HW5 – Theory + SVM

1. PAC Learning and VC dimension (30 pts)

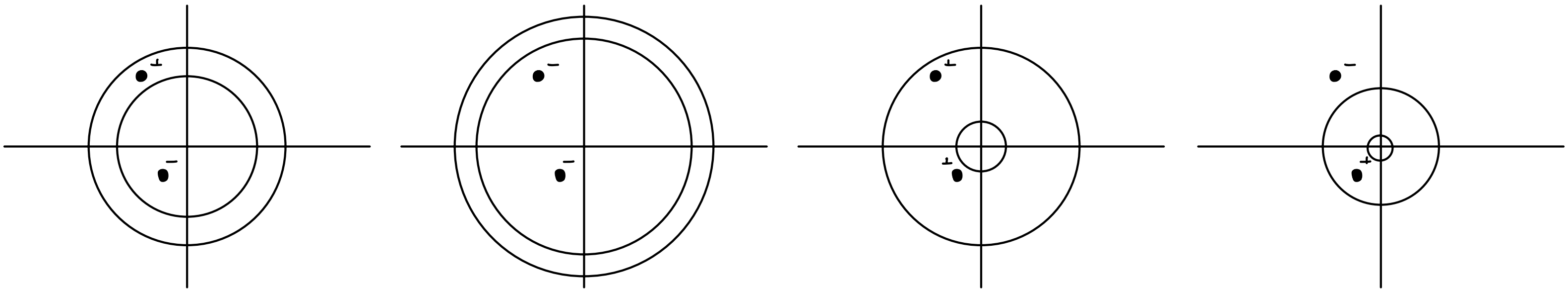
Let $X = \mathbb{R}^2$. Let

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

the set of all origin-centered rings.

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

(i) first, $VC(H) \geq 2$:



it suffisiant to find a set x of size 2 s.t $S(H,x) = \text{True}$

(ii) we prove that $vc(H) < 3$:

let X be a set of size 3 : $\{(a_1,a_2),(b_1,b_2),(c_1,c_2)\}$

- we assume WLOG that $d(a,0) \leq d(b,0) \leq d(c,0)$, where $a = (a_1,a_2)$, $b = (b_1,b_2)$, $c = (c_1,c_2)$ and $0 = (0,0)$. then the labeling $+ - +$ for a,b,c respectively is impossible.

a.t.c that this labeling is indeed possible. this means that $\exists\, h \in H$ s.t $h(a) = h(c) = +$ and $h(b) = -$. then, there would be some satisfying ring for the labeling, which is impossible for rings, since if a and c are both inside the ring, it has to be that b is inside the ring as well (because of our assumption)

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:

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

First, we want to propose a learning algorithm L and show that L is a consistent learner. L fits a hypothesis h to the training set by chosing r_1 to be the distance between the origin center and the closest data point in X with label +, and r_2 to be the distance between the origin center and the furthest data point in X with label +. in this case, all positive labled points in X are inside h and all negative labled points are outside h.

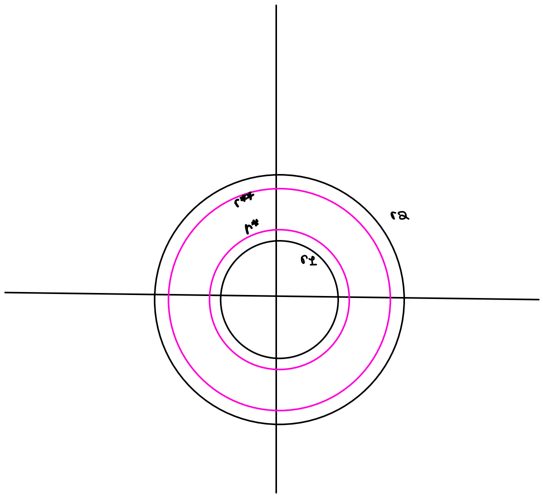
Now, we want to prove the consistency of L, that is- for all points in the training data $h(p) = c(p)$

consider a concept c in C. denote the radiuses defining the ring by r_1 and r_2 $(r_1 \leq r_2)$ by the definition of c, for all points P_in inside the ring, c(p_in) = + and for all points P_out outside the ring, c(P_out) = -

let P^* be the furthest point from the origin that is inside c and let P^{**} be the closest point to the origin that is inside c. that is, their lables are +. let r^* and r^{**} be the distances of P^* and P^{**} respectively from the origin.

