Machine Learning from Data –IDC

HW5 – Theory+ SVM

ID1: 203909320
ID2: 311132468

1. a.

> a. Let K, L be two kernels (operating on the same space) and let α, β be two positive scalars.
> Prove that $\alpha K + \beta L$ is a kernel.

Answer:

Given αK is a kernel hence there exists a mapper function ψ_1 such that:

$$\alpha K(x,y) = \langle \sqrt{\alpha}\, \psi_1(x), \sqrt{\alpha}\, \psi_1(y) \rangle$$

For the same way, for βL:

$$\beta L(x,y) = \langle \sqrt{\beta}\, \psi_2(x), \sqrt{\beta}\, \psi_2(y) \rangle$$

Now, we are requested to prove the following:

$$\widehat{K}(x,y) = \alpha K(x,y) + \beta L(x,y) \; \forall \; \alpha, \beta > 0$$

$$\widehat{K}(x,y) = \alpha K(x,y) + \beta L(x,y) = \langle \sqrt{\alpha}\, \psi_1(x), \sqrt{\alpha}\, \psi_1(y) \rangle + \langle \sqrt{\beta}\, \psi_2(x), \sqrt{\beta}\, \psi_2(y) \rangle =$$

Due to linearity of the inner product space (which is a vector space)

$$= \langle \sqrt{\boldsymbol{\alpha}}\, \boldsymbol{\psi_1}(\boldsymbol{x}) + \sqrt{\boldsymbol{\beta}}\, \boldsymbol{\psi_2}(\boldsymbol{x}), \sqrt{\boldsymbol{\alpha}}\, \boldsymbol{\psi_1}(\boldsymbol{y}) + \sqrt{\boldsymbol{\beta}}\, \boldsymbol{\psi_2}(\boldsymbol{y}) \rangle$$

We note that we expressed $\widehat{K}(x,y)$ as an inner product of the mappers to a given kernels, thus a kernel itself!

1. b.

> b. Provide (two different) examples of non-zero kernels K, L (operating on the same space), so that:
> i. $K - L$ is a kernel.
> ii. $K - L$ is not a kernel.

Answers:

1.b.i.

Assume K and L are both polynomial kernels with dimension of 1 as follows:

$K(x,y) = 2(x \cdot y)$, $L(x,y) = (x \cdot y)$

Applying $K - L \rightarrow 2(x \cdot y) - (x \cdot y) = (\boldsymbol{x} \cdot \boldsymbol{y})$ which is a polynomial kernel hence $\rightarrow$ **$\boldsymbol{K - L}$ is a kernel.**

1.b.ii.

Assume K and L are both polynomial kernels with dimension of 1 as follows:

$K(x, y) = (x \cdot y)\,, \quad L(x, y) = 2(x \cdot y)$

Applying $K - L \rightarrow \ (x \cdot y) - 2(x \cdot y) = \ -(\boldsymbol{x} \cdot \boldsymbol{y})$.
is this kernel a valid one? Let's check for positive semi- definite as it inherits the properties of the inner product vector space:

$K - L_{(x,x)} = \ -(\boldsymbol{x} \cdot \boldsymbol{x}) \overset{?}{\geq} \ \boldsymbol{0}$
we know that $(x \cdot x) = \|x\|^2 \geq 0$. So, for $-(\boldsymbol{x} \cdot \boldsymbol{x}) \geq \ \boldsymbol{0}$ to be true, the following has to exist→ $(x \cdot x) = 0$,**but!** $K, L \neq 0$ hence, $K - L_{(x,x)} < 0$ and it stands with contradiction to the positive semi-definite property of a kernel(and a inner product). Hence:

$$K - L_{(x,x)} < 0 \rightarrow \boldsymbol{NOT\ A\ KERNEL}$$

2.

