Kernel learning for data classification using support vector machine (SVM)

Authors:
Cédric Fontaine Sciper : 274871
Jérémie Arthur Maurice Pochon Sciper : 272761

24 nov 2021 Supervised by: Prof. Daniel Kuhn MGT-418

1 Introdction

The aim of this project is learn the kernel function for best classifying a radar dataset. This avoids one to have to perform grid-search for testing the different kernels to find the best one. For this, a convex - and thus both computationally efficient and globally optimal - optimisation problem will be formulated. The algorithm resulting performance will then be compared between the different kernels and the optimal one in order to ensure that it is indeed optimal.

First we will discuss the question related to the theoretical part then directly the results. The computational part should be understandable directly from the code attached to this work.

2 Question 1

The kernel matrix is defined as follow:

$$\mathcal{K}^L_{i,j} = K^L(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{1}$$

By definition of the positive semi definite (PSD) we have that: If there exists B such that: $M = B^T B$, then M is PSD (slide 14 - linear algebra review [1]). Hence, if we succeed in showing that the kernel matrix can be represented as such, it will ensured that it is PSD. Therefore, the kernel matrix can be re-expressed as shown in Equation 3.

$$\mathcal{K} = \begin{pmatrix} \phi(x_1)^T\phi(x_1) & \phi(x_1)^T\phi(x_2) & \dots & \phi(x_1)^T\phi(x_m) \\ \phi(x_2)^T\phi(x_1) & \phi(x_2)^T\phi(x_2) & \dots & \phi(x_1)^T\phi(x_m) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(x_m)^T\phi(x_1) & \phi(x_m)^T\phi(x_2) & \dots & \phi(x_m)^T\phi(x_m) \end{pmatrix} \tag{2}$$

$$= \begin{pmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_m)^T \end{pmatrix} \begin{pmatrix} \phi(x_1) & \phi(x_2) & \vdots & \phi(x_m) \end{pmatrix} = \phi(X)^T \phi(X) \tag{3}$$

It can clearly be seen that the kernel matrix is indeed the product of two identical matrices with one of those being transposed.

3 Question 2 : min-max transformation to maximization problem

Demonstration that we can use the *minimax Theorem* :

We start with a $min - max$ problem, as stated in Equation 4.

$$\begin{aligned} \min_{K \in \mathcal{K}} \max_{\lambda \in \mathbb{R}^m} \quad & \boldsymbol{\lambda}^T \mathbf{1} - \frac{1}{2\rho} \boldsymbol{\lambda}^T G(K) \boldsymbol{\lambda} \\ \text{s.t.} \quad & \boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ & 0 \leq \lambda_i \leq 1 \qquad \forall i = 1, \ldots, m \end{aligned} \tag{4}$$

In order to use the *minimax Theorem*, one must first show that the problem to be optimised is convex in the variable $K \in \mathcal{K}$ as well as concave in λ within its domain. Also, the domains for both K and λ must be closed, convex and bounded.

First, the objective is indeed convex in K since it depends linearly on the PSD matrix K. Its domain contains two constraints : K must be PSD (but this was demonstrated to always be the case for kernel matrices in question 1.) and $tr(K) = c$. Since the trace of a matrix is a convex function of this matrix (cf. slide 52 of the course), the set $\mathcal{K}$ is convex. It is also closed since it indeed contains its boundary (strict equality of the constraint) and it is thus bounded (only a finite number of matrices will satsisfy this equality).

Secondly, the part of the objective $\overbrace{-\overbrace{\frac{1}{2\rho} \boldsymbol{\lambda}^T G(K) \boldsymbol{\lambda}}^{convex}}^{negative\ of\ a\ convex\ function}$ is concave since the inner quadratic form is convex (cf. slide 59 of the course) in λ and taking the negative of any convex function make it then concave. The part $\boldsymbol{\lambda}^T \mathbf{1}$ is affine in λ and is thus both convex and concave. Therefore, the objective conisists in the sum of two concave functions of λ and is thus concave in λ. The feasible set of λ is made of one affine equality (a hyperplane) intersecting with $2m$ inequalities (half-spaces). Since the intersection of half-spaces with hyperplanes is convex (still a hyperplane, but restricted), this set is convex. It is bounded since each coordinate of the λ vector is upper- and lower-bounded, and since each of these is not strict, it is also closed.

