ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

EPFL COURSEWORK REPORT

Project 2: Learning the Kernel Function

JIANHAO ZHENG (SCIPER: 323146)

YUJIE HE (SCIPER: 321657)

24th November 2021

All the equations in this report have no number. Equation with number, e.g. problem (5), refers to the formulation in the question pdf.

Question 1

With matrix $\hat{P}^l = [\hat{\Phi}^l(\boldsymbol{x}_1), \hat{\Phi}^l(\boldsymbol{x}_2), ..., \hat{\Phi}^l(\boldsymbol{x}_m)] \in \mathbb{R}^{D\times m}$ for any feature map $\hat{\Phi}^l$, the candidate kernel matrix $\hat{K}^l$ can be written as:

$$\hat{K}^l = (\hat{P}^l)^T \hat{P}^l$$

which implies $\hat{K}^l$ is positive semidefinite in any case.

Question 2

The problem in (4) can be rewritten as:

$$\min_{K\in\mathcal{K}} \max_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\lambda}^T\boldsymbol{y}=0,\, \boldsymbol{\lambda}\leq\mathbf{1}} \quad \boldsymbol{\lambda}^T\mathbf{1} - \frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda}.$$

The objective function $f(K, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^T\mathbf{1} - \frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda}$ is a linear transform of $\boldsymbol{\lambda}$ minus a quadratic transform of $\boldsymbol{\lambda}$. Thus, it is concave in $\boldsymbol{\lambda}$. Furthermore, the constraints on $\boldsymbol{\lambda}$ is the intersection of a hyperplane and a hypercube. We at least have $\boldsymbol{\lambda} = \mathbf{0}$ satisfying the constraints, which shows the feasible set is non-empty. Thus, the feasible set of $\boldsymbol{\lambda}$ is convex, closed and bounded sets.

The term $-\frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda} = -\frac{1}{2\rho}\sum_{i,j=1}^m \lambda_i\lambda_j y_i y_j k(\boldsymbol{x}_i\boldsymbol{x}_j)$ is a linear functions on K. Thus, $f(K, \boldsymbol{\lambda})$ is convex in K, which is proven to be positive semidefinite in Question 1. For any $K \in \mathcal{K}$, K is the conic combinations of several positive semidefinite matrices. Therefore, the constraint $K \succeq 0$ is implicitly satisfied. The feasible set $\mathcal{K}$ is the intersection of the conic combinations set and a hyperplane ($\mathrm{tr}(K) = c$ is equivalent to $\sum_{i=1}^m K_{ii} = c$, which is a linear constraint). For an arbitrary $t \in \{1, 2, ..., L\}$, let $\mu_t = \frac{c}{\mathrm{tr}(\hat{K}^t)} \geq 0$ as $c \geq 0$ and $\hat{K}^t \succeq 0$, we have $K' = \mu_t\hat{K}^t \in \mathcal{K}$ since $\mathrm{tr}(K') = \mu_t\mathrm{tr}(\hat{K}^t) = c$. This proves $\mathcal{K}$ is not empty. We can therefore conclude that the feasible set $\mathcal{K}$ is convex, closed and bounded sets.

Use the theorem from hint, the minimization problem in (4) is equivalent to:

$$\begin{aligned}
&\max_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\lambda}^T\boldsymbol{y}=0,\, \boldsymbol{\lambda}\leq\mathbf{1}} \min_{K\in\mathcal{K}} \quad \boldsymbol{\lambda}^T\mathbf{1} - \frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda} \\
&\qquad\qquad \Updownarrow \\
&\max_{\boldsymbol{\lambda}\in\mathbb{R}^m_+,\, \boldsymbol{\lambda}^T\boldsymbol{y}=0,\, \boldsymbol{\lambda}\leq\mathbf{1}} \max_{z\in\mathbb{R}} \quad \boldsymbol{\lambda}^T\mathbf{1} - cz \\
&\qquad\qquad\qquad \text{s.t.} \quad -cz \leq -\frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda}, \quad \forall K \in \mathcal{K} \\
&\qquad\qquad \Updownarrow
\end{aligned}$$

$$\begin{aligned}\max_{\boldsymbol{\lambda}\in\mathbb{R}^m, z\in\mathbb{R}} \quad & \boldsymbol{\lambda}^T\mathbf{1} - cz \\ \text{s.t.} \quad & cz \geq \frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda} \quad \forall K \in \mathcal{K} \\ & \boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ & 0 \leq \lambda_i \leq 1 \qquad\qquad i = 1, ..., m\end{aligned}$$

Only thing left is to show $cz \geq \frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda}, \forall K \in \mathcal{K} \iff z\hat{r}_l \geq \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda}, l = 1, ..., L$.

