HW5:

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.

1a.

In terms of the mapping function φ, the relationship of a kernel to its mapping is given by $K(\mathrm{x}, \mathrm{y}) = \langle\varphi(\mathrm{x}), \varphi(\mathrm{y})\rangle$. For $K1$ and $K2$, we can write this as $K1(\mathrm{x}, \mathrm{y}) = \langle\varphi1(\mathrm{x}), \varphi1(\mathrm{y})\rangle$ and $K2(\mathrm{x}, \mathrm{y}) = \langle\varphi2(\mathrm{x}), \varphi2(\mathrm{y})\rangle$.

Define $\varphi(\mathrm{x}) = [\sqrt{5}\varphi1(\mathrm{x}), \sqrt{4}\varphi2(\mathrm{x})]$

Thus, we can express $K = 5K1 + 4K2$ as follows:

$K(\mathrm{x}, \mathrm{y}) = 5K1(\mathrm{x}, \mathrm{y}) + 4K2(\mathrm{x}, \mathrm{y})$

$= 5\langle\varphi1(\mathrm{x}), \varphi1(\mathrm{y})\rangle + 4\langle\varphi2(\mathrm{x}), \varphi2(\mathrm{y})\rangle = \begin{bmatrix}\sqrt{5\varphi1(\mathrm{x})} \\ \sqrt{4\varphi2(\mathrm{x})}\end{bmatrix} \cdot [\sqrt{5}\varphi1(\mathrm{y}), \sqrt{4}\varphi2(\mathrm{y})]$

Therefore, the resulting kernel K corresponds to the mapping $\varphi(\mathrm{x})$ as described above.

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 ($\mathrm{m} > \mathrm{n}$). 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.

1b. it is known that the data in the higher space R^m is separable by a linear classifier with weights vector w.

Hence we classify in the next form:

$$\begin{cases} 1 & w^T\varphi_1(x) > 0 \\ -1 & else \end{cases}$$

Now we'll find linear separator for $K = 5K_1 + 4K_2$.

φ_2 maps to dimension X.

We will choose $\varphi(x) = (\sqrt{5}\,\varphi_1(x), 2\varphi_2(x))$.
We will define vector W for our separator $W = (\frac{1}{\sqrt{5}} w(x), \underbrace{0{,}0{,}0{,}0}_{x})$

Hence, from the dot product $W^T \cdot \varphi(x) = w\varphi_1(x)$ and as seen previously if it's > 0 , we classify 1, else -1.

1c.

c. (10 pts) Consider the space $S = \{1,2, \ldots N\}$ 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:

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

Consider the following mapping function φ(x):

For a given x, φ(x) is a N-dimensional vector where the first min(x) entries are sqrt(9) and the remaining entries are zero.

Now, let's compute the dot product φ(x)·φ(y) and see if it equals K(x, y) for all x, y in S:

The dot product is computed by multiplying corresponding entries in each vector and then summing those products. Because the only non-zero entries in the vectors φ(x) and φ(y) are the first min(x) and min(y) entries respectively, the dot product essentially becomes the sum of the products of these non-zero entries.

If x <= y, then the non-zero entries in φ(x)·φ(y) are the first x entries. For those terms, φ(x)[i] = φ(y)[i] = sqrt(9). The dot product is then x * 9 = 9 * min(x, y).

If y < x, then the non-zero entries in φ(x)·φ(y) are the first y entries. For those terms, φ(x)[i] = φ(y)[i] = sqrt(9). The dot product is then y * 9 = 9 * min(x, y).

In both cases, φ(x)·φ(y) = 9 * min(x, y) = K(x, y), confirming that K is a valid kernel with the proposed feature mapping φ.

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?

2. The equation for revenue R is given by R(h, s) = 200 * h^(2/3) * s^(1/3), and we have a budget constraint of 20 * h + 170 * s = 20000, where h represents hours of labor and s represents tons of steel.

We can use the method of Lagrange multipliers to find the maximum revenue. In this case, the Lagrange function L(h, s, λ) is given by:

L(h, s, λ) = 200 * h^(2/3) * s^(1/3) + λ * (20h + 170s - 20000)

Taking the partial derivatives of L with respect to h, s, and λ and setting them equal to 0, we get the following equations:

∂L/∂h = (200 * 2/3 * h^(-1/3) * s^(1/3)) + 20λ = 0 (equation 1)

∂L/∂s = (200 * 1/3 * h^(2/3) * s^(-2/3)) + 170λ = 0 (equation 2)

∂L/∂λ = 20h + 170s - 20000 = 0 (equation 3)

We can solve this system of equations to find the optimal values of h and s.

