Classification with Smooth Hinge Loss

MGT-418 : Convex Optimization

Authors:

Vibhu Baibhav
344068

Jiří Bláha
415639

Javier de Ramón Murillo
314797

Renuka Singh Virk
326470

November 2025

1 QCQP

1.1 Equaling $h(z)$ with $L(z)$

We begin with the definition of the infimal convolution for the given functions $f(z) = z^2/2$ and $g(z) = \max\{0, 1 - z\}$. Specifically, it is defined as

$$h(z) = \inf_{t\in\mathbb{R}} \left(f(t) + g(z - t)\right) = \inf_{t\in\mathbb{R}} \left(\frac{1}{2}t^2 + \max\{0, 1 - (z - t)\}\right)$$

and can be rewritten as

$$\begin{aligned} \underset{t,u\in\mathbb{R}}{\text{minimize}} \quad & \frac{1}{2}t^2 + u \\ \text{subject to} \quad & u \geq 0, \\ & u \geq 1 - z + t, \end{aligned}$$

which is equivalent. Its Lagrangian reads

$$\mathcal{L}(t, u, \lambda, \gamma) = \frac{1}{2}t^2 + u - \lambda u + \gamma(1 - z + t - u)$$

and if we group by terms, we obtain

$$\mathcal{L}(t, u, \lambda, \gamma) = \frac{1}{2}t^2 + \gamma t + \gamma(1 - z) + u(1 - \lambda - \gamma),$$

for $\lambda \geq 0$, $\gamma \geq 0$. The dual objective is

$$g(\lambda, \gamma) = \inf_{t,u\in\mathbb{R}} \mathcal{L}(t, u, \lambda, \gamma),$$

and the Lagrangian is separable in the decision variables. This in turn allows us to go through the decision variables, finding the infimum of the Lagrangian with respect to that specific decision variable, and then combine all the infima together to form the dual objective $g(\lambda, \gamma)$.

1.1.1 Infimum over u

Notice $u(1 - \lambda - \gamma)$ is affine in u, therefore

$$\inf_{u\in\mathbb{R}} u(1 - \lambda - \gamma) = \begin{cases} 0 & \text{if } \lambda + \gamma = 1, \\ -\infty & \text{otherwise,} \end{cases}$$

which implies $\lambda, \gamma \in [0, 1]$, since both $\lambda \geq 0$, $\gamma \geq 0$.

1.1.2 Infimum over t

As $t^2/2 + \gamma t$ is quadratic in t, we have

$$\nabla_t \left(\frac{1}{2}t^2 + \gamma t\right) = t + \gamma = 0 \Leftrightarrow t^\star = -\gamma,$$

and also

$$\nabla_t^2 \left(\frac{1}{2}t^2 + \gamma t\right) = 1 > 0,$$

hence $t^2/2 + \gamma t$ convex in t.

1.1.3 Lagrangian Dual

Our previous results yield the dual objective

$$g(\gamma) = -\frac{1}{2}\gamma^2 + \gamma(1-z),$$

and the feasible set is bounded, i.e., $0 \leq \gamma \leq 1$, and the dual objective is a continuous concave function, we can invoke the equivalence

$$\sup_{0\leq\gamma\leq 1} g(\gamma) = \max_{0\leq\gamma\leq 1} g(\gamma),$$

and hence the dual problem is

$$\begin{aligned} \underset{\gamma\in\mathbb{R}}{\text{maximize}} \quad & -\frac{1}{2}\gamma^2 + \gamma(1-z) \\ \text{subject to} \quad & 0 \leq \gamma \leq 1, \end{aligned}$$

which we can decouple to three cases, as shown next.

