Assignment 1

Machine Learning

USI 26-Oct-23

Fabian Flores Gobet

gobetf@usi.ch

Problem 1. Ridge Regression (15 points).

In a regression task, we have vectors $\boldsymbol{x} \in \mathbb{R}^D$, target values $y \in \mathbb{R}$ associated with them, and some model $f(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}$ to predict the target values for arbitrary vectors in $\mathbb{R}^D$.

Suppose we have a training dataset $\{\Phi, \boldsymbol{t}\}$, where $\Phi_{N\times D}$ is the design matrix in which each row is a feature vector $\phi(\boldsymbol{x})$ of a training point $\boldsymbol{x}$, $\boldsymbol{t}_{N\times 1}$ is the vector with target values for the training points. N is the number of points in the training dataset, D is the dimensionality of the feature space. Suppose that each entry in the last column of Φ is equal to 1.

Derive the closed form solution for the optimal parameters of a ridge regression model:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x})$$

The optimal parameters give the minimum to the following loss function:

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(f(\phi(\boldsymbol{x}_n)) - t_n)^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

Here $\|\cdot\|$ is the Euclidean norm of a vector; $\phi(\boldsymbol{x}_n)$ and t_n are n-th rows of Φ and $\boldsymbol{t}$ respectively.

$$\frac{\partial E}{\partial w_i} = \sum_{n=1}^{N}\left[(w^T\phi(x_n) - t_n)\,\phi(x_n)_i\right] + \lambda w_i$$

Generalizing for $\underline{w}$ we have:

$$\nabla_w E(w) = \Phi^T(\Phi w - t) + \lambda w$$

Setting $\nabla_w E(w) = \bar{0}$ $\Leftrightarrow$

$\Leftrightarrow$ $\Phi^T(\Phi w - t) + \lambda w = \bar{0}$

$\Leftrightarrow$ $\Phi^T\Phi w - \Phi^T t + \lambda w = \bar{0}$

$\Leftrightarrow$ $(\Phi^T\Phi + \lambda I_D)w = \Phi^T t$

$\Leftrightarrow$ $w = (\Phi^T\Phi + \lambda I_D)^{-1}\Phi^T t$

Convention

I_D is the $D \times D$ identity matrix

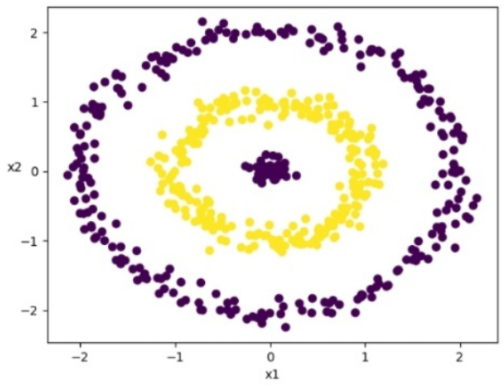
Note:

Assuming $\lambda > 0$ and an arbitrary D dimensional vector v

$v^T(\Phi^T\Phi + \lambda I_D)v = \underbrace{v^T\Phi^T\Phi v}_{\geq 0} + \underbrace{v^T\lambda I_D v}_{>0} > 0$, therefore $\Phi^T\Phi + \lambda I_D$ is invertible

Problem 2. Feature engineering (10 points).

Suppose you have the following set S of 2D points, $S_n = (x_n^{(1)}, x_n^{(2)})$



Color denotes the class attribution of a point: blue points belong to the class C_1, yellow points belong to the class C_2. **Propose the new features for points in S based on $x^{(1)}$ and $x^{(2)}$. In this new feature space, classes C_1 and C_2 should be linearly separable.**

We may project each point into a 3-dimensional space where their height is determined by their position in the x-y axis.

For all points $S_n^2 = (x_n, y_n) \in C^2$, we can determine $\underline{a}$ and $\underline{b}$ such that

- $b = \max(|x_1|, |y_1|, \ldots, |x_n|, |y_n|)$
- $a = \min(|x_1|, |y_1|, \ldots, |x_n|, |y_n|)$

where $(x_i, y_i) \in C^2$

Therefore, $\forall S_n^{(2)} \in C^2 : a^2 \leq x_n^2 + y_n^2 \leq b^2$.

This defines an annulus that contains all points of class C^2.

The mean of the interval $[a^2, b^2]$ is $\frac{a^2+b^2}{2}$.

Based on the RBF we can consider

$$z(x,y) = \exp\left(-\frac{(x^2+y^2-\frac{a^2+b^2}{2})^2}{2C^2}\right)$$

C is a constant that determines the steep descent of $z(x,y)$ when $b^2 < x^2+y^2 < a^2$.

We know that $z(x,y)$ never really reaches 0, but we want it to be proximate to 0 in the origin so linear separation is easier.