S = 2000/51 ≈ 39.21 tons

h ≈ 2000/3 ≈ 666.66 hours

So, the maximum revenue occurs with about 666.66 hours of labour and 39.21 tons of steel.

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

R(666.66, 39.21) = 200 * (666.66)^(2/3) * (39.21)^(1/3) ≈ $51,852.0

So, the maximum possible revenue is about $51,852.0

3. <u>PAC Learning and VC dimension (30 pts)</u>

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

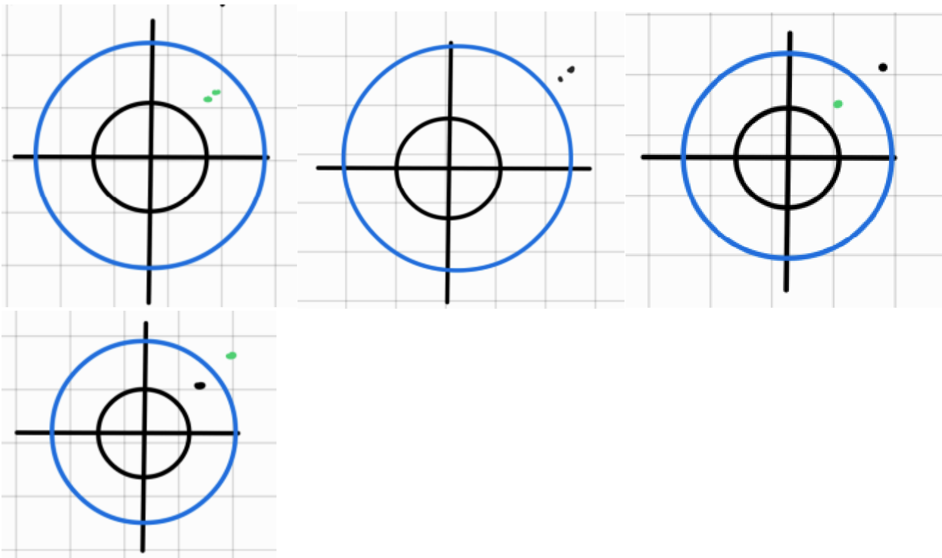
$$C = H = \left\{ h(r_1, r_2) = \left\{ (x_1, x_2) \middle| \begin{matrix} x_1^2 + x_2^2 \geq r_1 \\ x_1^2 + x_2^2 \leq r_2 \end{matrix} \right\} \right\}, \text{ for } 0 \leq r_1 \leq r_2,$$

the set of all origin-centered rings.

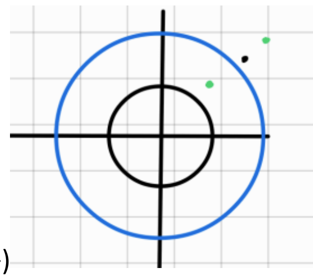
a. (8 pts) What is the $VC(H)$? Prove your answer.

3a. VC(H)=2. First we will see that $VC \geq 2$

There is a separation for every dichotomy:



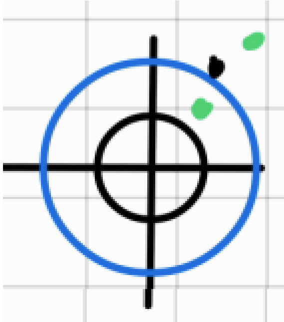
Now we will prove that $VC < 3$, let set of 3 different points $\{v_1, v_2, v_3\}$.



If the 3 points are colinear, we will choose labeling (+ - +)

And that is impossible.

If the three points are convex hull, we choose the labeling + - + again and that is impossible

to find a linear separator for them. 

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:

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

3b.

Algorithm:

We will go through training data D.

We will find the point that is classified as positive (1) and that its distance from the origin center is minimal, i.e., $r_1 = \min({x_1}^2 + {x_2}^2)$

and another point that is classified as positive (1) and that its distance from the origin center is maximal, i.e., $r_2 = \max\,(x_1^2 + {x_2}^2)$

We will draw the rings according to $r_1 and\ r_2$, hence return h=L(D) such that $\forall x \in X \Rightarrow 1 = c(x)$

- Note: Different training datasets will cause different results.