Now, we can use the *minimax* Theorem to invert the order of the optimisation problem :

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \min_{K \in \mathcal{K}} \quad & \boldsymbol{\lambda}^T \mathbf{1} - \frac{1}{2\rho} \boldsymbol{\lambda}^T G(K) \boldsymbol{\lambda} \\ \text{s.t.} \quad & \boldsymbol{\lambda}^T \boldsymbol{y} = 0 \\ & 0 \leq \lambda_i \leq 1 \qquad \forall i = 1, \ldots, m \end{aligned} \tag{5}$$

In the aim of removing the inner minimisation problem, an epigraphical variable $z \in \mathrm{R}$ can be used. Since the original inner problem is a minimisation, z shall be lower-bounded, so that at optimality its value is the same as the term it is supposed to replace, as will be demonstrated below. Since this removes the variable K from the objective, the inner minimisation problem does not hold anymore and the epigraphical variable z becomes thus a maximiser of the problem, as is λ. But we will first slightly reformulate the right part of the objective.

$$\begin{aligned}
\frac{1}{2\rho}\boldsymbol{\lambda}^T G(K)\boldsymbol{\lambda} &= \frac{1}{2\rho}\boldsymbol{\lambda}^T\{\sum_{l=1}^{3}\mu_l G(\hat{K}^l)\}\boldsymbol{\lambda} && (6)\\
&= \frac{1}{2\rho}\{\sum_{l=1}^{3}\mu_l \boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda}\} && (7)\\
&= \frac{\mu_1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^1)\boldsymbol{\lambda} + \frac{\mu_2}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^2)\boldsymbol{\lambda} + \frac{\mu_3}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^3)\boldsymbol{\lambda} && (8)\\
&= \mu_1 tr(\hat{K}^1)z + \mu_2 tr(\hat{K}^2)z + \mu_3 tr(\hat{K}^3)z && (9)\\
&= cz && (10)
\end{aligned}$$

From Equation 10 one can observe that the term $c = tr(K) = \sum_{l=1}^{3}\mu_l tr(\hat{K})^l$ has appeared, and the epigraphical variable z will be constrained, as shown below.

$$\begin{aligned}
\max_{\lambda\in\mathbb{R}^m}\quad & \boldsymbol{\lambda}^T\mathbf{1} - cz \\
\text{s.t.}\quad & z\hat{r}_l \le \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda} \qquad \forall l = 1,...,3 \\
& \boldsymbol{\lambda}^T\boldsymbol{y} = 0 \\
& 0 \le \lambda_i \le 1 \qquad\qquad \forall i = 1,\dots,m
\end{aligned} \tag{11}$$

Hence, since this is a maximisation problem, optimality would yield $z\hat{r}_l = \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda}$ $\forall l = 1,...,3$, which can be replaced as it is in Equation 9, which is also the objective of 11. A last step can be perform in order to move the constraints on λ directly on constraints on the domain of λ.

$$\begin{aligned}
\max_{\lambda\in\mathbb{R}^m_+,\boldsymbol{\lambda}^T\boldsymbol{y}=0,\boldsymbol{\lambda}\le 1}\quad & \boldsymbol{\lambda}^T\mathbf{1} - cz \\
\text{s.t.}\quad & z\hat{r}_l \le \frac{1}{2\rho}\boldsymbol{\lambda}^T G(\hat{K}^l)\boldsymbol{\lambda} \quad \forall l = 1,...,3
\end{aligned} \tag{12}$$

4 Question 3 : The dual problem

a) The Lagrangian of the problem yields:

$$\mathcal{L} = \boldsymbol{\lambda}^T \mathbf{1} - cz + \sum_{i=1}^{L} \mu_i \left(\frac{1}{2\rho} \boldsymbol{\lambda}^T G(K^i) \boldsymbol{\lambda} - z\hat{r}_i \right) \tag{13}$$

We define the dual objective function g as:

$$g(\lambda, \mu) = inf_{z \in \mathbb{R}}(\mathcal{L}(z, \lambda, \mu)) \tag{14}$$

By derivation we obtain that:

$$\frac{\partial \mathcal{L}}{\partial z} = -c - \sum_{i=1}^{L} \mu_i \hat{r}_i = 0 \tag{15}$$