1.1.4 Establishing Equality

Finding the unconstrained optimum of the dual yields

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\left(-\frac{1}{2}\gamma^2 + \gamma(1-z)\right) = -\gamma + (1-z) = 0 \Leftrightarrow \gamma^\star = 1 - z,$$

and can be broken up into three cases, i.e.:

1. If $z \geq 1$, then $\gamma^\star \leq 0$, i.e., it lies on the left boundary or outside of the feasible set, and the dual optimum is therefore $g(0) = 0$.
2. If $0 < z < 1$, then $\gamma^\star \in (0, 1)$, i.e., it lies within the interior of the feasible set, and the dual optimum is therefore $g(1-z) = (1-z)^2/2$.
3. If $z \leq 0$, then $\gamma^\star \geq 1$, i.e., it lies on the right boundary or outside the feasible set, and the dual optimum is therefore $g(1) = 1/2 - z$.

Additionally, since the primal problem is convex with linear constraints and satisfies Slater's condition (by setting $t = 0$, and assuming z is finite, we can choose u big enough so that in all three cases mentioned above, the inequalities are strict inequalities), strong duality holds. Therefore

$$h(z) = \begin{cases} 1/2 - z & \text{if } z \leq 0, \\ (1-z)^2/2 & \text{if } 0 < z < 1, \\ 0 & \text{if } z \geq 1, \end{cases}$$

which is exactly the smooth hinge loss function $L(z)$.

1.2 QCQP Reformulation

Starting from

$$\underset{w\in\mathbb{R}^d,\, b\in\mathbb{R}}{\text{minimize}} \quad \frac{1}{m}\sum_{i=1}^{m} L(y_i(w^T x_i - b)) + \frac{\varrho}{2}||w||_2^2,$$

we introduce epigraphical variables $s_i \in \mathbb{R}, s_i \geq L(y_i(w^T x_i - b))$, for all $i = 1, \ldots, m$, and rewrite the aforementioned optimization problem as

$$\begin{aligned} \underset{w\in\mathbb{R}^d,\, b\in\mathbb{R}}{\text{minimize}} \quad & \frac{1}{m}\sum_{i=1}^{m} s_i + \frac{\varrho}{2}\|w\|_2^2 \\ \text{subject to} \quad & L(y_i(w^T x_i - b)) \leq s_i, \quad \forall i = 1, \ldots, m. \end{aligned}$$

We have established before that

$$L(y_i(w^T x_i - b)) = \inf_{t_i \in \mathbb{R}} \left(\frac{1}{2} t_i^2 + \max\{0, 1 - (y_i(w^T x_i - b) - t_i)\} \right),$$

which allows us to rewrite the constraint as

$$\inf_{t_i \in \mathbb{R}} \left(\frac{1}{2} t_i^2 + \max\{0, 1 - (y_i(w^T x_i - b) - t_i)\} \right) \leq s_i, \quad \forall i = 1, \ldots, m.$$

By definition of the infimum, we can rewrite the above to

$$\frac{1}{2} t_i^2 + \max\{0, 1 - (y_i(w^T x_i - b) - t_i)\} \leq s_i, \quad \forall i = 1, \ldots, m,$$

and furthermore note that $\max\{a, b\} \leq c \Leftrightarrow (a \leq c \wedge b \leq c)$. When applied to the above inequality, it leads to the following two constraints:

$$\begin{aligned} \frac{1}{2} t_i^2 + 1 - (y_i(w^T x_i - b) - t_i) &\leq s_i, \qquad \forall i = 1, ..., m, \\ \frac{1}{2} t_i^2 &\leq s_i, \qquad \forall i = 1, ..., m. \end{aligned}$$

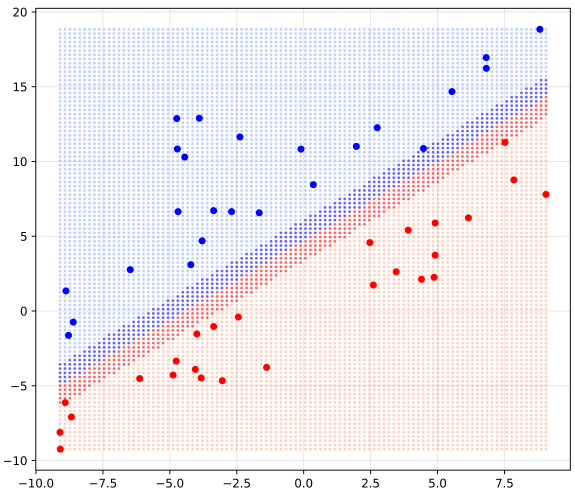
Finally, we can rewrite the initial problem as