Let $\xi > 0$ be an arbitraty small distance. Then

$$z(0,0) = \exp\left(-\frac{1}{2C^2}\cdot\left(\frac{a^2+b^2}{2}\right)^2\right) = \xi \iff$$

$$\iff -\frac{1}{2C^2}\left(\frac{a^2+b^2}{2}\right)^2 = \ln(\xi) \iff$$

$$\iff C^2 = -\frac{1}{2\ln(\xi)}\left(\frac{a^2+b^2}{2}\right)^2$$

Let's consider that $\xi = 0.1$ and that

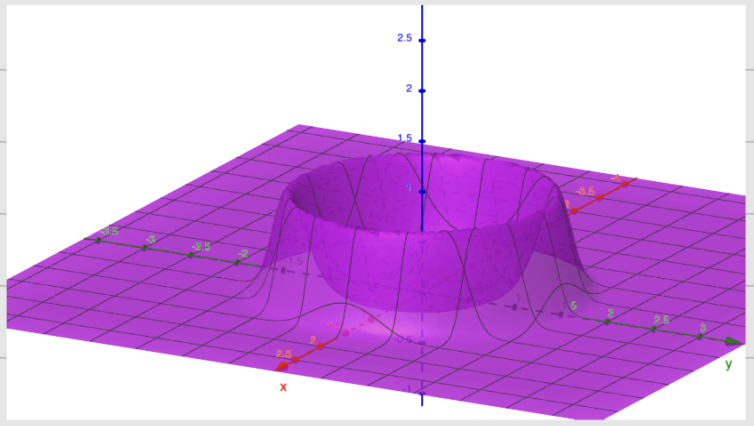
$$C^2 = -\frac{1}{2\ln(0.1)}\left(\frac{a^2+b^2}{2}\right)^2$$

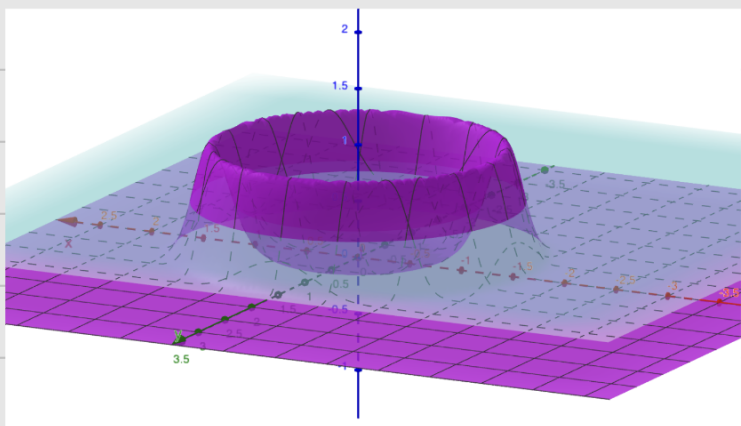
Given the above conditions, we can now design a feature space ϕ, s.t.

$$\phi(s_n) = \phi((x_n^{(1)}, x_n^{(2)})) = \left[x_n^{(1)}, x_n^{(2)}, z(x_n^{(1)}, x_n^{(2)}) \right]$$

This feature space effectively permits to find $w \in \mathbb{R}^3$ such that through a linear function $f(s_n) = w^T \phi(s_n) + b$ we can find an hyperplane with normal vector w that linearly separates C^1 and C^2 in the feature space.

The following images are mere ilustrative examples.





Problem 3. Kernel functions (12 points).

Consider the following function $f : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$

$$f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T \boldsymbol{x} \boldsymbol{x}^T \boldsymbol{y} \boldsymbol{y}^T \boldsymbol{y}$$

Prove that f is a valid kernel or prove the opposite.

The only rules allowed to use without a proof:
$k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is a valid kernel if

$$k(\boldsymbol{x}, \boldsymbol{y}) = c k_1(\boldsymbol{x}, \boldsymbol{y})$$
$$k(\boldsymbol{x}, \boldsymbol{y}) = k_1(\boldsymbol{x}, \boldsymbol{y}) + k_2(\boldsymbol{x}, \boldsymbol{y})$$
$$k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T A \boldsymbol{y}$$
$$k(\boldsymbol{x}, \boldsymbol{y}) = k_3(\phi(\boldsymbol{x}), \phi(\boldsymbol{y}))$$

where k_1 and k_2 are valid kernels in $\mathbb{R}^D$, $c > 0$ is a constant, ϕ is a function from $\mathbb{R}^D$ to $\mathbb{R}^M$, k_3 is a valid kernel in $\mathbb{R}^M$, A is a symmetric positive semidefinite matrix.

Since it would be helpful to have a property stating that the products of kernels is a kernel, let's prove it!

