Adi Dereviani Prager, 305674731
Nitzan Gov, 203639646

HW5 - Theory + SVM - Solutions
Machine Learning from Data

Question 1

a. VC(H) is equal to 2.

Given any two points $P_1=(x_1, y_1)$ and $P_2=(x_2,y_2)$, and WLOG $x_1^2 + y_1^2 \leq x_2^2 + y_2^2$.
We have four possible dichotomies, and we will show that we can find r_1 and r_2 for each:

1. $P_1 \to -, P_2 \to -: r_1 \leq r_2 < x_1^2 + y_1^2 \; OR \; x_2^2 + y_2^2 < r_1 \leq r_2$
2. $P_1 \to +, P_2 \to -: r_1 \leq x_1^2 + y_1^2 \leq r_2 < x_2^2 + y_2^2$
3. $P_1 \to -, P_2 \to +: x_1^2 + y_1^2 < r_1 \leq x_2^2 + y_2^2 \leq r_2$
4. $P_1 \to +, P_2 \to +: r_1 \leq x_1^2 + y_1^2 \leq x_2^2 + y_2^2 \leq r_2$

Meaning we have proved that $VC(H) \geq 2$. Let's show that $VC(H) < 3$:

Let $P_1=(x_1, y_1)$, $P_2=(x_2,y_2)$ and $P_3=(x_3,y_3)$ and WLOG $x_1^2 + y_1^2 \leq x_2^2 + y_2^2 \leq x_3^2 + y_3^2$

Now, let $P_1 \to +, P_2 \to -, P_3 \to +$: Since P_1 is positive it enforces $r_1 \leq x_1^2 + y_1^2$. In addition, since P_3 is also positive we must conclude that $x_3^2 + y_3^2 \leq r_2$.

Since we have set WLOG $x_1^2 + y_1^2 \leq x_2^2 + y_2^2 \leq x_3^2 + y_3^2$, it means that $r_1 \leq x_2^2 + y_2^2 \leq r_2$ which would make P_2 positive in contradiction to the dichotomy. This proves $VC(H) < 3$ and determines that $VC(H) = 2$.

b. A possible polynomial sample complexity algorithm L that learns C using H is an algorithm that would find the smallest ring which contains all the positive points.

Let $\Delta = \Delta^m = (x_i, y_i)_{i=1}^m$ be a set of points, labeled positive and negative. Our algorithm seeks to return a hypothesis $h \in H$ and it would do it in the following steps:

I. Initialize $r_1' = r_2' = 0$
II. For every point in Δ^m:
    i. if the point is positive compute $x_2^2 + y_2^2$.
        1. If $x_i^2 + y_i^2 > r_2'$ update $r_2' = x_i^2 + y_i^2$
        2. If $r_1' < x_i^2 + y_i^2$ update $r_1' = x_i^2 + y_i^2$

Adi Dereviani Prager, 305674731
Nitzan Gov, 203639646

III. Return the ring with r_1' as the internal radius and r_2' as the external radius.

Hence the learning algorithm will produce the inner and outer radiuses: $L(\Delta) = h(r_1', r_2')$

Proof the algorithm is <u>consistent</u>:
Since we iterate over all the points in Δ^m and adjust the radiuses according to the positive points, we create two circles that are centered over the same point that would form a ring we would know contains all the positive points.

<u>Sample Complexity:</u>
Consider $c \in C$ and let $\Delta^m(c) = (x_i(c), y_i(c))_{i=1}^m$ training data generated from c without errors and by drawing m independent points according to a probability distribution π.

Given $\varepsilon > 0$ and $\delta > 0$ we wil now compute a number $m(\varepsilon, \delta)$ so that:

$$m \geq m(\varepsilon, \delta) \Rightarrow \pi^m(err_\pi(L(\Delta^m(c)), c) > \varepsilon) \leq \delta$$

Let $C = h(r_1, r_2)$ the concept we are trying to estimate. We will define two area so that the probability to be in those areas would be defined as $\frac{\varepsilon}{2}$:

$$S_1(\varepsilon) = \{(x, y): r_1 \leq x^2 + y^2 \leq r_1 + \alpha\}$$
$$S_2(\varepsilon) = \{(x, y): r_2 - \beta \leq x^2 + y^2 \leq r_2\}$$

And α, β are chosen so that $\pi(S_1(\varepsilon)) = \pi(S_2(\varepsilon)) = \frac{\varepsilon}{2}$.

