Machine Learning from Data – IDC – 2023HW5 – Theory + SVM

1. Kernels and mapping functions (30 pts)
   a. (10 pts) Consider two kernels K_1 and K_2, with the mappings φ_1 and φ_2 respectively. Show that $K = 5K_1 + 4K_2$ is also a kernel and find its corresponding φ.

$K_1(x, y)$:

Defined: $\varphi_1(x_1, x_2 \ldots., x_n) = (x'_1, x'_2 \ldots., x'_k)$

$K_1(x, y) = \varphi_1(x)^T \varphi_1(y)$

$K_2(x, y)$:

Defined: $\varphi_2(x_1, x_2 \ldots., x_n) = (x''_1, x''_2 \ldots., x''_m)$

Let $(\varphi(x_1, x_2 \ldots., x_n)) = ((\sqrt{5}x'_1, (\sqrt{5}x'_2 \ldots (\sqrt{5}x'_k, 2x''_1 .., 2x''_m)$

Hence:

$$\varphi(x)^T \cdot (\varphi(y) = \left((\sqrt{5}x'_1, (\sqrt{5}x'_2 \ldots (\sqrt{5x}'_k, 2x''_1 .., 2x''_m\right)$$
$$\cdot (((\sqrt{5}y'_1, (\sqrt{5}y'_2 \ldots (\sqrt{5}y'_k, 2y''_1 .., 2y''_m)$$

$$= 5x'_1y'_1 + 5x'_2y'_2 + \cdots 5x'_ky'_k + 4x''_1y''_1 + \cdots + 4x''_my''_m$$

$$= 5(x'_1y'_1 + x'_2y'_2 + \cdots x'_ky'_k) + 4(x''_1y''_1 + \cdots + x''_my''_m)$$

$$= 5K_1 + 4K_2 = K(x, y)$$

Hence, φ is the mapping function of K, and he valid kernel.

b. (10 pts) Consider a kernel K_1 and its corresponding mapping φ_1 that maps from the lower space R^n to a higher space R^m (m > n). We know that the data in the higher space R^m, is separable by a linear classifier with the weights vector w.

Given a different kernel K_2 and its corresponding mapping φ_2, we create a kernel $K = 5K_1 + 4K_2$ as in section a above. Can you find a linear classifier in the higher space to which φ, the mapping corresponding to the kernel K, is mapping?

If YES, find the linear classifier weight vector.

If NO, prove why not.

<u>YES:</u>

We get the φ mapping function from the last question $5K_1 + 4K_2$. It given to use $sign(W \cdot \varphi_1(x))$ linear separates data in the space R^m.

We'll find a vector w_1 s.t for any x vector in R^n , $\left(w_1 \cdot \varphi_1(x)\right) = \left(w \cdot \varphi_1(x)\right)$, and so by extension $sign(w_1 \cdot \varphi_1(x)) = sign(w \cdot \varphi_1(x))$

proving that w_1 linearly separates data in the space R^{m+k} with the mapping φ.

Let $w = (w_1, \dots, w_m)$, we remind that $w_1 \in R^{m+k}$ s.t- $w_1 = (1/\sqrt{5})\ (w_1, w_2, \dots., w_m, 0, 0, 0, \dots, 0)$.

$$\left(w_1 \cdot \varphi_1(x)\right) = \left(\frac{1}{\sqrt{5}}\right)(w_1, w_2, \dots., w_m, 0, 0, 0, \dots, 0) \cdot$$
$$\cdot \left(\sqrt{5}x'1, \sqrt{5}x'2, \sqrt{5}x'3, \dots, \sqrt{5}x'm, 2x''1, 2x''2, 2x''3, \dots, 2x''k\right)$$
$$= \left(\frac{\sqrt{5}}{\sqrt{5x}}x'_1 w_1 + \frac{\sqrt{5}}{\sqrt{5x}}x'_2 w_2 + \dots + \frac{\sqrt{5}}{\sqrt{5x}}x'_m w_m\right) =$$
$$= (x'_1 w_1 + x'_2 w_2 + \dots + x'_m w_m) = \left(\boldsymbol{w} \cdot \boldsymbol{\varphi_1}(\boldsymbol{x})\right).$$

Hence: $\left(\boldsymbol{w} \cdot \boldsymbol{\varphi_1}(\boldsymbol{x})\right) = \left(\boldsymbol{w_1} \cdot \boldsymbol{\varphi_1}(\boldsymbol{x})\right)$

c. (10 pts) Consider the space $S = \{1,2, \dots N\}$ for some finite N (each instance in the space is a 1-dimension vector and the possible values are 1, 2, …, N) and the function $K(x, y) = 9 \cdot f(x, y)$ for every $x, y \in S$.