So the dual problem of the inner maximization yields, with cz and $z\hat{r}_i$ cancelling each other:

$$\begin{aligned} \min_{\mu \in \mathbb{R}} \quad & g(\lambda, \mu) = \boldsymbol{\lambda}^T \mathbf{1} + \sum_{i=1}^{L} \mu_i \left(\frac{1}{2\rho} \boldsymbol{\lambda}^T G(K^i) \boldsymbol{\lambda} \right) \\ \text{s.t.} \quad & \mu \geq 0 \end{aligned} \tag{16}$$

To prove that the strong duality holds, we will use the Slater condition (Slide 115 - Lagrangian duality [1]). We know that strong duality holds for convex problem if there exists a $\boldsymbol{z_s}$ satisfying the constraints of the primal problem. In our case, the epigraphical variable (see part 3) $\boldsymbol{z_s}$ is only bounded from below by $\frac{1}{2\rho} \boldsymbol{\lambda}^T G(K^i) \boldsymbol{\lambda}$, this means that we can always find a $\boldsymbol{z_s}$ that is greater than the constraints. It will lead to a solution far from the optimum, but be feasible nonetheless. Note that this holds because z_s is an epigraphical variable, which means that its purpose was to replace a minimisation problem by several inequalities such that, at optimality, these become equalities. But since this is only valid at optimality, one can deduce that there must exist other non-optimal values of z_s which make the inequalities strictly hold.

b) To use Sion's theorem, we first need to validate its hypothesis. First we trivially have that the objective function is both convex and concave in μ as it consists of the sum of affine functions $\sum_{i=1}^{L} \mu_i \left(\frac{1}{2\rho} \boldsymbol{\lambda}^T G(K^i) \boldsymbol{\lambda} \right)$. In $\boldsymbol{\lambda}$, the function is indeed convex as proven earlier. Therefore we have a convex and a concave functions to satisfy the first hypothesis. Now we need to show that the set:

$$\mathcal{C} = \{ \lambda \in \mathbb{R}_+^m \quad s.t. \quad \lambda^T y = 0, \quad \lambda \leq 1 \} \tag{17}$$

is convex, closed and bounded. This set can be seen as the intersection of the space $0 \leq \lambda \leq 1$ and the hyper-plane $\lambda^T y = 0$ which are both convex, closed and bounded. The intersection is therefore convex, closed and bounded.

The feasible set for μ_i is $\hat{r}^T \mu = -c$ is a hyper-plane bounded below by $\mu_i \geq 0$. The hyper-plane is convex and close but not bounded in itself. Below, μ is bounded by the condition $\mu \geq 0$. To have it bounded above we can rewrite the problem with a finite upper bound. We have that the two problems are equivalent because μ multiplies a positive scalar. We know that $\frac{1}{2\rho}\boldsymbol{\lambda}^T G(K^i)\boldsymbol{\lambda}$ is positive because $G(K)$ is PSD. In this case minimizing over μ will always leads to μ being the smaller feasible value. Therefore, adding an upper bound greater than any feasible value will not change the minimisation problem.

We can therefore apply Sion's theorem and finally obtain:

$$\begin{aligned} \min_{\mu \in \mathbb{R}^+} \quad & \max_{\lambda \in \mathbb{R}^+, \lambda \leq 1} \quad g(\lambda, \mu) = \boldsymbol{\lambda}^T \mathbf{1} + \sum_{i=1}^{L} \mu_i \left(\frac{1}{2\rho} \boldsymbol{\lambda}^T G(K^i) \boldsymbol{\lambda} \right) \\ & \text{s.t.} \qquad \lambda^T y = 0 \end{aligned} \tag{18}$$

5 Results

The results obtains after 100 iterations are shown in table 1.

Table 1: Average accuracies of the difference kernel function and the optimal kernel function combination

Kernel function	$\sum_{l=1}^{L} \mu_l^* \hat{k}^l$	$\hat{k}^1$	$\hat{k}^2$	$\hat{k}^3$
Average accuracy	90.5%	90.9%	63.8%	86.4%

Bibliography

Courses

[1] KUHN, Daniel.(2021). Convex optimisation, EPFL.