There are two possible cases:

- The dataset visit $S_1(\varepsilon)$ and/or $S_2(\varepsilon)$. The probability in that case is at most ε from the way we have constructed the areas, so that means the error is bounded by ε.
- No samples are inside the $S_1(\varepsilon)$ or $S_2(\varepsilon)$. In that case the probability is:

$$\left(1 - \frac{\varepsilon}{2}\right)^m + \left(1 - \frac{\varepsilon}{2}\right)^m = 2\left(1 - \frac{\varepsilon}{2}\right)^m \leq 2e^{-\frac{\varepsilon m}{2}}$$

We wish for the probability for error to be bounded by δ:

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

Adi Dereviani Prager, 305674731
Nitzan Gov, 203639646

Time complexity:

Our algorithm initializes the radiuses ($O(1)$), goes over all the samples ($O(m)$) and computes the size for the positive points ($O(m^+)$). Overall, the time complexity is:

$$O(1) + O(m) + O(m^+) \approx O(m).$$

c. Using $\varepsilon = 0.05$ and $\delta = 1 - 0.95 = 0.05$ in both the expression found in b and in the general formula for infinitie |H|:

- $Formula\ found\ in\ b: m \geq \frac{2}{0.05}\ln\left(\frac{2}{0.05}\right) = 147.55 \rightarrow 148\ samples$
- $Formula\ for\ infinite\ |H|$:

$$m \geq \frac{1}{0.05}(4\log_2\left(\frac{2}{0.05}\right) + 8 * 2 * \log_2(\frac{13}{0.05})) = 2992.9 \approx 2993\ samples$$

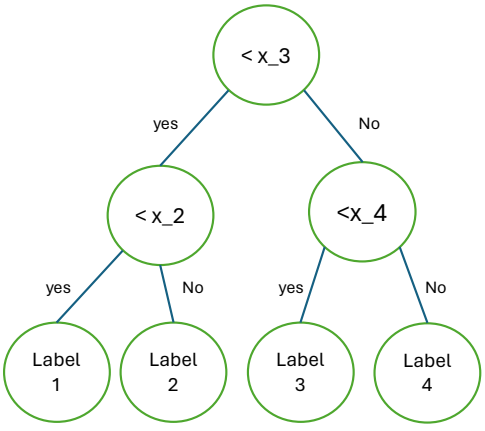
We have received a tighter bound using the formula from b since the equation for the general infinite |H| makes no assumptions on the data besides VC(H). On the contrary, in b we have utilized the known information to find a better bound.

Question 2

a. From the definition, the number of nodes in decision trees with m=3 can be:
- $n = 1: 2^1 - 1 = 1$
- $n = 2: 2^2 - 1 = 4 - 1 = 3$
- $n = 3: 2^3 - 1 = 8 - 1 = 7$

First, let's show $VC(H3) \geq 4$:

H_3 has at most $2^{n-1} = 2^2 = 4$ leaves. Let's define a dataset of points $\{x_1, x_2, x_3, x_4\}$. WLOG let's assume $x_1 < x_2 < x_3 < x_4$. A possible tree from H_3 would be:



Adi Dereviani Prager, 305674731
Nitzan Gov, 203639646

We can see that regardless of the dichotomy, we can separate any 4 points dataset to different leaves using $h \in H_3$ meaning $VC(H3) \geq 4$.

However, since we will have at most 4 leaves, we will now show that $VC(H_3) < 5$:

Let's define a dataset of points $\{x_1, x_2, x_3, x_4, x_5\}$. WLOG let's assume $x_1 < x_2 < x_3 < x_4 < x_5$ and we will choose the following dichotomy: $c(x_1) = c(x_3) = c(x_5) = 0$, $c(x_2) = c(x_4) = 1$.

Since this is a binary search tree, we know that two points must go to the same leaf. WLOG let's assume the points who share the leave are $x_i < x_j$:

- Option 1: the points are consecutive points, meaning $c(x_i) \neq c(x_j)$ that were labeled to the same leave.
- Option 2: the points are not consecutive, so they share the same label. However, this implies there is at least one point x_{i+1} between the two points with different label than them. From the behavior of decision trees, we know x_{i+1} must also be labeled to the same leaf – once again we have two labels in one leaf.

Therefore, we must conclude $VC(H_3) = 4$.

b. H_m has at most 2^{m-1} leaves. Therefore, like shown for H_4 we can conclude that a data set of $2^{m-1} + 1$ points would have for sure two points at the same leaf, and we would be able to find a dichotomy that cannot be separated using $h \in H_m$. Meaning $VC(H_m) < 2^{m-1} + 1$.

However, since H_m has at most 2^{m-1} leaves, for any dataset of 2^{m-1} points we can generalize the way we have built the tree in (a) to ensure each data point is classified to a different leaf and we would be able to choose the label of the leaf according to the point.

That way for each possible dichotomy we would be able to find a hypothesis $h \in H_m$ meaning $VC(H_m) \geq 2^{m-1}$.

In conclusion, we must conclude $VC(H_m) = 2^{m-1}$.