We know that a sum of kernels is a kernel. We also know that the identity matrix is symmetric posite definite, hence

$$K(x, y) = x^T I y = x^T y$$

Let ϕ_a and ϕ_b be a M and N dimensional feature space, respectively, such that

$$K_a(x, y) = \phi_a(x)^T \cdot \phi_a(y) \quad \text{and} \quad K_b(x, y) = \phi_2(x)^T \cdot \phi(y)$$

Then

$$K_a(x, y) \cdot K_b(x, y) = \left(\phi_a(x)^T \cdot \phi_a(y)\right)\left(\phi_2(x)^T \phi_2(y)\right) =$$

$$= \left(\sum_{m=1}^{M} \phi_a(x)_m \cdot \phi_a(y)_m \right) \left(\sum_{n=1}^{N} \phi_b(x)_n \cdot \phi_b(y)_m \right) =$$

$$= \sum_{m=1}^{M} \sum_{n=1}^{N} \phi_a(x)_m \phi_b(x)_n \cdot \phi_a(y)_m \cdot \phi_b(y)_n, \qquad (1)$$

let $C(x)_{mn} = \phi_a(x)_m \cdot \phi_b(x)_n$ and

$$C(x)_k = [C(x)_{k1}, C(x)_{k2}, \ldots, C(x)_{kN}] \quad N\text{-vector}$$

From (1) comes

$$\sum_{m=1}^{M} \sum_{n=1}^{N} C(x)_{mn} C(y)_{mn} = \underbrace{\sum_{m=1}^{M} C_m(x)^T \cdot C_m(y)}_{\text{sum of inner products}}$$

By the sum of kernels and inner product properties we conclude that

$$K_a(x,y) \cdot K_b(x,y) \quad \text{is also a kernel} \quad \square$$

Knowing this property we have that

$$f(x,y) = x^T x \, x^T y \, y^T y = (x^T x)(x^T y)(y^T y)$$

which is the product of kernels.

Therefore $f(x,y)$ is a kernel ■

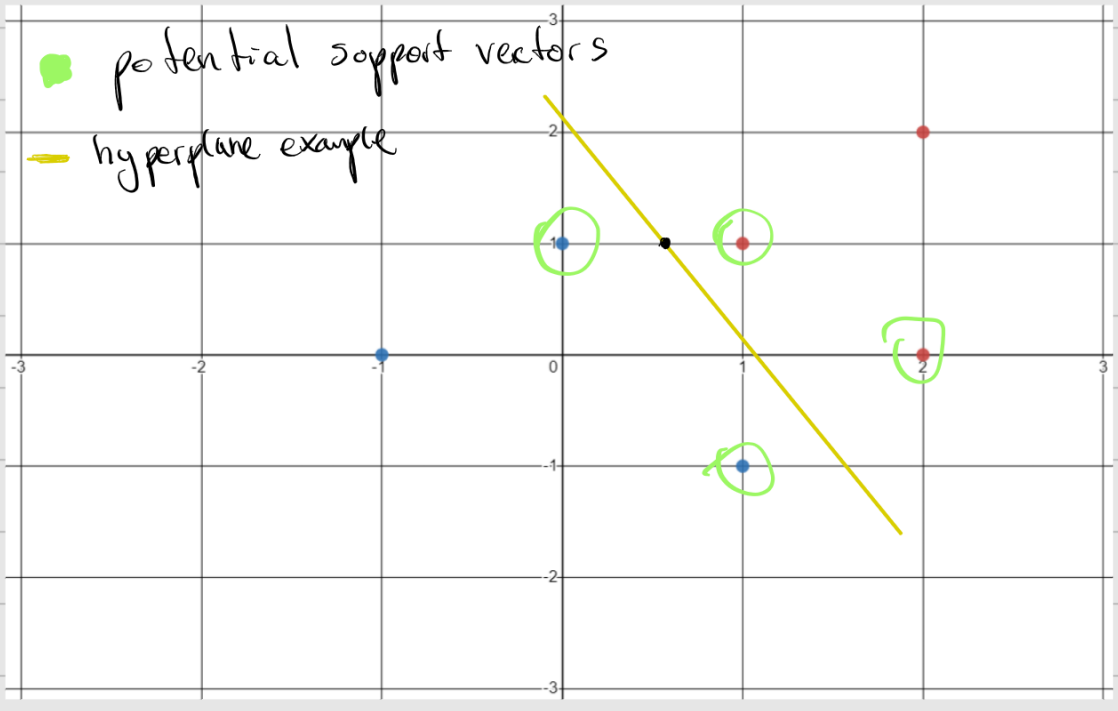
Problem 4. SVM (15 points).

Consider the following training data.

