Ido Shenbach
Machine Learning Ex 05

Question 1

a) We have that $K_1 = \varphi_1(x)^T\varphi_1(y)$ and $K_2 = \varphi_2(x)^T\varphi_2(y)$ and we want to show that $K = 5K_1 + 4K_2$ is a Kernel.
Proof:

$$\begin{aligned} K = 5K_1 + 4K_2 &= 5\varphi_1(x)^T\varphi_1(y) + 4\varphi_2(x)^T\varphi_2(y) \\ &= \sqrt{5}\varphi_1(x)^T\sqrt{5}\varphi_1(y) + 2\varphi_2(x)^T 2\varphi_2(y) \\ &= (\sqrt{5}\varphi_1(x)^T 2\varphi_2(x)^T)\begin{pmatrix}\boldsymbol{\sqrt{5}\varphi_{1(}(y)} \\ \boldsymbol{2\varphi_2(y)}\end{pmatrix} \\ &= \varphi^T\varphi \end{aligned}$$

We can thus conclude that K is a Kernel, and $\varphi = \begin{pmatrix}\boldsymbol{\sqrt{5}\varphi_1} \\ \boldsymbol{2\varphi_2}\end{pmatrix}$

b) Yes, this is possible.
We are given that the data in the higher space R^m is separable by a linear classifier with weights vector w.
We'll define f(x): $R^m \rightarrow \{-1,1\}\ s.t\ \forall x \in R^m, f(x) = sgn(\langle w, \varphi_1(x)\rangle)$
Now, let $k \in N\ s.t\ \varphi_2 : R^n \rightarrow R^k$ and let w' = $\left\{\frac{1}{\sqrt{5}}w_1, \dots, \frac{1}{\sqrt{5}}w_m, 0,0,\dots,0\right\} \in R^{m+k}$ (where there are an k number of 0's)
We can now define g(x): $R^{m+k} \rightarrow \{-1,1\}$ to be the linear classifier s.t $\forall x \in R^n$ where $g(x) = sgn(\langle w', \varphi(x)\rangle)$ = $sgn(\frac{1}{\sqrt{5}}w_1\sqrt{5}\varphi_1(x_1), \dots, \frac{1}{\sqrt{5}}w_m\sqrt{5}\varphi_1(x_m), 0\cdot 2\cdot\varphi_2(x_1), \dots,\ 0\cdot 2\cdot\varphi_2(x_k),)$ = $sgn(\langle w, \varphi_1(x)\rangle)$ = f(x)
We have thus found a linear classifier weight vector.

c) Define $\varphi(x): S \rightarrow R^n$, where n is the maximal number in S.
$\forall x, \varphi(x) = \{3,3,\dots,3,0,0,\dots,0\}$ s.t there are an x number of 3's and a n-x number of 0's.
Then, $K(x,y) = \varphi(x)\ \varphi(y) = \sum_{i=1}^{\min(x,y)} 3\cdot 3 + \sum_{i=1+\min(x,y)}^{\max(x,y)} 3\cdot 0 + \sum_{i=1+\max(x,y)}^{\mathrm{n}} 0 = \sum_{i=1}^{\min(x,y)} 3\cdot 3 = 9\cdot\min(x,y)$
We have thus proved that K is a valid Kernel.

Question 2

Since the maximum budget is \$20000, we need to maximize L where $R(h,s) = 200h^{\frac{2}{3}}s^{\frac{1}{3}}$ and the constant g(h,s) = 20h + 170s – 20000 = 0

$$L(h,s) = 200h^{\frac{2}{3}}s^{\frac{1}{3}} + \lambda(20h + 170s - 20000)$$

$$\frac{\partial}{\partial_h}L(h,s) = 200s^{\frac{1}{3}}\frac{2}{3}h^{-\frac{1}{3}} + 20\lambda = 0\ldots(1)$$

$$\frac{\partial}{\partial_s}L(h,s) = 200h^{\frac{2}{3}}\frac{1}{3}s^{-\frac{2}{3}} + 170\lambda = 0\ldots(2)$$

