Machine Learning from Data – IDC

HW5 – Theory + SVM: 304866551, 204118616

1. a. K, L are two kernels, therefore there are two mappings ϕ_K, ϕ_L accordingly such that:

$$\forall x, y, K(x,y) = \langle \phi_K(x), \phi_K(y) \rangle$$

$$\forall x, y, L(x,y) = \langle \phi_L(x), \phi_L(y) \rangle$$

Now we will show that exists an additional mapping ϕ, such that:

$$\forall x, y, (\alpha K + \beta L)(x,y) = \langle \phi(x), \phi(y) \rangle \quad \alpha > 0, \beta > 0$$

Considering the following mapping ϕ as:

$$\phi(x) = (\sqrt{\alpha} \cdot \phi_K(x), \sqrt{\beta} \cdot \phi_L(x))$$

And since ϕ_K, ϕ_L are mappings, thus ϕ is a mapping following:

$$\forall x, y, \langle \phi(x), \phi(y) \rangle = \langle \left(\sqrt{\alpha} \cdot \phi_K(x), \sqrt{\beta} \cdot \phi_L(x)\right), \left(\sqrt{\alpha} \cdot \phi_K(y), \sqrt{\beta} \cdot \phi_L(y)\right) \rangle =$$

$$= \langle \sqrt{\alpha} \cdot \phi_K(x), \sqrt{\alpha} \cdot \phi_K(y) \rangle + \langle \sqrt{\beta} \cdot \phi_L(x), \sqrt{\beta} \cdot \phi_L(y) \rangle =$$

$$= \alpha \cdot \langle \phi_K(x), \phi_K(y) \rangle + \beta \cdot \langle \phi_L(x), \phi_L(y) \rangle =$$

$$= \alpha \cdot K(x,y) + \beta \cdot L(x,y) = (\alpha K + \beta L)(x,y)$$

Therefore, $\boldsymbol{\alpha K + \beta L}$ is a kernel.

b. i. $\boldsymbol{K - L}$ **is a kernel:**

Let's define,

- $K = (x \cdot y + 1)^2$
- $L = (x \cdot y + 1)$

Then,

$$K - L = (x \cdot y + 1)^2 - (x \cdot y + 1) =$$

$$= (x \cdot y)^2 + 2(x \cdot y) + 1 - (x \cdot y + 1) =$$

$$= (x \cdot y)^2 + 2(x \cdot y) + 1 - (x \cdot y) - 1 =$$

$$= (x \cdot y)^2 + (x \cdot y)$$

Since $(x \cdot y)^2$ is a kernel, and $(x \cdot y)$ is a kernel, and sums of kernels are kernels as well, therefore $(x \cdot y)^2 + (x \cdot y)$ is also a kernel.

ii. $\boldsymbol{K - L}$ **is <u>not</u> a kernel:**

Let's define:

- $K = (x \cdot y + 1)$
- $L = (x \cdot y + 1)^2$

Then,

$$K - L = (x \cdot y + 1) - (x \cdot y + 1)^2 =$$

$$= (x \cdot y) + 1 - ((x \cdot y)^2 + 2(x \cdot y) + 1) =$$

$$= (x \cdot y) + 1 - (x \cdot y)^2 - 2(x \cdot y) - 1 =$$

$$= -(x \cdot y)^2 - (x \cdot y) =$$

$$= -((x \cdot y)^2 + (x \cdot y))$$

Let's define $W = (x \cdot y)^2 + (x \cdot y)$, and since $(x \cdot y)^2$ is a kernel and $(x \cdot y)$ is also a kernel, then $(x \cdot y)^2 + (x \cdot y)$ is also a kernel, according to the sum of kernels.

Therefore, $W = (x \cdot y)^2 + (x \cdot y)$ is a kernel, however $-W$ is <u>not</u> a kernel.

$$-W = -((x \cdot y)^2 + (x \cdot y))$$

Proof:

Using contradiction, let's say that $-W$ is a kernel, then there is a mapping such that:

$$\forall x, y, -W(x, y) = \langle \phi(x), \phi(y) \rangle$$

Since W is a non-zero kernel, then exists x such that $W(x, x) > 0$, hence:

$$-W(x, x) < 0$$

But,

$$-W(x, x) = \langle \phi(x), \phi(x) \rangle = \|\phi(x)\| \geq 0$$

Which causes to contradiction, hence $-W = K - L$ is not a kernel.