$$\begin{aligned} \underset{w \in \mathbb{R}^d,\, b \in \mathbb{R},\, t, s \in \mathbb{R}^m}{\text{minimize}} \quad & \frac{1}{m} \sum_{i=1}^{m} s_i + \frac{\varrho}{2} \|w\|_2^2 \\ \text{subject to} \quad & \frac{1}{2} t_i^2 + 1 - y_i\big(w^T x_i - b\big) + t_i \leq s_i, \quad \forall i = 1, \ldots, m, \\ & \frac{1}{2} t_i^2 \leq s_i, \qquad \forall i = 1, \ldots, m, \end{aligned}$$

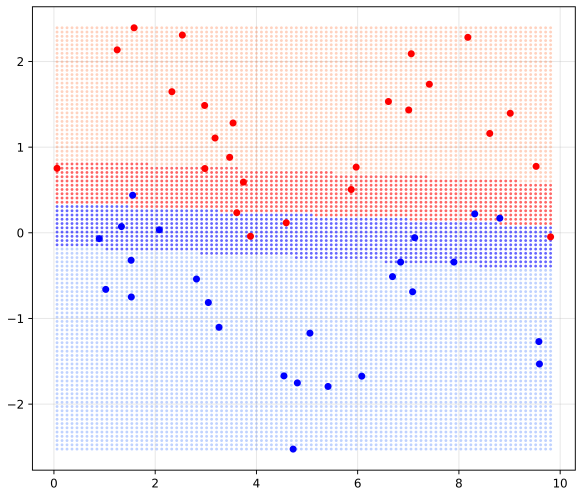
which is exactly equivalent to (2).

2 Linear SVM

After solving problem (2), we find the following decision boundaries:



(a) SVM decision boundary for p2data2.mat



(b) SVM decision boundary for p2data2.mat

Figure 1: SVM decision boundaries for question 2

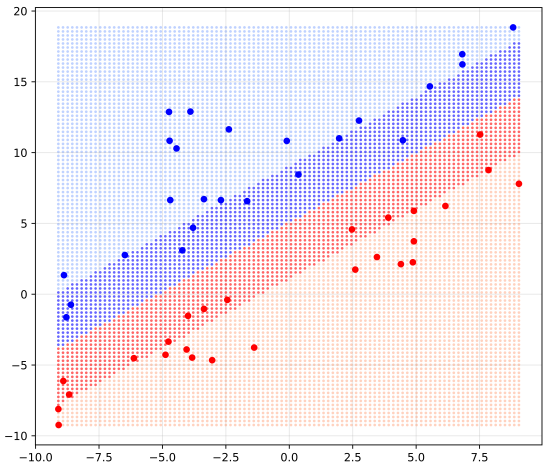


Figure 2: SVM decision boundary for p2data1.mat with $\varrho = 0.5$

The data from the first matrix is linearly separable, we are thus able to find a hyperplane that perfectly separates the data, as is showed in Figure 1a.

However, the data from the second matrix is not linearly separable, which is why no hyperplane can separate the data (Figure 1b.

One solution to overcome this problem could be to use the kernel trick to map the data to a higher dimension, where it might be linearly separable.

Note that for this exercise, we set $\varrho = 10^{-4}$. We saw in the lecture that the parameter ϱ determines how strict we are about allowing some misclassifications. For $\varrho \approx 0$, we are not allowing any misclassifications, which means that the maximum margin we will find is the one without any points inside of it, but some on its boundary.

Had we allowed a larger value of ϱ, the maximum margin might have allowed some points to be inside of it.

If we rerun the experiment with $\varrho = 0.5$, we obtain the results displayed in Figure 2.