Class	x_1	x_2
+	1	1
+	2	2
+	2	0
−	1	-1
−	-1	0
−	0	1

1. Plot the six training points. Are the classes {+,−} linearly separable?
2. Construct the weight vector of the maximum margin hyperplane by inspection and identify the support vectors.
3. If you remove one of the support vectors does the size of the optimal margin decrease, stay the same, or increase?
4. Is your answer to (3) also true for any dataset? Provide a counterexample or give a short proof.

(1)



yes, they are linearly seperable

(2) the points with green circles are potential support vectors. By computing all distances between potential support vectors of different classes we know that points (0,1) and (1,1) have the minimum distance, so we'll define an hyperplane $r \equiv w_1 x + w_2 y + c$ that has equal distance to these points. that is

$$\frac{|w_1 \cdot 0 + w_2 \cdot 1 + c|}{\|w\|} = \frac{|1 \cdot w_1 + 1 \cdot w_c + c|}{\|w\|} =$$

$$\Leftrightarrow |w_2 + c| = |w_1 + (w_2 + c)| \Leftrightarrow$$

$$\Leftrightarrow (w_2 + c)^2 = w_1^2 + 2(w_2 + c) w_1 \qquad (w_2 + c)^2$$

$$\Leftrightarrow w_1 (w_1 + 2(w_2 + c)) = 0$$

$$\Leftrightarrow w_1 = 0 \ \vee \ w_1 = -2(w_2 + c)$$

if $w_1 = 0$ then y is constant $\Rightarrow$ horizontal line hyperplane

Thus $w_1 = -2(w_2 + c)$

$$r \equiv w_1 x + w_2 y + c = -2(w_2 + c) x + w_2 y + c$$

Now looking at $d(r, (0,1)) = \frac{|w_2 + c|}{\sqrt{4(w_2 + c)^2 + w_2^2}}$

we want to chose the third support vector that maximizes this distance

By observation of current r we know that the midle point between $(0,1)$ and $(1,1)$ is part of all possible hyperplanes.

$$M = \left(\frac{0+1}{2}, \frac{1+1}{2}\right) = \left(\frac{1}{2}, 1\right), \quad r\left(\frac{1}{2}, 1\right) = -2(w_2 + c)\frac{1}{2} + w_2 \cdot 1 + c = 0$$

Now, we know that $\underline{\underline{r}}$ passes between points $(1,-1)$ and $(2,0)$. We can also see that the closer $\underline{\underline{r}}$ is to $(1,-1)$, the greater the distance between $\underline{\underline{r}}$ and our 2 chosen support vectors. This suggests that $(1,-1)$ is the candidate for the third support vector. But $\underline{\underline{r}}$ can only be so close to $(1,-1)$ such that it doesn't violate the distance of $\underline{\underline{r}}$ to the former support vectors. In other words, the hyperplane has to be such that it's parallel to the line that passes through $(1,-1)$ and $(0,1)$.

Let $r' \equiv a'x + b'y + c' = 0$ such that

$(1,-1), (0,1) \in r'$. then

$$\begin{cases} b' + c' = 0 \\ a' - b' + c = 0 \end{cases} \quad \begin{cases} c' = -b \\ a' = 2b \end{cases}, \; b' \neq 0$$

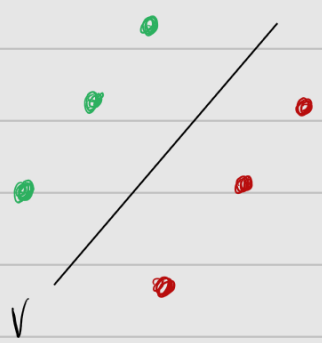
Let's choose $b' = 1 \Rightarrow r' \equiv 2x + y - 1 = 0$

Then $r' \parallel r \Rightarrow \begin{cases} -2(w_2 + c) = 2 \\ w_2 = 1 \end{cases} \quad \begin{cases} c = -2 \\ w_2 = 1 \end{cases}$

Thus $r \equiv 2x + y - 2 = 0$ and the margin width $M = 2d(r, (0,1)) = 2\dfrac{|1-2|}{\sqrt{4+1}} = \dfrac{2}{\sqrt{5}}$

(3) If we remove either (0,1) or (1,1) support vector, the hyperplane chages and the margin increases.
If we remove (1,-1) support vector the hyperplane changes and the also increases. In any case it increases

(4) The prior statement is not always true. Let's consider a set of data points linearly seperable, and let's consider that there are 3 points of class 1 as support vectors, and 3 points of class 2 as support vectors. Removing any one of these will not change the hyperplane and the Margin width.



In general removing points from a linearly separable set of points either increases the margin or it remains the same, but never decreses. This is because if the margin has been set, when we remove a support vector, depending on all other support vectors, the margin can only increase or remain the same because we are removing one of it's limiting constraints.