$$\frac{\partial}{\partial_\lambda}L(h,s) = 20h + 170s - 20000 = 0\ldots(3)$$

$$(1)\times 8.5 : \frac{3400}{3}s^{\frac{1}{3}}\left(\frac{1}{h^{\frac{1}{3}}}\right) + 170\lambda = 0\ldots(4)$$

$$(4)-(2): \frac{3400}{3}s^{\frac{1}{3}}\left(\frac{1}{h^{\frac{1}{3}}}\right) - \frac{200}{3}h^{\frac{2}{3}}\left(\frac{1}{s^{\frac{2}{3}}}\right) = 0$$

$$17s^{\frac{1}{3}}\left(\frac{1}{h^{\frac{1}{3}}}\right) - h^{\frac{2}{3}}\left(\frac{1}{s^{\frac{2}{3}}}\right) = 0$$

$$17s - h = 0 \rightarrow 17s = h\ldots(5)$$

$$substitute\ (5)\ into\ (3): 20(17s) + 170s - 20000 = 0$$

$$510s = 20000 \rightarrow s = 39.216$$

$$h = 17(39.216) = 666.67$$

Thus, $R(h,s) = 200(666.67)^{\frac{2}{3}}(39.216)^{\frac{1}{3}} = \$51\,854.95$

Question 3

a) VC(H) = 2. Proof:

VC(H) ≥ 2: Show by example:

Choose two points, $X_1(1{,}0)\ and\ X_2(1{,}1)$. We will show that this can be divided into every possible set of labels.

(i) Assume $X_1 and\ X_2\ are\ positive.$

Consider $r_1 = 1, r_2 = \sqrt{2}$. Clearly both points satisfy the ring given by these r's.

(ii) Assume $X_1 and\ X_2\ are\ negative.$

Consider $r_1 = 2, r_2 = 3$. Clearly both points satisfy the ring given by these r's.

(iii) Assume $X_1\ is\ positive\ and\ X_2\ is\ negative.$

Consider $r_1 = 0.5, r_2 = 1$. Clearly both points satisfy the ring given by these r's.

(iv) Assume $X_1\ is\ negative\ and\ X_2\ is\ positive.$

Consider $r_1 = \sqrt{2}, r_2 = 2$. Clearly both points satisfy the ring given by these r's.

Since we have shown a set which shatters, we can say that VC(H) ≥ 2.

VC(H) < 3:

Consider any set of 3 distinct points $\{v_1, v_2, v_3\}$.

Consider an annulus where r_1 (the radius of the inner circle) is the distance of the point that is closest to the origin, and r_2 (the radius of the outer circle) is the distance of the point that is furthest from the origin.

However, there are 2 such points, call this set of points $S \subset \{v_1, v_2, v_3\}$.

Any origin centered ring/annulus that contains S must also contain all the points $\{v_1, v_2, v_3\}$, and there is atleast one v_i that was not used in S but must still be in the annulus.

Therefore, the label assignment that labels all points in S with positive and v_i with negative can't be consistent with any origin centered annulus.

This means that there is no set of size 3 that can be shattered by H, and therefore VC(H) < 3.

Thus, since $2 \leq VC(H) < 3 \rightarrow$ VC(H) = 2

b) **The Learning algorithm:**

L fits a hypothesis (from H) to the training set by choosing r1 to be the distance between the origin (the center of the circle), and the closest data point in D that is positive (i.e., that the point belongs to the concept c), and choosing r2 to be the distance between the origin, and the furthest data point in D that is positive (i.e., that the point belongs to the concept c).

It's also clear that the time complexity is polynomial in terms of the number of samples: finding a maximum distance and a minimum distance is linear in the number of points.

Correctness:

To show correctness of L, we must show that L is a consistent learner (consistent with the concept it's trying to learn, (c). We show that for all points in the training data, p, and for L(D) = h, h(p) = c(p).

Let ϵ>0,δ>0. Consider c∈C, and denote the inner radius of c by r_{c_1} and the outer radius by r_{c_2}. By the definition of c, all the interior points (inside the annulus) p_{in} are positive, and all the exterior points, p_{out} are negative. Let p_1^* be the closest point to the origin in c, and let p_2^* be the furthest point to the origin in c, i.e. p_1^* and p_2^* are positive, and denote r_1 as the distance of p_1^* from the origin, and r_2 as the distance of p_2^* from the origin.