Prove that K is a valid kernel by finding a mapping φ such that:

$$\varphi(x) \cdot \varphi(y) = 9 \min(x, y) = K(x, y)$$

For example, if the instances are $x = 4, y = 8$, for some $N \geq 8$, then:

$$\varphi(x) \cdot \varphi(y) = \varphi(4) \cdot \varphi(8) = 9 \cdot \min(4,8) = 36$$

To prove that $K(x, y) = 9 \cdot f(x, y)$ is a valid kernel, we need to find a mapping $\varphi(x)$ such that $\varphi(x) \cdot \varphi(y) = K(x, y)$ for every x, y in the given space S.

Let's define the mapping $\varphi(x)$ as follows:

$\varphi(x) = \sqrt{}9x$

Now, let's calculate $\varphi(x) \cdot \varphi(y)$:

$\varphi(x) \cdot \varphi(y) = (\sqrt{}9x) \cdot (\sqrt{}9y)$
$= (\sqrt{}9x)(\sqrt{}9y)$
$= 9\ (xy)$

On the other hand, let's calculate $K(x, y) = 9 \cdot f(x, y)$:

$K(x, y) = 9 \cdot f(x, y)$
$= 9 \cdot \min(x, y)$

Now, we need to show that $\varphi(x) \cdot \varphi(y) = 9 \cdot \min(x, y)$:

$\varphi(x) \cdot \varphi(y) = 9\sqrt{}(xy) = 9 \cdot \min(x, y)$

Therefore, $\varphi(x) \cdot \varphi(y) = 9 \cdot \min(x, y)$ holds true for all x, y in the given space S.

Hence, we have shown that $K(x, y) = 9 \cdot f(x, y)$ is a valid kernel, and the corresponding mapping $\varphi(x) = \sqrt{}9x$ satisfies $\varphi(x) \cdot \varphi(y) = 9 \cdot \min(x, y) = K(x, y)$.

2. Lagrange multipliers (20 pts)

Suppose you are running a factory, producing some sort of widget that requires steel as a raw material. Your costs are predominantly human labor, which is \$20 per hour for your workers, and the steel itself, which runs for \$170 per ton.

Suppose your revenue R is modeled by the following equation:

$$R(h,s) = 200 \cdot h^{\frac{2}{3}} \cdot s^{\frac{1}{3}}$$

Where:

- h represents hours of labor
- s represents tons of steel

If your budget is \$20,000, what is the maximum possible revenue?

Let's define the objective function as the revenue function:

$$f(h,s) = 200 * \mathrm{h}^{\frac{2}{3}} * \mathrm{s}^{1/3}$$

Subject to the budget constraint:

$$g(h,s) = 20h + 170s - 20000 = 0$$

We introduce a Lagrange multiplier λ to incorporate the constraint into the objective function. The Lagrangian function is given by:

$$L(h,s,\lambda) = f(h,s) - \lambda * (g(h,s))$$

Now, we need to find the critical points of L(h, s, λ) by taking partial derivatives and setting them equal to zero:

$$\partial L/\partial h = (400/3) * h^{-\frac{1}{3}} * s^{\frac{1}{3}} - 20\lambda = 0\ (1)$$
$$\partial L/\partial s = (200/3) * h^{\frac{2}{3}} * s^{-\frac{2}{3}} - 170\lambda = 0\ (2)$$
$$\partial L/\partial \lambda = 20h + 170s - 20000 = 0\ (3)$$

$$(400/3) * h^{-\frac{1}{3}} * s^{\frac{1}{3}} = 20\lambda\ (1)$$

$$(200/3) * h^{\frac{2}{3}} * s^{-\frac{2}{3}} = 170\lambda\ (2)$$

/ (2)/(1)