2. 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 $\alpha > \beta > 0$.

   Therefore,

$$\exists\lambda \ \ s.t \ \ \nabla f(\vec{v}) + \lambda\nabla g(\vec{v}) = 0$$

$$L(x, y, z) = x^2 + y^2 + z^2 + \lambda\left(\frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} + \frac{z^2}{\beta^2}\right)$$

$$\rightarrow \frac{\partial}{\partial x}L(x, y, z) = \frac{\partial}{\partial x}f(x, y, z) + \lambda\frac{\partial}{\partial x}g(x, y, z) = 0 \rightarrow$$
$$\rightarrow 2x + \lambda\frac{2x}{\alpha^2} = 0 \rightarrow$$
$$\rightarrow \mathbf{2\boldsymbol{x}\left(1 + \frac{\boldsymbol{\lambda}}{\boldsymbol{\alpha}^2}\right) = 0}$$

$$\rightarrow \frac{\partial}{\partial y}L(x, y, z) = \frac{\partial}{\partial y}f(x, y, z) + \lambda\frac{\partial}{\partial y}g(x, y, z) = 0 \rightarrow$$
$$\rightarrow 2y + \lambda\frac{2y}{\beta^2} = 0 \rightarrow$$
$$\rightarrow \mathbf{2\boldsymbol{y}\left(1 + \frac{\boldsymbol{\lambda}}{\boldsymbol{\beta}^2}\right) = 0}$$

$$\rightarrow \frac{\partial}{\partial z}L(x, y, z) = \frac{\partial}{\partial z}f(x, y, z) + \lambda\frac{\partial}{\partial z}g(x, y, z) = 0 \rightarrow$$
$$\rightarrow 2z + \lambda\frac{2z}{\beta^2} = 0 \rightarrow$$
$$\rightarrow \mathbf{2\boldsymbol{z}\left(1 + \frac{\boldsymbol{\lambda}}{\boldsymbol{\beta}^2}\right) = 0}$$

$$\rightarrow \frac{\partial}{\partial\lambda}L(x, y, z) = \boldsymbol{\frac{x^2}{\alpha^2} + \frac{y^2}{\beta^2} + \frac{z^2}{\beta^2} = 0}$$

According to the equations above, λ cannot revoke any of them, therefore the solutions are:

$$\frac{x^2}{\alpha^2}+\frac{y^2}{\beta^2}+\frac{z^2}{\beta^2}=1$$

$$\boldsymbol{x=0, y=0 \Rightarrow z= \pm\beta}$$

$$\boldsymbol{x=0, z=0 \Rightarrow y= \pm\beta}$$

$$\boldsymbol{y=0, z=0 \Rightarrow x= \pm\alpha}$$

Then the maximum or minimum values of the function are amongst the points below:

$$(x,y,z)=(\pm\alpha,0,0)$$

$$(x,y,z)=(0,\pm\beta,0)$$

$$(x,y,z)=(0,0,\pm\beta)$$

Since it is given that $\alpha > \beta > 0$,
The maximal points are: $(\pm\alpha,0,0)$, and the minimal points are: $=(0,0,\pm\beta)$.

Hence:

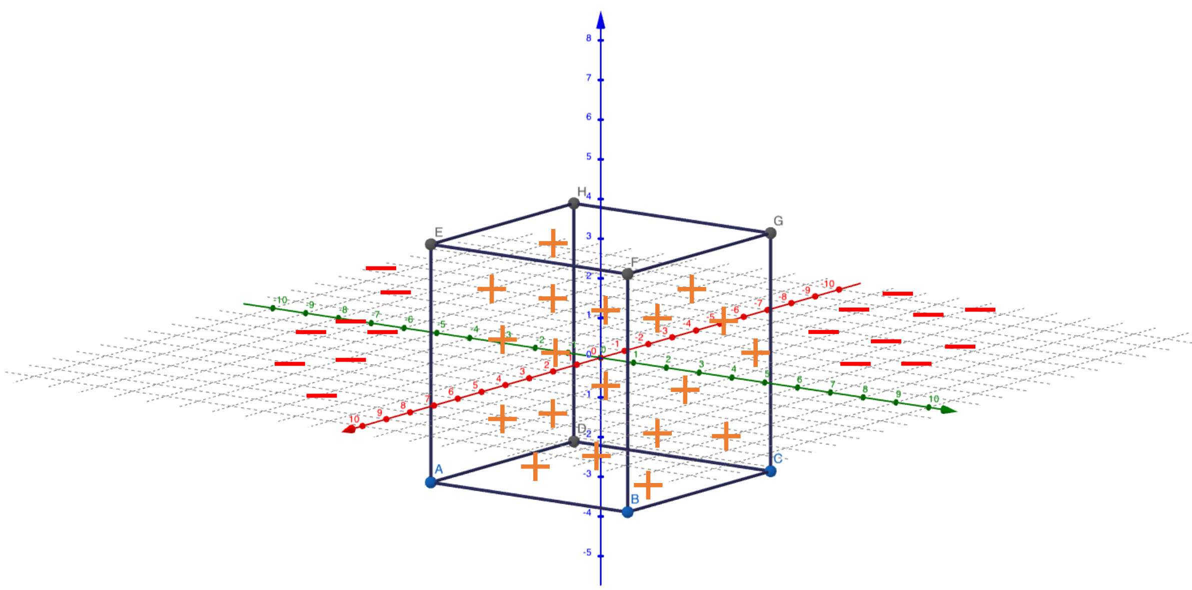
The <u>maximum</u> value of the function subject to the given constraints is: $\boldsymbol{\alpha^2}$**.**