The algorithm is polynomial because it will cost O(m) to find the max and min out of m samples.

Correctness of the algorithm – if we block all positive points with the minimal length up to the maximal length, then the classification was correct. $\forall x\ h(x) = 1 \rightarrow c(x) = 1$

Now we will find the sample complexity:

Define $\varphi(b_1)$ – the probability that one point is located inside r_1 circle.

$\varphi(b_2)$ – the probability that one point is located outside r_2 circle.

$$\varphi(B_i) \geq \frac{\epsilon}{2}$$
$$\varphi(\{D \in X^m : Err(L(D), C) > \varepsilon\}) \leq \Sigma_{i=1}^{2}(\varphi(x - Bi)^m \leq 2\left(1 - \frac{\varepsilon}{2}\right)^m \leq 2\exp\left(\frac{-m\varepsilon}{2}\right) \Rightarrow$$

Hence the sample size $m(\varepsilon, \delta) = \frac{2}{\varepsilon} \cdot \overbrace{ln}^{*}\frac{2}{\delta}$

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.

3c.

Define $\varphi(b_1)$ – the probability that one point is located inside r_1 circle.

$\varphi(b_2)$ – the probability that one point is located outside r_2 circle.

$$\varphi(B_i) \geq \frac{\epsilon}{2}$$
$$\varphi(\{D \in X^m : Err(L(D), C) > \varepsilon\}) \leq \Sigma_{i=1}^{2}(\varphi(x - Bi)^m \leq 2\left(1 - \frac{\varepsilon}{2}\right)^m \leq 2\exp\left(\frac{-m\varepsilon}{2}\right) \Rightarrow$$

Hence the sample size $m(\varepsilon, \delta) = \frac{2}{\varepsilon} \cdot \overbrace{ln}^{*}\frac{2}{\delta}$

Substitute $\varepsilon = 0.05$, $\delta = 0.05$ in the two bound formulas

I) $\frac{1}{0.05} \cdot \left(4log\frac{2}{0.05} + 8 \cdot 2 \cdot \log_2 \frac{13}{0.05}\right) = 2992$

instance 2992

II) $* \Rightarrow$ m(0.05,0.05)=$\frac{2}{0.05} \cdot \ln\left(\frac{2}{0.05}\right) = 147$

We got two bounds for number of samples. Choose the bound 147 since it is tighter.

4. <u>VC dimension (20 pts)</u>

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

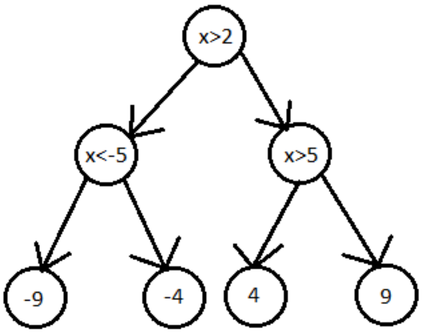
Define "x-node decision tree" for any $x = 2^n - 1$ to be a full binary decision tree with x nodes (including the leaves).

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.

4a. VC(H)=4

Firstly, for 4 points: -9, -4, 4, 9 we build the decision tree:



We can assign 2 options for every leaf which gives us 2^4different dichotomies.

We show that VC(H)<5.

Assume towards contradiction that we can shatter 5 points, denote by x_1, x_2, x_3, x_4, x_5.
$There\ exists\ a\ full\ binary\ desicion\ tree\ with\ 7\ nodes\ that\ shatters\ x_1, x_2, \ldots, x_5.$

But our tree has 4 leaves and from the pigeonhole principle there exists a leaf that classifies more than one point, and as we need to show a dichotomy for every label assigning, there exists a case where the classified points contradict.

Hence, $\exists x_i, x_j$ while $i \neq j$ that are classified by the same leaf and then we can classify + to x_i and – to x_j. ⇒We didn't reach a valid dichotomy, contradiction.

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

4b. we will show that VC (H_m)= 2^{m-1}

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

We look at points $x_1, x_2, .. x_{m-1}$. There exists a tree that classifies each point to a different leaf. And overall, there are 2^{m-1} leaves.

We can assign 2^m different dichotomies.

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

Assume toward contradiction that we can shatter the points $x_1, x_2, \ldots x_{2^{m-1}+1}$

We will use pigeonhole principle, hence there are at least two points which will be at the same leaf, and we could label them as different classes, and therefore we get a contradiction.