\boldsymbol{x},\boldsymbol{x}' \in \mathbb{R}^d$. As discussed in the lecture (slide 183), the label of any test sample $\boldsymbol{x} \in \mathbb{R}^d$ is predicted as $\operatorname{sign}(\frac{1}{\rho}\sum_{i=1}^m \lambda_i^\star y_i k^\star(\boldsymbol{x}_i,\bar{\boldsymbol{x}}) + b^\star)$, where $\boldsymbol{\lambda}^\star$ is an optimizer of (5) and $b^\star$ is the optimal dual variable of the constraint $\boldsymbol{\lambda}^\top\boldsymbol{y} = 0$ in (5). Use the code skeletons `SVM_predict` (in `main.ipynb`) or `SVM_predict.m`.

5. **(5 points)** Repeat the steps 4(a)–(c) 100 times with different seeds for the random partition of the data into training and test sets, and report the average test accuracy (correct classification rate) to Table 1.
6. **(10 points)** For each of the 100 training and test sets constructed in 5., solve (2) using the kernels functions $\hat{k}^1$, $\hat{k}^2$ and $\hat{k}^3$, respectively, and report the average test accuracies in Table 1. Use the code skeletons `svm_fit` (in `main.ipynb`) or `svm_predict.m`.

Kernel function	$\hat{k}^1$	$\hat{k}^2$	$\hat{k}^3$	$\sum_{l=1}^3 \mu_l^\star \hat{k}^l$
Average accuracy				

Table 1: Average test accuracies over 100 data splits.