The <u>minimum</u> value of the function subject to the given constraints is: $\boldsymbol{\beta^2}$.

3. $X = \mathbb{R}^3, C = H = \{h(a,b,c) = \{(x,y,z)\ s.t\ |x| \leq a, |y| \leq b, |z| \leq c\}\ s.t\ a,b,c \in \mathbb{R}_+$

<u>The Algorithm</u>
The algorithm will produce a hypothesis which is the smallest relevant area within the centered box which contains all the positive points (label: +). Our algorithm seeks to return a hypothesis $h \in H$.

- Let D be set of points (considering *m* sampled data points, $D \in \Omega^m$) in the plain labeled as positive (+) and negative (-) classes.
- **Time complexity**: the algorithm can be done in O(m).



<u>Consistent Learner</u>
Denotes as L, find points follows as below, and draws edges accordingly:

- L(D) = $h \in H$.
- Max and min x.
- Max and min y.
- Max and min z.

Returns h such that $h(x) = 1 => c(x) = 1$, but not necessarily the opposite. Both directions are true on the training data.

The Concept

For every $c \in C$, we will now bound all training datasets, D, that can lead to h = L(D) with $err(h, c) > \varepsilon$ into a union of sets (subsets of X^m, where $X = \mathbb{R}^3$) characterizable by regions that they do not visit.

We will estimate the probability of each such set of instances and finally their union. From that we will infer a bound-on sample complexity, **as a function of ε and δ** ($\varepsilon > 0$ and $\delta > 0$).

The Learning Axes Aligned Rectangles

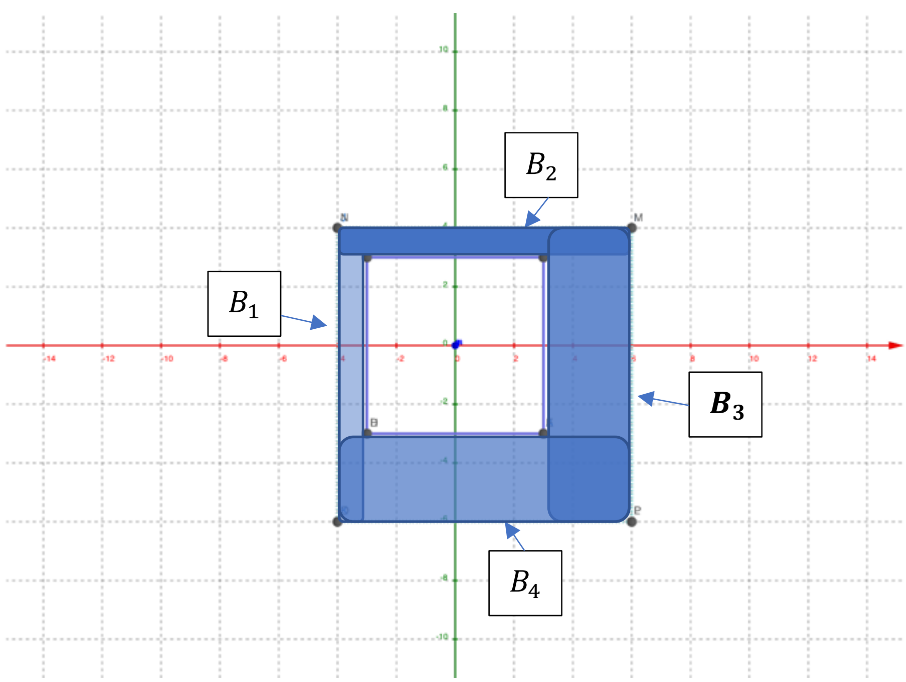
Considering each region B_i, generated by the concept c, **as a wall** that wraps the consistent hypothesis area.