Adi Dereviani Prager, 305674731
Nitzan Gov, 203639646

Question 3

$$K(x,y) = (x \cdot y + 1)^3 = (x_1y_1 + x_2y_2 + 1)^3 =$$
$$\big((x_1y_1 + x_2y_2 + 1)(x_1y_1 + x_2y_2 + 1)\big)(x_1y_1 + x_2y_2 + 1) =$$
$$(x_1^2y_1^2 + x_1x_2y_1y_2 + x_1y_1 + x_1x_2y_1y_2 + x_2^2y_2^2 + x_2y_2 + x_1y_1 + x_2y_2 + 1)(x_1y_1 + x_2y_2 + 1) =$$
$$(x_1^2y_1^2 + 2x_1x_2y_1y_2 + x_2^2y_2^2 + 2x_1y_1 + 2x_2y_2 + 1)(x_1y_1 + x_2y_2 + 1) =$$
$$x_1^3y_1^3 + x_1^2x_2y_1^2y_2 + x_1^2y_1^2 + 2x_1^2x_2y_1^2y_2 + 2x_1x_2^2y_1y_2^2 + 2x_1x_2y_1y_2 + x_1x_2^2y_1y_2^2 + x_2^3y_2^3 + x_2^2y_2^2 +$$
$$2x_1^2y_1^2 + 2x_1x_2y_1y_2 + 2x_1y_1 + 2x_1x_2y_1y_2 + 2x_2^2y_2^2 + 2x_2y_2 + x_1y_1 + x_2y_2 + 1 =$$
$$x_1^3y_1^3 + 3x_1^2x_2y_1^2y_2 + 3x_1x_2^2y_1y_2^2 + 6x_1x_2y_1y_2 + 3x_2^2y_2^2 + 3x_1^2y_1^2 + 3x_1y_1 + 3x_2y_2 + x_2^3y_2^3 + 1 =$$
$$x_1^3y_1^3 + x_2^3y_2^3 + 3x_1^2x_2y_1^2y_2 + 3x_1x_2^2y_1y_2^2 + 6x_1x_2y_1y_2 + 3x_2^2y_2^2 + 3x_1^2y_1^2 + 3x_1y_1 + 3x_2y_2 + 1$$

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

b. Full rational variety.

c. We have 10 multiplication operations when using $\psi(x) \cdot \psi(y)$.
When using $K(x, y)$ we have only 4 multiplication operations, 2 from the inner product and another 2 from the power calculation.
Hence, by using $K(x, y)$ instead of $\psi(x) \cdot \psi(y)$ we are saving 6 multiplication operations.

Question 4

$$f(x,y) = 2x - y, g(x,y) = \frac{x^2}{4} + y^2 = 1 \Rightarrow g(x,y) = \frac{x^2}{4} + y^2 - 1 = 0$$
$$\mathcal{L}(x,y,\lambda) = 2x - y + \lambda\left(\frac{x^2}{4} + y^2 - 1\right)$$

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

From 1:
$$2 + \frac{2\lambda x}{4} = 0 \Rightarrow 2\lambda x = -8 \Rightarrow x = \frac{-4}{\lambda}$$

From 2:
$$-1 + 2\lambda y = 0 \Rightarrow 2\lambda y = 1 \Rightarrow y = \frac{1}{2\lambda}$$

From 1+2+3:
$$\frac{\left(\frac{-4}{\lambda}\right)^2}{4} + \left(\frac{1}{2\lambda}\right)^2 - 1 = 0 \Rightarrow \frac{16}{4\lambda^2} + \frac{1}{4\lambda^2} - 1 = \frac{17}{4\lambda^2} - 1 = 0 \Rightarrow \frac{17}{4\lambda^2} = 1 \Rightarrow \frac{17}{4} = \lambda^2 \Rightarrow \lambda_1, \lambda_2 = \pm\sqrt{\frac{17}{4}}$$

Adi Dereviani Prager, 305674731
Nitzan Gov, 203639646

$$x_1 = \frac{-4}{\sqrt{\frac{17}{4}}} = -\frac{8\sqrt{17}}{17} \quad y_1 = \frac{1}{\sqrt{\frac{17}{4}} \cdot 2} = \frac{\sqrt{17}}{17} \qquad x_2 = \frac{-4}{-\sqrt{\frac{17}{4}}} = \frac{8\sqrt{17}}{17}, \quad y_2 = \frac{1}{-\sqrt{\frac{17}{4}} \cdot 2} = -\frac{\sqrt{17}}{17}$$

$$f(x_1, y_1) = 2 \cdot -\frac{8\sqrt{17}}{17} - \frac{\sqrt{17}}{17} = -\sqrt{17}$$

$$f(x_2, y_2) = 2 \cdot \frac{8\sqrt{17}}{17} + \frac{\sqrt{17}}{17} = \sqrt{17}$$

(x1,y1) is the minimum, (x2,y2) is the maximum of the function with the given constraint.

Question 5: The answers for this question are in the Jupyter Notebook.