$$\frac{\left(\frac{200}{3}\right) * h^{\frac{2}{3}} * s^{-\frac{2}{3}}}{\left(\frac{400}{3}\right) * h^{-\frac{1}{3}} * s^{\frac{1}{3}}} = \frac{1}{2} * \frac{h^{\frac{2}{3}} * s^{-\frac{2}{3}}}{h^{-\frac{1}{3}} * s^{\frac{1}{3}}} = \frac{1}{2} * \frac{h^{\frac{2}{3}} * h^{\frac{1}{3}}}{s^{\frac{2}{3}} * s^{\frac{1}{3}}} = \frac{1}{2} * \frac{h}{s} = \frac{170\lambda}{20\lambda}$$

$$\frac{h}{2s} = \frac{170}{20}$$

$$\frac{h}{s} = \frac{2 * 17}{2} = 17$$

$$h = 17s$$

$$(3) 20h + 170s - 20000 = 0$$

$$20 * (17s) + 170s - 20000 = 0$$

$$340s + 170s - 20000 = 0$$

$$510s = 20000$$

$$s \approx 39.22$$

⇨ h=17*39.22 = 666.66

Therefore, the optimal values for hours of labor (h) and tons of steel (s) that maximize the revenue within the budget constraint are approximately h ≈ 666.66 and s ≈ 39.22.

To find the maximum revenue, we substitute these values into the revenue function:

$$R(h, s) = 200 \cdot h^{\frac{2}{3}} \cdot s^{\frac{1}{3}}$$

$$R(666.66, 39.22) = 200 \cdot 666.66^{\frac{2}{3}} \cdot 39.22^{\frac{1}{3}} \approx \$\,309{,}458.62$$

Therefore, the maximum possible revenue within the given budget is approximately $ 309,458.62

3. PAC Learning and VC dimension (30 pts)

Let $X = \mathbb{R}^2$. Let

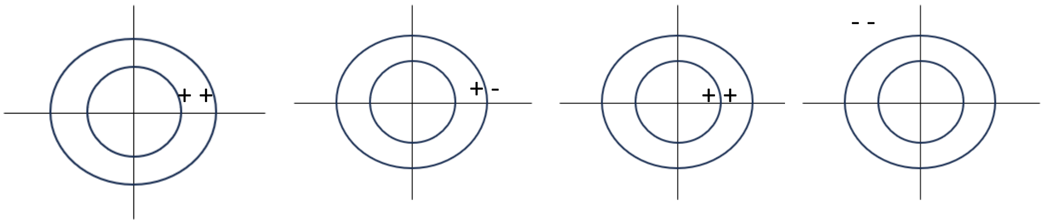
$$C = H = \left\{ h(r_1, r_2) = \left\{ (x_1, x_2) \middle| \begin{matrix} x_1^2 + x_2^2 \geq r_1 \\ x_1^2 + x_2^2 \leq r_2 \end{matrix} \right\} \right\}, \text{ for } 0 \leq r_1 \leq r_2,$$

the set of all origin-centered rings.

a. (8 pts) What is the $VC(H)$? Prove your answer.

H is the set of all origin centered rings, we shall prove that $VC(H) = 2$.

First $VC(H) \geq 2$:



We notice any dichotomy can be shattered for a set with size 2.

Now for $VC(H) < 3$:

Let x_1, x_2, x_3. If $distance(x_1) = distance(x_2) = distance(x_3)$ then there is no two points for which they can be classified heterogeneously (one of them +, other -).

Let $distance(x_1) < distance(x_2) < distance(x_3)$ wlog and $x_1 = +, x_2 = -, x_3 = +$. We note the inner, outer radius of the ring as r_1, r_2 respectively.

Then $r_1 \leq x_1 < x_2 < x_3 \leq r_2$. Then x_2 is within the hypothesis and needs to be categorized as +, contradiction.

Therefore, no hypothesis can shatter the latter points.

b. (14 pts) 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.

In class we saw a bound on the sample complexity when H is finite.

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

When $|H|$ is infinite, we have a different bound:

$$m \geq \frac{1}{\varepsilon}\left(4\log_2\frac{2}{\delta} + 8VC(H)\log_2\frac{13}{\varepsilon}\right)$$

The algorithm described should provide an output of hypothesis H, the ring (outer and inner radiuses) that encompasses all the positive points.