3 Kernel Trick

To begin, we restate the lifted regression formulation as presented in the assignment in order to establish a clear analytical foundation for the subsequent discussion. The problem is expressed as the optimization program

$$\begin{aligned} \underset{w\in\mathbb{R}^D,\, b\in\mathbb{R},\, t,s\in\mathbb{R}^m}{\text{minimize}} \quad & \frac{1}{m}\sum_{i=1}^{m} s_i + \frac{\varrho}{2}\|w\|_2^2 \\ \text{subject to} \quad & \frac{1}{2}t_i^2 + 1 - y_i\big(w^T\phi(x_i) - b\big) + t_i \le s_i, \quad \forall i = 1,\dots,m, \\ & \frac{1}{2}t_i^2 \le s_i, \qquad\qquad\qquad\qquad\qquad\quad \forall i = 1,\dots,m, \end{aligned}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^D$ denotes a feature mapping that lifts each input vector $x_i \in \mathbb{R}^d$ into a higher-dimensional feature space $\mathbb{R}^D$ with $D \ge d$.

3.1 Lagrangian Derivation

Let $\lambda_i \ge 0$, $\gamma_i \ge 0$ denote the Lagrange multipliers corresponding to the inequality constraints $\frac{1}{2}t^2 + 1 - y_i(w^\top x_i - b) + t_i \le s_i$ and $\frac{1}{2}t^2 \le s_i$, respectively. Then we have, grouped by terms,

$$\begin{aligned} \mathcal{L}(w,b,t,s,\lambda,\gamma) = \frac{\varrho}{2}\|w\|_2^2 - w^T\sum_{i=1}^{m}\lambda_i y_i\phi(x_i) \\ + b\sum_{i=1}^{m}\lambda_i y_i + \sum_{i=1}^{m}\left(\frac{1}{m} - \lambda_i - \gamma_i\right)s_i \\ + \sum_{i=1}^{m}\frac{\lambda_i+\gamma_i}{2}t_i^2 + \sum_{i=1}^{m}\lambda_i t_i + \sum_{i=1}^{m}\lambda_i, \end{aligned}$$

completing the Lagrangian derivation.

3.2 Lagrangian Dual

Recall the dual objective, defined as

$$g(\lambda,\gamma) = \inf_{w\in\mathbb{R}^D,\, b\in\mathbb{R},\, t,s\in\mathbb{R}^m} \mathcal{L}(w,b,t,s,\lambda,\gamma),$$

and also notice the Lagrangian is separable in the decision variables. This fact allows us to go through the decision variables, finding the infimum of the Lagrangian with respect to that specific decision variable, and then combine all the infima together to form the dual objective $g(\lambda,\gamma)$.

3.2.1 Infimum over s

Starting with the infimum over s, notice

$$\sum_{i=1}^{m}\left(\frac{1}{m} - \lambda_i - \gamma_i\right)s_i$$

is affine in s, therefore

$$\begin{aligned} \inf_{s\in\mathbb{R}^m}\sum_{i=1}^{m}\left(\frac{1}{m} - \lambda_i - \gamma_i\right)s_i &= \sum_{i=1}^{m}\inf_{s_i\in\mathbb{R}}\left(\frac{1}{m} - \lambda_i - \gamma_i\right)s_i \\ &= \begin{cases} 0 & \text{if } m(\lambda_i+\gamma_i) = 1, \\ -\infty & \text{otherwise,} \end{cases} \end{aligned}$$

and since $\lambda_i \ge 0$, $\gamma_i \ge 0$, we have $0 \le \lambda_i \le 1/m$, $\forall i = 1,\dots,m$.