Recall that $L(X) = h^*$ is the ring created by r^* and r^{**}

therefore, for all the interior points in c given in the training data P-in, h^*(P_in) = + and for all exterior points of c given in the training data P_out, h^*(P_out) = -. Thus, h^* is consistent with c.



now, we'll adress the Polynomiality sample complexity.

consider the two following sets:

$S\_eps_1 = \{x = (x_1, x_2) \mid r_1 \leq d(0,x) \leq r_1 + c_1\}$

$S\_eps_2 = \{x = (x_1, x_2) \mid r_2 - c_2 \leq d(0,x) \leq r_2\}$

we chose c_1, c_2 s.t. $\pi(S\_eps_1) = \pi(S\_eps_2) = \frac{\varepsilon}{2}$

now, consider the training data D^m where $|D^m| = m$

we have 2 cases:

let $h = L(D)$

(1) the training data visits both S-eps1 and S-eps2. due to L's consistency, $h \Delta c \subseteq$ S-eps1 $\cup$ S-eps2. thus, $\text{True\_Err}(h,c) =$

$$P(h \Delta c) \leq P(\text{Seps1} \cup \text{Seps2}) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

(2) the training doesn't visit at least one of S-eps1 and S-eps2. in this case

$$D^m \in \bigcup_{i=1}^{2} \{ \Delta \in X^m \mid \Delta \cap \text{Seps}_i = \emptyset \}$$

this is the case where $\text{True\_err}(h,c) > \varepsilon$, i.e. if $\text{True\_err}(h,c) > \varepsilon$ then,

$$D^m \in \bigcup_{i=1}^{2} \{ \Delta \in X^m \mid \Delta \cap \text{Seps}_i = \emptyset \}$$

$$P(D \in X^m \mid \text{True\_Err}(h,c) > \varepsilon) \leq \sum_{i=1}^{2} P(X - \text{Seps}_i)^m \leq 2\left(1 - \frac{\varepsilon}{2}\right)^m \leq 2 \cdot e^{\left(-\frac{m\varepsilon}{2}\right)}$$

so we can tune m in order to bound this Probability -

$$2 \cdot e^{\left(-\frac{m\varepsilon}{2}\right)} < \delta / 2$$

$$e^{\left(\frac{-m\varepsilon}{2}\right)} < \frac{\delta}{2}$$

$$\ln(e^{-\frac{m\varepsilon}{2}}) < \ln(\delta) - \ln(2)$$

$$\frac{-m\varepsilon}{2} < \ln(\delta) - \ln(2)$$

$$-m < \frac{2}{\varepsilon}(\ln(\delta) - \ln(2))$$

$$m > \frac{2}{\varepsilon}(\ln(2) - \ln(\delta))$$

that is $\boxed{m(\varepsilon, \delta) = \frac{2}{\varepsilon}(\ln(2) - \ln(\delta))}$

time complexity - in order to find r_1 and r_2, use euclidian distance for all positive labled points. distance calculation is in $O(1)$ for each distance, and thus we get an $O(|X|)$, i.e. a polinimial in m.

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.

(i) $m(0.05, 0.05) = \frac{2}{0.05}(\ln(2) - \ln(0.05)) = 147.555$

(ii) $m \geq \frac{1}{0.05}\left(4\log_2\left(\frac{2}{0.05}\right) + 8 \cdot 2 \cdot \log_2\left(\frac{13}{0.05}\right)\right) = 2992.911$

we got a smaller m in the first formula. this is because our formula takes advantage of geometric properties where the second formula doesn't make any further assumptions beside VC(H).

2. 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 $n \leq m$.