We define $D = \left(x_1^i, x_2^i\right)_{i=1}^{m}$.

We define the distance of each point from the center as $d = \sqrt{x_1^2 + x_2^2}$.

We define the positive points as $D^+ = \left(x_1^i, x_2^i\right)_{i=1}^{n}$.

We keep count of the running variables – r_{min}, r_{max}. Each time a new point is calculated, the variables consider the new point and update the two variables accordingly (two checks for each new data point), therefore it is $O(m)$ in complexity.

Let $\delta > 0, \epsilon > 0$.

We define:

- Inner concept radius - r_1^*
- Inner hypothesis radius - r_1
- Inner expanding radius - r_1^ϵ
- Outer concept radius - r_2^*
- Outer hypothesis radius - r_2
- Outer contracting radius - r_2^ϵ
- Circle with radius r_1^ϵ - c_1^ϵ
- Circle with radius r_2^ϵ - c_2^ϵ
- Area of the two rings made with the radius of hypothesis and concept - A^ϵ

We have two areas of mistake:

- The ring of all points for which $r_1^* \leq d \leq r_1$
- The ring of all points for which $r_1 \leq d \leq r_2^*$

Together these rings comprise A^ϵ.

The probability to be inside each of the rings is $\frac{\epsilon}{2}$, then we have the probability to be in A^ϵ is ϵ.

Two cases –

- First is if there are no points in A^ϵ:

Because the points are i.i.d then for each point x_i the probability to be outside of A^ϵ is $1 - \epsilon$.

Then by Taylor we have: $2 \cdot e^{\frac{-\epsilon \cdot m}{2}} \geq 2(1-\epsilon)^m$.

- Second case – some points are in A^ϵ:

The algorithm will classify points and provide $L(D) = h$ s.t. h is inside the concept - $r_1^\epsilon \leq r_1^* \leq r_1 \leq r_1^\epsilon \leq r_2^\epsilon \leq r_2 \leq r_2^*$ then the probability for a mistake is less than ϵ.

We choose m with respect to δ.

$$2e^{-\frac{\epsilon m}{2}} < \delta$$

$$-\frac{\epsilon m}{2} < ln\left(\frac{2}{\delta}\right)$$

$$\frac{\epsilon m}{2} > ln\left(\frac{2}{\delta}\right)$$

$$m > \frac{2ln\left(\frac{2}{\delta}\right)}{\epsilon}$$

c. (8 pts) You want to get with 95% confidence a hypothesis with at most 5% error. Calculate the sample complexity with the bound that you found in b and the above bound for infinite $|H|$. In which one did you get a smaller m? Explain.

We note that $\delta = \epsilon = 0.05$.

And with section b:

$$m > \frac{2ln\left(\frac{2}{\delta}\right)}{\epsilon} = \frac{2ln\left(\frac{2}{0.05}\right)}{0.05} = 147.6$$

$$m \geq 148$$

We can also use with the VC bound:

$$m \geq \frac{1}{\epsilon}\left(4\log_2\frac{2}{\delta} + 8VC(H)\log_2\frac{13}{\epsilon}\right) = \frac{1}{0.05}\left(4\log_2\frac{2}{0.05} + 8\cdot 2\cdot\log_2\frac{13}{0.05}\right) = 2992.9$$

$$m \geq 2993$$

We notice that we have a tighter bound by the results we got over the bound we have for infinite solutions.