3.2.2 Infimum over t

Using the previous step, we have

$$\sum_{i=1}^{m} \frac{\lambda_i + \gamma_i}{2} t_i^2 + \sum_{i=1}^{m} \lambda_i t_i = \sum_{i=1}^{m} \left(\frac{\lambda_i + \gamma_i}{2} t_i^2 + \lambda_i t_i \right) = \sum_{i=1}^{m} \left(\frac{1}{2m} t_i^2 + \lambda_i t_i \right)$$

and again,

$$\inf_{t \in \mathbb{R}^m} \sum_{i=1}^{m} \left(\frac{1}{2m} t_i^2 + \lambda_i t_i \right) = \sum_{i=1}^{m} \inf_{t_i \in \mathbb{R}} \underbrace{\left(\frac{1}{2m} t_i^2 + \lambda_i t_i \right)}_{:= \psi_i(t_i)} = \sum_{i=1}^{m} \inf_{t_i \in \mathbb{R}} \psi_i(t_i),$$

where

$$\nabla_{t_i} \psi_i(t_i) = \frac{1}{m} t_i + \lambda_i = 0 \Leftrightarrow t_i^\star = -m\lambda_i,$$

and also

$$\nabla_{t_i}^2 \psi_i(t_i) = \frac{1}{m} > 0,$$

hence $\psi_i(t_i)$ is convex in t_i. All of the above yield the result

$$\begin{aligned} \inf_{t_i \in \mathbb{R}} \psi_i(t_i) = \psi_i(t_i^\star) &= -\frac{m}{2} \lambda_i^2 \\ \Rightarrow \inf_{t \in \mathbb{R}^m} \sum_{i=1}^{m} \left(\frac{1}{2m} t_i^2 + \lambda_i t_i \right) &= \sum_{i=1}^{m} \inf_{t_i \in \mathbb{R}} \left(\frac{1}{2m} t_i^2 + \lambda_i t_i \right) \\ = \sum_{i=1}^{m} \inf_{t_i \in \mathbb{R}} \psi_i(t_i) &= -\frac{m}{2} \sum_{i=1}^{m} \lambda_i^2, \end{aligned}$$

for all $i = 1, \ldots, m$.

3.2.3 Infimum over w

Define $v \in \mathbb{R}^d$ as

$$v := \sum_{i=1}^{m} \lambda_i y_i \phi(x_i)$$

so that the w-term of the Lagrangian reads

$$\frac{\varrho}{2} \|w\|_2^2 - w^T v$$

which is quadratic in w. Furthermore, we have

$$\nabla_w \Big(\frac{\varrho}{2} \|w\|_2^2 - w^T v \Big) = \varrho w - v = 0 \Leftrightarrow w^\star = \frac{1}{\varrho} v,$$

and additionally

$$\nabla_w^2 \Big(\frac{\varrho}{2} \|w\|_2^2 - w^T v \Big) = \varrho I \succ 0$$

since $\varrho > 0$, hence $\frac{\varrho}{2}\|w\|_2^2 - w^T v$ is convex. Substituting $w^\star$ back into the ℓ^2-norm and calculating its inner product with v yields the result

$$\begin{aligned}
\|w^\star\|_2^2 &= \left\| \frac{1}{\varrho} \sum_{i=1}^m \lambda_i y_i \phi(x_i) \right\|_2^2 \\
&= \frac{1}{\varrho^2} \sum_{i=1}^m \sum_{i'=1}^m \lambda_i \lambda_{i'} y_i y_{i'} \phi^T(x_i)\phi(x_{i'}) \\
(w^\star)^T v &= \left(\frac{1}{\varrho} \sum_{i=1}^m \lambda_i y_i \phi(x_i) \right)^T \left(\sum_{i'=1}^m \lambda_{i'} y_{i'} \phi(x_{i'}) \right) \\
&= \frac{1}{\varrho} \sum_{i=1}^m \sum_{i'=1}^m \lambda_i \lambda_{i'} y_i y_{i'} \phi^T(x_i)\phi(x_{i'}) \\
\Rightarrow \frac{\varrho}{2}\|w^\star\|_2^2 - (w^\star)^T v &= -\frac{1}{2\varrho} \sum_{i=1}^m \sum_{i'=1}^m \lambda_i \lambda_{i'} y_i y_{i'} \phi^T(x_i)\phi(x_{i'})
\end{aligned}$$

so therefore

$$\inf_{w \in \mathbb{R}^D} \left(\frac{\varrho}{2}\|w\|_2^2 - w^T v \right) = -\frac{1}{2\varrho} \sum_{i=1}^m \sum_{i'=1}^m \lambda_i \lambda_{i'} y_i y_{i'} \phi^T(x_i)\phi(x_{i'})$$

for all $i = 1, \ldots, m$.