($\Longrightarrow$) Assume $cz \geq \frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda}$ holds for $\forall K \in \mathcal{K}$. Suppose there exists a $t \in \{1, 2, ..., L\}$ such that $z\hat{r}_t < \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^t)\boldsymbol{\lambda}$. For $K' = \frac{c}{tr(\hat{K}^t)}\hat{K}^t \in \mathcal{K}$, we have:

$$\frac{1}{2\rho}\boldsymbol{\lambda}^T G(K')\boldsymbol{\lambda} = \frac{c}{2\rho\, tr(\hat{K}^t)}\boldsymbol{\lambda}^T G(\hat{K}^t)\boldsymbol{\lambda} > cz$$

which shows contradiction.

($\Longleftarrow$) Assume $z\hat{r}_l \geq \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda}$ holds for $\forall l \in \{1, 2, ..., L\}$. Then, for $\forall K \in \mathcal{K}$, we have:

$$\begin{aligned}\frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda} &= \sum_{l=1}^{L}\frac{1}{2\rho}\mu_l\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda} \\ &\leq z\sum_{l=1}^{L}\mu_l tr(\hat{K}^l) \\ &= zc\end{aligned}$$

With all the proof above, we can conclude that the optimization problem in (4) is equivalent to the following convex max problem:

$$\begin{aligned}\max_{\boldsymbol{\lambda}\in\mathbb{R}^m, z\in\mathbb{R}} \quad & \boldsymbol{\lambda}^T\mathbf{1} - cz \\ \text{s.t.} \quad & z\hat{r}_l \geq \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda} \quad l = 1, ..., L \\ & \boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ & 0 \leq \lambda_i \leq 1 \qquad\qquad i = 1, ..., m\end{aligned}$$

Question 3

(**a**) The inner maximization problem in (6) is:

$$\begin{aligned}\max_{z\in\mathbb{R}} \quad & \boldsymbol{\lambda}^T\mathbf{1} - cz \\ \text{s.t.} \quad & z\hat{r}_l \geq \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda} \quad l = 1, ..., L\end{aligned}$$

We can derive the Lagrangian function for the above primal problem assuming $\mu_l \in \mathbb{R}_+$ denotes the dual variable of the constraint $z\hat{r}_l - \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda} \geq 0$ for $l = 1, ..., L$

$$L(z, \boldsymbol{\mu}) = \boldsymbol{\lambda}^T\mathbf{1} - cz + \sum_{l=1}^{L}\mu_l\left[z\hat{r}_l - \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda}\right]$$

$$= z\left(-c + \sum_{l=1}^{L} \mu_l \hat{r}_l\right) + \boldsymbol{\lambda}^T \mathbf{1} - \sum_{l=1}^{L} \frac{1}{2\rho} \mu_l \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda}$$

which is a linear function of z.

Thus the dual objective function is:

$$g(\boldsymbol{\lambda}) = \begin{cases} \boldsymbol{\lambda}^T \mathbf{1} - \sum_{l=1}^{L} \frac{1}{2\rho} \mu_l \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda}, & c = \sum_{l=1}^{L} \mu_l \hat{r}_l \\ +\infty, & \text{otherwise} \end{cases}$$

We can then derive the dual problem below:

$$\begin{aligned} \min_{\boldsymbol{\mu} \in \mathbb{R}^L} \quad & \boldsymbol{\lambda}^T \mathbf{1} - \sum_{l=1}^{L} \frac{1}{2\rho} \mu_l \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda} \\ \text{s.t.} \quad & c = \sum_{l=1}^{L} \mu_l \hat{r}_l \\ & \boldsymbol{\mu} \geq \mathbf{0} \end{aligned}$$

The primal problem is a linear programming, thus convex. We have

$$z' = \max_{l=1,\ldots,L} \left(\frac{1}{2\rho \hat{r}_l} \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda}\right) + 1,$$

which strictly satisfies the inequality constraint. This means the Slater's constraint qualification holds. Henece, strong duality holds for this problem