2. Use Lagrange Multipliers to find the maximum and minimum values of the function subject to the given constraints:

Function: $f(x,y,z) = x^2 + y^2 + z^2$. Constraint: $g(x,y,z) = \frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} + \frac{z^2}{\beta^2} = 1$,

where $\alpha > \beta > 0$

Answer:

$$f(x,y,z) = x^2 + y^2 + z^2 \implies \nabla f = (2x, 2y, 2z)$$

$$s.t \quad g(x,y,z) = \frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} + \frac{z^2}{\beta^2} - 1 = 0 \implies \nabla g = \left(\frac{2x}{\alpha^2}, \frac{2y}{\beta^2}, \frac{2z}{\beta^2}\right)$$

f and g continuously differentiable real valued functions hence there exists a number λ which for him the following holds:

$$\vec{\nabla} f = -\lambda \vec{\nabla} g$$

$$\begin{cases} 1)2x = -\lambda \cdot \dfrac{2x}{\alpha^2} \\ 2)2y = -\lambda \cdot \dfrac{2x}{\beta^2} \\ 3)2z = -\lambda \cdot \dfrac{2x}{\beta^2} \\ 4)\dfrac{x^2}{\alpha^2} + \dfrac{y^2}{\beta^2} + \dfrac{z^2}{\beta^2} = 1 \end{cases} \implies \begin{matrix} 1)\ 2x\left(1 + \dfrac{\lambda}{\alpha^2}\right) = 0 \implies \boxed{x = 0 \ \boldsymbol{or}\ \lambda = -\alpha^2} \\ 2)\ 2y\left(1 + \dfrac{\lambda}{\beta^2}\right) = 0 \implies \boxed{y = 0 \ \boldsymbol{or}\ \lambda = -\beta^2} \\ 3)\ 2z\left(1 + \dfrac{\lambda}{\beta^2}\right) = 0 \implies \boxed{z = 0 \ \boldsymbol{or}\ \lambda = -\beta^2} \\ 4)\dfrac{x^2}{\alpha^2} + \dfrac{y^2}{\beta^2} + \dfrac{z^2}{\beta^2} - 1 = 0 \end{matrix}$$

We conclude that the multiplier λ yields us a degree of freedom in equations 2) & 3). If , $\lambda = -\beta^2$, then we can choose any y,z we want due to that DOF.
Let's look for some optional solutions:
we'll observe the constrain g and apply the trivial solutions $x = 0$ **or** $y = z = 0$:

1) $\boldsymbol{Case1}$: $x = 0 \implies \frac{y^2}{\beta^2} + \frac{z^2}{\beta^2} - 1 = 0 \overset{y=z}{\implies} 2 \cdot y^2 = \beta^2 \implies y = z = \pm\frac{\beta}{\sqrt{2}}$

$$\left(0, \pm\frac{\beta}{\sqrt{2}}, \pm\frac{\beta}{\sqrt{2}}\right)$$

2) $\boldsymbol{Case2}$: $y = z = 0 \implies \frac{x^2}{\alpha^2} - 1 = 0 \implies x^2 = \alpha^2 \implies x = \pm\alpha$

$$(\pm\alpha, 0, 0)$$

Using the given input inequality where $\alpha > \beta > 0$ we get that the points $(\pm\alpha, 0,0)$ yields maximal f whereas the points $\left(0, \pm\frac{\beta}{\sqrt{2}}, \pm\frac{\beta}{\sqrt{2}}\right)$ yields minimal f :

$$Max(f) = \alpha^2$$
$$Min(f) = \beta^2$$

3.

3. Let $X = \mathbb{R}^3$. Let

$C = H = \{h(a,b,c) = \{(x,y,z)\ s.t\ |x| \le a, |y| \le b, |z| \le c\}\ s.t.\ a,b,c \in \mathbb{R}_+\}$ the set of all origin centered boxes. Describe a polynomial sample complexity algorithm L that learns C using H. State the time complexity and the sample complexity of your suggested algorithm. Prove all your steps.

Answer:

We look to find the hypothesis which is a bounding box that separates the samples to binary groups: 1 or 0.