3.2.4 Infimum over b

Similar to the infimum over s, notice that the term

$$b \sum_{i=1}^m \lambda_i y_i$$

is affine in b, therefore

$$\inf_{b \in \mathbb{R}} b \sum_{i=1}^m \lambda_i y_i = \begin{cases} 0 & \text{if } \sum_{i=1}^m \lambda_i y_i = 0, \\ -\infty & \text{otherwise,} \end{cases}$$

providing another constraint for the dual.

3.2.5 Dual Objective

Coming back to the dual objective, we now have

$$\begin{aligned}
g(\lambda) &= \sum_{i=1}^m \lambda_i - \frac{m}{2} \sum_{i=1}^m \lambda_i^2 - \frac{1}{2\varrho} \sum_{i=1}^m \sum_{i'=1}^m \lambda_i \lambda_{i'} y_i y_{i'} \phi^T(x_i)\phi(x_{i'}) \\
&= \sum_{i=1}^m \left(\lambda_i - \frac{m}{2} \lambda_i^2 \right) - \frac{1}{2\varrho} \sum_{i=1}^m \sum_{i'=1}^m \lambda_i \lambda_{i'} y_i y_{i'} \phi^T(x_i)\phi(x_{i'}),
\end{aligned}$$

which we are looking to maximize.

3.2.6 Dual Problem

All of the above results in the dual problem

$$\begin{aligned}
\underset{\lambda \in \mathbb{R}^m}{\text{maximize}} \quad & \sum_{i=1}^m \left(\lambda_i - \frac{m}{2} \lambda_i^2 \right) - \frac{1}{2\varrho} \sum_{i=1}^m \sum_{i'=1}^m \lambda_i \lambda_{i'} y_i y_{i'} \phi^T(x_i)\phi(x_{i'}) \\
\text{subject to} \quad & \sum_{i=1}^m \lambda_i y_i = 0, \quad 0 \le \lambda_i \le \frac{1}{m}, \quad \forall i = 1, \ldots, m,
\end{aligned}$$

which is exactly equivalent to (4) from the assignment.

3.3 KKT Conditions A

As for the first result, we have

$$\begin{aligned}\nabla_w \mathcal{L}(w,b,t,s,\lambda,\gamma) &= \varrho w - \sum_{i=1}^{m} \lambda_i y_i \phi(x_i) \\ &= 0 \Leftrightarrow w^\star = \frac{1}{\varrho}\sum_{i=1}^{m} \lambda_i^\star y_i \phi(x_i) \\ &\Rightarrow w_j^\star = \frac{1}{\varrho}\sum_{i=1}^{m} \lambda_i^\star y_i \phi_j(x_i), \quad \forall j = 1,\ldots,D,\end{aligned}$$

and as for the second result, it is

$$\begin{aligned}\nabla_{t_i} \mathcal{L}(w,b,t,s,\lambda,\gamma) &= (\lambda_i + \gamma_i)t_i + \lambda_i \\ &= 0 \Leftrightarrow t_i^\star = -\frac{\lambda_i}{\lambda_i^\star + \gamma_i^\star} \\ &\Rightarrow t_i^\star = -m\lambda_i^\star, \quad \forall i = 1,\ldots,m,\end{aligned}$$

since $\lambda_i^\star + \gamma_i^\star = 1/m$.

3.4 KKT Conditions B

To start, let us mention an elementary observation, which we show is true (see the following Corollary), as we derive other statements from it hereafter.