Recall that then L(D) = h* is the annulus with inner radius r_1 and outer radius r_2. Therefore, for all interior points of c given in the training data, p_{in}, h*(p_{in}) is positive and for all exterior points of c given in the training data, p_{out}, h*(p_{out}) is negative.

⇨ *h* is consistent with c*

Sample Complexity:

Given the desired parameters ϵ and δ, the number of training samples m that is required to guarantee the desired error and confidence, is polynomial in $\frac{1}{\epsilon} > 0\,, \frac{1}{\delta} > 0.$

Let $\epsilon > 0, \delta > 0$. Consider $c \in C$, and denote the inner radius of c by r_{c_1} and the outer radius by r_{c_2}.

Now consider the annulus $c^{(\epsilon)}$ for which the inner radius of $c^{(\epsilon)}$, denoted by $r_{c_1^{(\epsilon)}}$ where $r_{c_1^{(\epsilon)}} > r_{c_1}$, and the outer radius of $c^{(\epsilon)}$, denoted by $r_{c_2^{(\epsilon)}}$, where $r_{c_2^{(\epsilon)}} < r_{c_2}$. This annulus $c^{(\epsilon)}$ creates the annulus A_ϵ, with outer radius of $r_{c_1^{(\epsilon)}}$ and inner radius of r_{c_1} that satisfies $\pi(A_\epsilon) = \frac{\epsilon}{2}$, and creates the annulus B_ϵ, with outer radius of r_{c_2} and inner radius of $r_{c_2^{(\epsilon)}}$ that satisfies $\pi(B_\epsilon) = \frac{\epsilon}{2}$.

More formally:

For $s_1 > r_{c_1}$ define the annulus A_s = $\{(x_1, x_2) |\, r_{c_1} \leq d\big((x_1, x_2), (0,0)\big) \leq s_1\}$, and for $s_2 < r_{c_2}$, define the annulus B_s = $\{(x_1, x_2) |\, s_2 \leq d\big((x_1, x_2), (0,0)\big) \leq r_{c_2}\}$.

Now,

$r_{c_1^{(\epsilon)}}$ = $\inf\{s_1 \mid \pi(A_s) \leq \frac{\epsilon}{2}\}$ and $r_{c_2^{(\epsilon)}}$ = $\inf\{s_2 \mid \pi(B_s) \leq \frac{\epsilon}{2}\}$

Now consider training data $D^{(m)}$, where $|D^m|$ = m.

We have 2 cases:

(1) The "bad" case: No points in the training data falls in the two annuluses that we created.

⇨ Happens with probability $2\left(1 - \frac{\epsilon}{2}\right)^m \leq 2e^{-\frac{m\epsilon}{2}}$, since for a single point $d \in D^m$, we have $\pi(d \notin A_\epsilon) = 1 - \frac{\epsilon}{2}$ and $\pi(d \notin B_\epsilon) = 1 - \frac{\epsilon}{2}$ and the points are independent.

(2) The "good" case: There exists a point in the training set that falls in the two annuluses that we created.

⇨ Happens with probability $1 - 2\left(1 - \frac{\epsilon}{2}\right)^m$.

Consider the good case, that is there exists a point in at least one of the annuluses. So the error is as follows:

$err(L(D^m), c) = \pi(h(D^m)\Delta c)$ by error definition from the general set up.

Note that, $\pi(h(D^m)\Delta c) \leq \pi(A_\epsilon \cup B_\epsilon) = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$ as A_ϵ and B_ϵ are disjoint. This implies that $err(L(D^m), c) \leq \epsilon$.

Finally, assume we can tune m to be so that in the bad case:

- ⇨ $\pi(D^m \; will \; yield \; the \; bad \; case) < \delta$

Therefore:
$\pi(err(L(D^m), c) \leq \epsilon) \geq \pi(D^m \; will \; yield \; the \; good \; case) = 1 - \pi(D^m \; will \; yield \; the \; bad \; case) > 1 - \delta$.