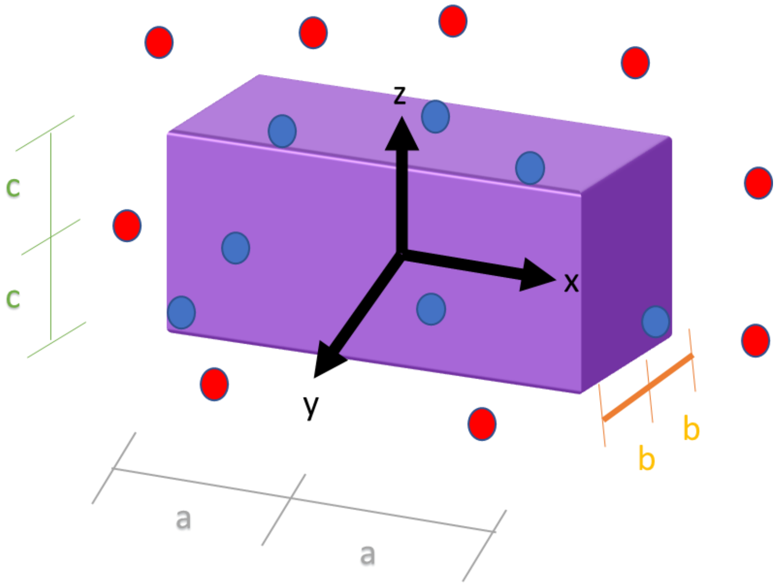
The learner will be defined as follows:

Find a,b,c :

- $a = \max(|x_i|)\ s.t\ C(x_i, y_j, z_j) = 1\ for\ each\ i \in [1,2,3, \dots, m]$
- $b = \max(|y_i|)\ s.t\ C(x_i, y_j, z_j) = 1\ for\ each\ i \in [1,2,3, \dots, m]$
- $c = \max(|z_i|)\ s.t\ C(x_i, y_j, z_j) = 1\ for\ each\ i \in [1,2,3, \dots, m]$

Now with learnt a,b,c the bounding box can be defined as the box originated in (0,0,0), and stretches accordingly to all axis:

- $x^+_{bound} = a;\ \ x^-_{bound} = -a$
- $y^+_{bound} = b;\ \ y^-_{bound} = -b$
- $z^+_{bound} = c;\ \ z^-_{bound} = -c$



Finally, Return $h(a,b,c) \in H$

Time complexity analysis:

We iterated over the samples 3 to find each time an optimal parameter (first a, then b, and last c) thus we bound our time complexity with big O notation s.t $O(3m) = O(m)$

Sample complexity analysis:

We'll divide the space between the concept and the hypothesis into 6 parts. Let there be a', b', c' that will represent the concept centered box s.t for each instance X the following holds:

$$\forall X \in R^3, concept(X) = 1 \; for \; (|X_x| \leq a' \wedge |X_y| \leq b' \wedge |X_z| \leq c')$$

Now it is possible to define the space of each of these bounding boxes:

$$B_1 = B_2 = (a' - a) \cdot b \cdot c$$
$$B_3 = B_4 = a \cdot (b' - b) \cdot c$$
$$B_5 = B_6 = a \cdot b \cdot (c' - c)$$

Such that the probability of the data D to be in either B_1 or B_2 or B_3 or B_4 or B_5 or B_6 overall the space X^m is $P_{B1} = P_{B2} = P_{B3} = P_{B4} = P_{B5} = P_{B6} = \frac{\varepsilon}{6}$

Now, assuming the data D visits each of the 6 spaces, the error between the hypothesis and the concept denoted:

$$Err(L(D), concept) = Err(h, concept) = \varepsilon$$

For a given ε and δ , the required number of samples will be yielded from the following:

$$P(\{D \in X^m: Err(h = L(D), concept) > \varepsilon\}) \leq \delta$$

$$P(\{D \in X^m: Err(h = L(D), concept) > \varepsilon\}) \leq \sum_{i=1}^{6} \left(P(X - B_i)\right)^m \leq 6\left(1 - \frac{\varepsilon}{6}\right)^m \leq 6e^{-\frac{m\varepsilon}{6}} \Rightarrow$$

$$6e^{-\frac{m\varepsilon}{6}} \leq \delta \Rightarrow \ln(6) - \frac{m\varepsilon}{6} \leq \ln(\delta) \Rightarrow \ln\left(\frac{6}{\delta}\right) \leq \frac{m\varepsilon}{6}$$

$$\boldsymbol{m} \geq \frac{6}{\varepsilon} \ln\left(\frac{6}{\delta}\right)$$

We note also that our sample complexity is polynomial on all the inspected parameters.

We conclude that if we want a confidence of $1 - \delta$ that our hypothesis will have the Err of ε, we require at least $\frac{6}{\varepsilon} \ln\left(\frac{6}{\delta}\right)$ instances.