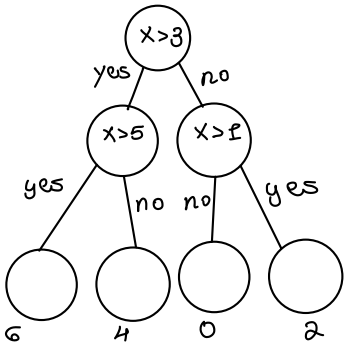
a. (5 pts) What is the $VC(H_3)$? Prove your answer.

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

a) $VC(H_3) = 4$.

→ for H_3, we have x node decision tree with 4 leaves, since for $n=3$ the number of leaves is according to the formula $\left\lceil \frac{2^3-1}{2} \right\rceil = 4$.

therefore, let $P_1 = 0$, $P_2 = 2$ $P_3 = 4$ $P_4 = 6$.



now, for every dichotomy of these 4 points, we can seperate the points.

← now, for any set of 5 points, there will always be at least 2 points that will end up in the same leave (according to the pigeon hole principle) and therefore there is always an assignment of labels that x node decision tree with $n=3$ cannot be consistent with.

b) → for H_m we have an x-node decision tree with 2^{m-1} leaves. we show that $vc(H_m) \geq 2^{m-1}$

let $P_1, \dots P_{2^{m-1}}$ points in $\mathbb{R}$. assume WLOG that $P_1 < P_2 < \dots < P_{2^{m-1}}$. we can construct an x-node decision tree as follows - the root of the tree will split the points equally. then, in each sub-tree we split again equally untill we'll end up with one point in each leave.

← again, as in the previous question, for every set of $2^{m-1}+1$ points, there are always at least 2 points that will end up in the same leave. therefore, there is an assignment in which the points that end up in the same leave don't have the same label, and therefore $VC(H_m) < 2^{m-1}+1$

3. Kernels and mapping functions (25 pts)

   a. (20 pts) Let $K(x, y) = (x \cdot y + 1)^3$ be a function over $\mathbb{R}^2 \times \mathbb{R}^2$ (i.e., $x, y \in \mathbb{R}^2$).

      Find ψ for which $K$is a kernel. (It may help to first expand the above term on the right-hand side).

   b. (2 pts) What did we call the function ψ in class if we remove all coefficients?

   c. (3 pts) How many multiplication operations do we save by using $K(x, y)$ versus $\psi(x) \cdot \psi(y)$?

(a

$$\left((xy)^2 + 2xy + 1\right)(xy + 1) = \left((x_1y_1 + x_2y_2)^2 + 2(x_1y_1 + x_2y_2) + 1\right)(x_1y_1 + x_2y_2 + 1)$$

$$\left(\overset{\checkmark}{x_1^2y_1^2} + \overset{\checkmark}{2x_1y_1x_2y_2} + \overset{\checkmark}{x_2^2y_2^2} + \overset{\checkmark}{2x_1y_1} + 2x_2y_2 + 1\right)(x_1y_1 + x_2y_2 + 1) =$$

$$\cancel{x_1^3y_1^3} + x_1^2y_1^2x_2y_2 + \cancel{x_1^2y_1^2} + 2x_1^2y_1^2x_2y_2 + 2x_1y_1x_2^2y_2^2 + 2x_1y_1x_2y_2 +$$

$$x_1y_1x_2^2y_2^2 + \cancel{x_2^3y_2^3} + \cancel{x_2^2y_2^2} + \cancel{2x_1^2y_1^2} + 2x_1y_1x_2y_2 + \cancel{2x_1y_1} + 2x_1y_1x_2y_2 + \cancel{2x_2^2y_2^2} + \cancel{2x_2y_2}$$

$$+ \cancel{x_1y_1} + \cancel{x_2y_2} + 1 \quad =$$

$$1 + \sum_{i=1}^{2} 3x_iy_i + \sum_{i=1}^{2} 3x_i^2y_i^2 + \sum_{i=1}^{2} x_i^3y_i^3 + 6x_1y_1x_2y_2 + 3x_1^2y_1^2x_2y_2 + 3x_1y_1x_2^2y_2^2 \quad =$$