To see what m needs to be for $\pi(D^m \; will \; yield \; the \; bad \; case) < \delta$ to hold:

- ⇨ $\pi(D^m \; will \; yield \; the \; bad \; case) \leq 2e^{-\frac{m\epsilon}{2}} < \delta$
- ⇨ $ln2 - \frac{m\epsilon}{2} < ln\delta$
- ⇨ $2ln2 - m\epsilon < 2ln\delta$
- ⇨ $-m\epsilon < 2ln\delta - 2ln2$
- ⇨ $-m\epsilon < 2(ln\delta - ln2)$
- ⇨ $m\epsilon > -2\ln\left(\frac{\delta}{2}\right)$
- ⇨ $m > \frac{2}{\epsilon}\ln\left(\frac{2}{\delta}\right)$

QED

c) 95% confidence → $1 - \delta = 0.95$ → $\delta = 0.05$ and 5% error → $\epsilon = 0.05$.

- ⇨ $m > \frac{2}{0.05}\ln\left(\frac{2}{0.05}\right) \approx 147.56 = 148$ samples
- ⇨ $m \geq \frac{1}{0.05}(4\log_2\left(\frac{2}{0.05}\right) + 8 \; * 2 * \log_2\left(\frac{13}{0.05}\right) \approx 2992.91 = 2993$ samples

Even though the hypothesis space may be infinite, the fact that we know more details about how the hypothesis space looks, i.e. we know that we are looking at a ring, we can use this to make conclusions that allow us to be more precise with a tighter bound.

Question 4

a) VC(H_3) = 4. Proof:
VC(H_3) ≥ 4: Since we are looking at H_3, this is the hypothesis space of all "x-node decision trees" with $n \leq 3$, meaning that the maximal tree will have x = $2^3 - 1 = 7$ nodes, which means 4 leaves.
Since we are looking at 4 leaves, we have $2^4 = 16$ different options for the labels of the leaves, just as we have 16 different options for the labels of the points. We will always be capable of asking the correct questions in the internal nodes to ensure that every point corresponds to a leaf which will contain the corresponding label. Therefore, the hypothesis space H_3 shatters a set with 4 points and thus, VC (H_3) ≥ 4.
VC(H_3) < 5: As we have 5 points, but only 4 leaves, then by the pigeonhole principle, we can conclude that 2 points will both be labelled at 1 leaf. However since a leaf cannot hold more than one label, then we can say for sure that for any set of 5 points, H_3 doesn't shatter it. Concluding that VC(H_3) < 5.
Thus, since $4 \leq VC(H_3) < 5$ → VC(H_3) = 4

b) $VC(H_m) = 2^{m-1}$
VC(H_m) $\geq 2^{m-1}$: Since we are looking at H_m, this is the hypothesis space of all "x-node decision trees" with $n \leq m$, meaning that the maximal tree will have x = $2^m - 1$ nodes, which means 2^{m-1} leaves. Since we are looking at 2^{m-1}leaves, we have $2^{2^{m-1}}$ different options for the labels of the leaves, just as we have $2^{2^{m-1}}$ different options for the labels of the points. We will always be capable of asking the correct questions in the internal nodes to ensure that every point corresponds to a leaf which will contain the corresponding label. Therefore, the hypothesis space H_m shatters a set with 2^{m-1} points and thus, VC (H_m) $\geq 2^{m-1}$.
VC(H_m) $< 2^{m-1} + 1$: Generalizing our proof from above, as we have $2^{m-1} + 1$ points, but only 2^{m-1} leaves, then by the pigeonhole principle, we can conclude that 2 points will both be labelled at 1 leaf. However since a leaf cannot hold more than one label, then we can say for sure that for any set of $2^{m-1} + 1$ points, H_m doesn't shatter it. Concluding that VC(H_3) $< 2^{m-1} + 1$.
Thus, since $2^{m-1} \leq VC(H_3) < 2^{m-1} + 1$ → VC(H_3) = 2^{m-1}