Corollary. *Previous results have shown that*

$$\gamma_k^\star = \frac{1}{m} - \lambda_k^\star \quad \text{for any } k = \{1,\ldots,m\},$$

which has to be strictly positive, since $0 < \lambda_k^\star < 1/m$. ◇

Using complementary slackness, we have

$$\begin{aligned}\lambda_i\left(\frac{1}{2}t_i^2 + 1 - y_i(w^T\phi(x_i) - b) + t_i - s_i\right) &= 0, \\ \gamma_i\left(\frac{1}{2}t_i^2 - s_i\right) &= 0,\end{aligned}$$

and because $\lambda_k^\star > 0$, $\gamma_k^\star > 0$,

$$\begin{aligned}\frac{1}{2}(t_k^\star)^2 + 1 - y_k((w^\star)^T\phi(x_k) - b^\star) + t_k^\star - s_k^\star &= 0, \\ \frac{1}{2}(t_k^\star)^2 - s_k^\star = 0 \Rightarrow s_k^\star &= \frac{1}{2}(t_k^\star)^2,\end{aligned}$$

therefore

$$\cancel{\frac{1}{2}(t_k^\star)^2} + 1 - y_k((w^\star)^T\phi(x_k) - b^\star) + t_k^\star - \cancel{\frac{1}{2}(t_k^\star)^2} = 0,$$

so we have

$$\begin{aligned}1 + t_k^\star &= y_k((w^\star)^T\phi(x_k) - b^\star) \\ \Rightarrow (w^\star)^T\phi(x_k) - b^\star &= y_k(1 + t_k^\star) \\ \Rightarrow b^\star &= (w^\star)^T\phi(x_k) - y_k(1 + t_k^\star),\end{aligned}$$

since $y_k^{-1} = y_k$ because $y_k \in \{+1, -1\}$ by definition. At this point, we need to make use of more previously discovered results. Specifically, we use

$$t_k^\star = -m\lambda_k^\star \quad \text{and} \quad w^\star = \frac{1}{\varrho}\sum_{i=1}^{m} \lambda_i y_i \phi(x_i)$$

to finally arrive at

$$\begin{aligned} b^\star &= \frac{1}{\varrho}\sum_{i=1}^{m} \lambda_i y_i \phi^T(x_i)\phi(x_k) - y_k(1 - m\lambda_k^\star) \\ &= \frac{1}{\varrho}\sum_{i=1}^{m} \lambda_i y_i \phi^T(x_i)\phi(x_k) + y_k(m\lambda_k^\star - 1) \end{aligned}$$

for any $k \in \{1, \ldots, m\}$, such that $\lambda_k^\star \in (0, 1/m)$.

3.5 Python solution

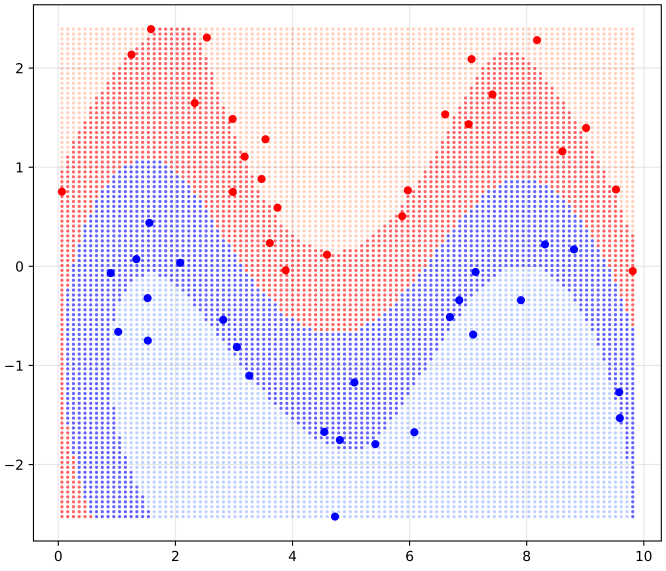


Figure 3: SVM decision boundary for p2data2.mat with Kernel trick.

As we can see on Figure 3, the data is now being properly separated compared to the case of Figure 1b. As expected, the data was mapped to a higher dimension where it became linearly separable.