$$1 + 3x_1y_1 + 3x_2y_2 + 3x_1^2y_1^2 + 3x_2^2y_2^2 + x_1^3y_1^3 + x_2^3y_2^3 + 6x_1y_1x_2y_2 + 3x_1^2y_1^2x_2y_2 + 3x_1y_1x_2^2y_2^2$$

$$\varphi(x) = \left(1, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, x_1^3, x_2^3, \sqrt{6}\, x_1x_2, \sqrt{3}\, x_1^2x_2, \sqrt{3}\, x_1x_2^2\right)$$

$$\varphi(y) \quad \left(1, \sqrt{3}y_1, \sqrt{3}y_2, \sqrt{3}y_1^2, \sqrt{3}y_2^2, y_1^3, y_2^3 \sqrt{6}y_1y_2, \sqrt{3}y_1^2y_2, \sqrt{3}y_1y_2^2\right)$$

$$\varphi(x) \cdot \varphi(y) = (x \cdot y + 1)^3$$

full rational variaty of degree 3 (b

. 10 Vs. 6 (c

we saved 4 multiplications operations

4. Lagrange multipliers (15 pts)

Let $f(x,y) = 2x - y$. Find the minimum and the maximum points for f under the constraint $g(x,y) = \frac{x^2}{4} + y^2 = 1$.

$$L(x,y,\lambda) = 2x - y + \lambda\left(\frac{x^2}{4} + y^2 - 1\right) = 2x - y + \frac{\lambda x^2}{4} + \lambda y^2 - \lambda$$

$$\frac{\partial}{\partial x} L(x,y,\lambda) = 2 + \frac{\lambda x}{2} = 0$$

$$\frac{\partial}{\partial y} L(x,y,\lambda) = -1 + 2\lambda y = 0$$

$$\frac{\partial}{\partial \lambda} L(x,y,\lambda) = \frac{x^2}{4} + y^2 - 1 = 0$$

(i) $4 + \lambda x = 0$

$\lambda x = -4$

$\boxed{x = \frac{-4}{\lambda}}$ $\quad \lambda \neq 0$

(ii) $2\lambda y = 1$

$\lambda y = \frac{1}{2}$

$\boxed{y = \frac{1}{2\lambda}}$ $\quad \lambda \neq 0$ (*)

$\longrightarrow$

$$\frac{4}{\lambda^2} + \frac{1}{4\lambda^2} - 1 = 0$$

$$\frac{16}{4\lambda^2} + \frac{1}{4\lambda^2} - 1 = 0$$

$$\frac{17}{4\lambda^2} - 1 = 0$$

$$\frac{17}{4\lambda^2} = 1$$

$$17 = 4\lambda^2$$

$$\lambda^2 = \frac{17}{4}$$

$$\boxed{\lambda = \pm \frac{\sqrt{17}}{2}}$$

$\longrightarrow$

$$\boxed{x = \pm \frac{8}{\sqrt{17}}}$$

$$y = \frac{1}{\cancel{2}\frac{\sqrt{17}}{\cancel{2}}} \qquad \boxed{y = \pm \frac{1}{\sqrt{17}}}$$

$$(\lambda_1, x_1, y_1) \longrightarrow \left(\frac{\sqrt{17}}{2}, \frac{-8}{\sqrt{17}}, \frac{1}{\sqrt{17}}\right) \longrightarrow f(x_1,y_1) = -\frac{16}{\sqrt{17}} - \frac{1}{\sqrt{17}} = \frac{-17}{\sqrt{17}} = -\sqrt{17} \longrightarrow \text{min Point}$$

$$(\lambda_2, x_2, y_2) \longrightarrow \left(-\frac{\sqrt{17}}{2}, \frac{8}{\sqrt{17}}, \frac{-1}{\sqrt{17}}\right) \longrightarrow f(x_2,y_2) = \frac{16}{\sqrt{17}} - \left(\frac{-1}{\sqrt{17}}\right) = \frac{17}{\sqrt{17}} = \sqrt{17} \longrightarrow \text{max Point}$$

(

*) λ cannot be 0, since in such case equation 1,2 are inconsistent

5. See notebook exercise (10 pts)