(**b**) The max-min problem is formulated as following:

$$\begin{aligned} \max_{\boldsymbol{\lambda} \in \mathbb{R}^m_+,\, \boldsymbol{\lambda}^T \boldsymbol{y} = 0,\, \boldsymbol{\lambda} \leq \mathbf{1}} \; \min_{\boldsymbol{\mu} \in \mathbb{R}^L} \quad & \boldsymbol{\lambda}^T \mathbf{1} - \sum_{l=1}^{L} \frac{1}{2\rho} \mu_l \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda} \\ \text{s.t.} \quad & c = \sum_{l=1}^{L} \mu_l \hat{r}_l \\ & \boldsymbol{\mu} \geq \mathbf{0} \end{aligned}$$

The objective function $f'(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{\lambda}^T \mathbf{1} - \sum_{l=1}^{L} \frac{1}{2\rho} \mu_l \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda}$ consists of a linear term and a negative quadratic transform on $\boldsymbol{\lambda}$, which is concave in $\boldsymbol{\lambda}$. We already show in Question 2 that the feasible set of $\boldsymbol{\lambda}$ is convex, closed and bounded.

$f'(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is a linear transform on $\boldsymbol{\mu}$, thus convex. The two constraints on $\boldsymbol{\mu}$ make the corresponding feasible set to be the intersection of halfspaces and a hyperplane. Hence, it is convex, closed and bounded.

All the statements above guarantee that the minmax theorem holds. The optimization problem can reformulated as:

$$\min_{\boldsymbol{\mu} \in \mathbb{R}^L} \; \max_{\boldsymbol{\lambda} \in \mathbb{R}^m_+,\, \boldsymbol{\lambda}^T \boldsymbol{y} = 0,\, \boldsymbol{\lambda} \leq \mathbf{1}} \quad \boldsymbol{\lambda}^T \mathbf{1} - \sum_{l=1}^{L} \frac{1}{2\rho} \mu_l \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda}$$

$$\text{s.t.} \quad c = \sum_{l=1}^{L} \mu_l \hat{r}_l$$
$$\boldsymbol{\mu} \geq \mathbf{0}$$

Based on the previous proof, the optimal value of problem (4) is equal to:

$$\begin{aligned}
\min_{K \in \mathcal{K}} \omega(K) &= \min_{K \in \mathcal{K}} \max_{\substack{\boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ \boldsymbol{\lambda} \leq \mathbf{1} \\ \boldsymbol{\lambda} \in \mathbb{R}_+^m}} \boldsymbol{\lambda}^T \mathbf{1} - \frac{1}{2\rho} \boldsymbol{\lambda}^T G(K) \boldsymbol{\lambda} \\
&= \max_{\substack{\boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ \boldsymbol{\lambda} \leq \mathbf{1} \\ \boldsymbol{\lambda} \in \mathbb{R}_+^m}} \max_{\substack{z \hat{r}_l \geq \frac{1}{2\rho} \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda} \\ \forall l = 1, \ldots, L \\ z \in \mathbb{R}}} \boldsymbol{\lambda}^T \mathbf{1} - cz \quad \text{(proven in question 2)} \\
&= \max_{\substack{\boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ \boldsymbol{\lambda} \leq \mathbf{1} \\ \boldsymbol{\lambda} \in \mathbb{R}_+^m}} \min_{\substack{\sum_{l=1}^{L} \mu_l \hat{r}_l = c \\ \boldsymbol{\mu} \in \mathbb{R}_+^L}} \boldsymbol{\lambda}^T \mathbf{1} - \sum_{l=1}^{L} \frac{1}{2\rho} \mu_l \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda} \quad \text{(strong duality)} \\
&= \min_{\substack{\sum_{l=1}^{L} \mu_l \hat{r}_l = c \\ \boldsymbol{\mu} \in \mathbb{R}_+^L}} \max_{\substack{\boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ \boldsymbol{\lambda} \leq \mathbf{1} \\ \boldsymbol{\lambda} \in \mathbb{R}_+^m}} \boldsymbol{\lambda}^T \mathbf{1} - \sum_{l=1}^{L} \frac{1}{2\rho} \mu_l \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda} \quad \text{(just proven)} \\
&= \min_{\substack{\sum_{l=1}^{L} \mu_l \hat{r}_l = c \\ \boldsymbol{\mu} \in \mathbb{R}_+^L}} \omega'(\boldsymbol{\mu}) = \omega'(\boldsymbol{\mu}^*)
\end{aligned}$$