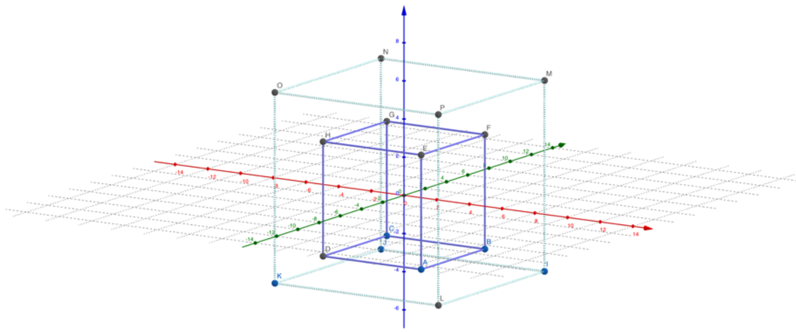
Each region could contain the errors, and is defined as:

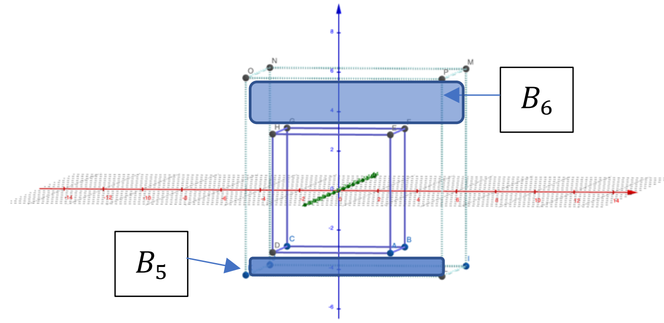
Let's define k, k', l, l', w, w' $\in \mathbb{R}_+$ as an arbitrary number.

- $B_1 = \{(x, y, z)\ s.t\ -a - k \leq x \leq a,\ |y| \leq b,\ |z| \leq c\}\ s.t\ a, c, b \in \mathbb{R}_+$.
- $B_2 = \{(x, y, z)\ s.t\ |x| \leq a,\ -b \leq y \leq b + l,\ |z| \leq c\}\ s.t\ a, c, b \in \mathbb{R}_+$.
- $B_3 = \{(x, y, z)\ s.t\ -a \leq x \leq a + k',\ |y| \leq b,\ |z| \leq c\}\ s.t\ a, c, b \in \mathbb{R}_+$.
- $B_4 = \{(x, y, z)\ s.t\ |x| \leq a,\ -b - l' \leq y \leq b,\ |z| \leq c\}\ s.t\ a, c, b \in \mathbb{R}_+$.
- $B_5 = \{(x, y, z)\ s.t\ |x| \leq a, |y| \leq b,\ -c - w \leq z \leq c\}\ s.t\ a, c, b \in \mathbb{R}_+$.
- $B_6 = \{(x, y, z)\ s.t\ |x| \leq a, |y| \leq b,\ -c \leq z \leq c + w'\}\ s.t\ a, c, b \in \mathbb{R}_+$.

$$P(B_i) = \frac{\varepsilon}{6}$$







<u>The Sample Complexity</u>

Consider training data, $D \in X^m$.
Assume that D visits each one of the 6 sets B_i (defined above), we can evaluate $err(h,c)$ as following:

$$P(B_1UB_2UB_3UB_4UB_5UB_6) \leq \sum_{i=1}^{6} B_i \leq \varepsilon \rightarrow Union\ of\ bound.$$

$$P(B_i) \geq \frac{\varepsilon}{6} \rightarrow\ Needs\ to\ have\ never\ visited\ at\ least\ one\ of\ them.$$

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

$$=> 6e^{-\frac{m\varepsilon}{6}} \leq\ \delta$$

$$=>\ m \geq \frac{6}{\varepsilon}(\ln(6) + \ln\left(\frac{1}{\delta}\right).$$

Q.E.D