4. VC dimension (20 pts)

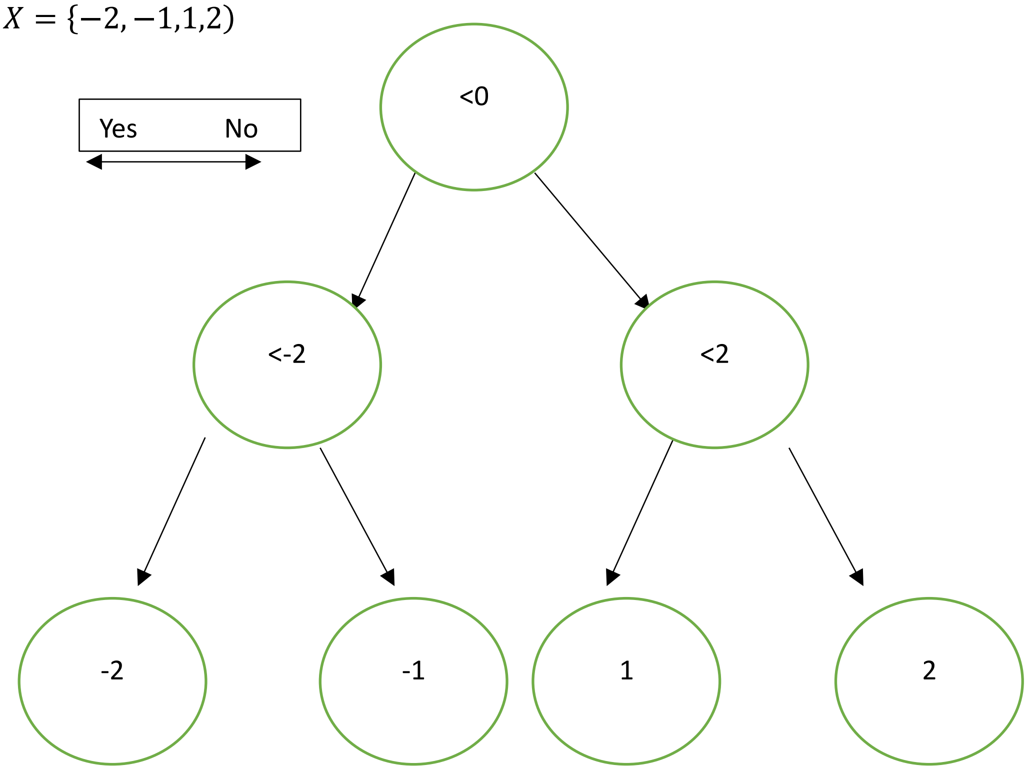
Let $X = \mathbb{R}$ and $n \in \mathbb{N}$.

Define "x-node decision tree" for any $x = 2^n - 1$ to be a full binary decision tree with x nodes (including the leaves).

Let H_m be the hypothesis space of all "x-node decision tree" with $n \leq m$.

a. (5 pts) What is the $VC(H_3)$? Prove your answer.

We will show that $VC(H_3) = 4$.
First $VC(H_3) \geq 4$:
We will find a set of 4 for which the hypothesis is shattered.
Let $X = \{-2, -1, 1, 2)$



We can see there is 16 ways to classify the points, therefore with the hypothesis we can shatter the set, then $VC(H_3) \geq 4$.

Next we will show $VC(H_3) < 5$:

Let $X = \{x_1, x_2, x_3, x_4, x_5\}$. With a decision tree algorithm with $m = 3$ then we have no more than 4 leaves in the tree. For any hypothesis h there is at least two points $x_i \in X$ s.t. both are in the same leaf by the pigeonhole principal. Therefore we cannot shatter the hypothesis. Therefore $VC(H_3) < 5$ and therefore $VC(H_3) = 4$.

b. (15 pts) What is the $VC(H_m)$? Prove your answer.

We will show that $VC(H_m) = 2^{m-1}$

First, we will show that $VC(H_m) \geq 2^{m-1}$:

A full binary tree with n nodes has $\frac{n+1}{2}$. If $n = 2^m - 1$ then $\frac{2^m-1+1}{2} = 2^{m-1}$ leaves.

Let $X = \{x_1, x_2, .., x_{2^{m-1}}\}$. Then we have $2^{2^{m-1}}$ optional classifications. We choose a hypothesis $h_1 \in H$ s.t. each leaf has exactly one point, then we cannot shatter X and $VC(H_m) \geq 2^{m-1}$

Second, we will show that $VC(H_m) \leq 2^{m-1}$:

Assume $|X| = 2^{m-1} + 1$. Assume there is some sorting for the points in X. A binary tree of 2^{m-1} leaves then by the pigeonhole principal we have at least one leaf with 2 points. We find a dichotomy that separates each two adjacent points (adjacent by sorting) into two different classes (…,+,-,+,-…) then we have the two points on the same leaf with different classification, and we have a contradiction. A binary tree with 2^{m-1} cannot shatter that group. Therefore $VC(H_m) \leq 2^{m-1}$.

All together we have $VC(H_m) = 2^{m-1}$.