Therefore, in order to prove $K^* = \sum_{l=1}^{L} \mu_l^* \hat{K}^l$ solves the problem, we just need to show $\omega(K^*) = \omega'(\boldsymbol{\mu}^*)$. This is quite obvious since we have:

$$\begin{aligned}
\omega(K^*) &= \max_{\substack{\boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ \boldsymbol{\lambda} \leq \mathbf{1} \\ \boldsymbol{\lambda} \in \mathbb{R}_+^m}} \boldsymbol{\lambda}^T \mathbf{1} - \frac{1}{2\rho} \boldsymbol{\lambda}^T G(K^*) \boldsymbol{\lambda} \\
&= \max_{\substack{\boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ \boldsymbol{\lambda} \leq \mathbf{1} \\ \boldsymbol{\lambda} \in \mathbb{R}_+^m}} \boldsymbol{\lambda}^T \mathbf{1} - \frac{1}{2\rho} \sum_{l=1}^{L} \mu_l^* \boldsymbol{\lambda}^T G(\hat{K}^l) \boldsymbol{\lambda} \\
&= \omega'(\boldsymbol{\mu}^*)
\end{aligned}$$

Question 4-6

The optimal dual variables computed by training data split with $randomseed = 0$ is reported as below:

$$\mu_1^* = 0.12, \quad \mu_1^* = 234.09, \quad \mu_1^* = 0.91, \quad b^* = 0.58$$

these values are also reported in the jupyter notebook.

The mean and standard deviation of accuracies on the test data with different kernel functions is reported in Table 1.

Kernal function	$\hat{k}^1$	$\hat{k}^2$	$\hat{k}^3$	$\sum_{l=1}^{3} \mu_l^* \hat{k}^l$
Accuracy	0.909 ± 0.034	0.899 ± 0.034	0.865 ± 0.037	0.937 ± 0.028

TABLE 1: Mean and standard deviation of test accuracies over 100 data splits

We further visualize the distribution of the accuracy over 100 data splits in Figure 1 where the learned optimal kernal outperforms others by a substantial margin.

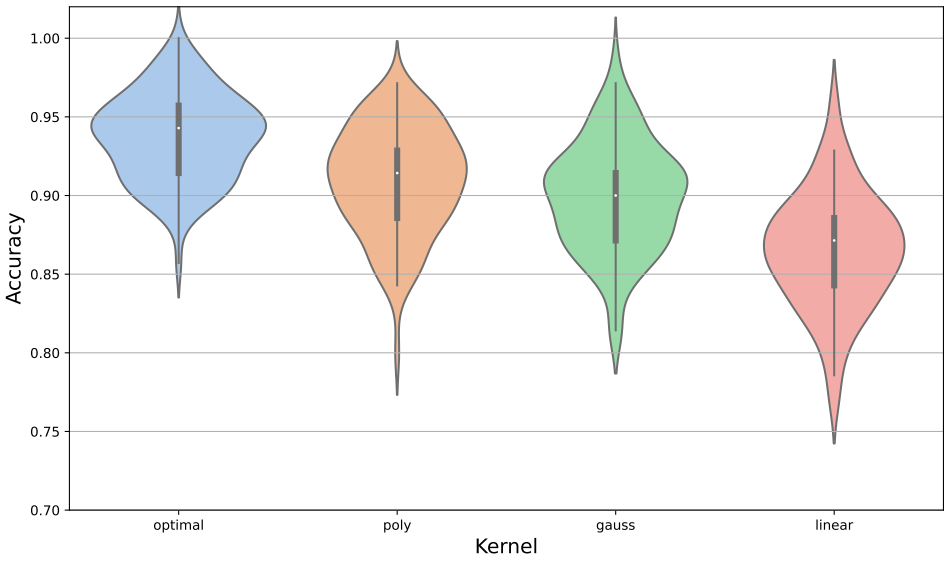

FIGURE 1: Accuracy distribution over 100 randomly repeated experiments