HOMEWORK 2

YUNLIN ZHANG

A. Problem A

A.1. VC-dimension of convex combinations.

$$\mathcal{F} = \{\text{sgn}(\sum_{t=1}^{T} \alpha_t h_t) : h_t \in H, \alpha_t \geq 0, \sum_{i=1}^{T} \alpha_t \leq 1\}$$

Let VCdim(H) $= d$, and the VC-dimension of set of linear threshold functions in $\mathbb{R}^T$ is $T + 1$.

$\mathcal{F}$ is like a neural network with 1 hidden layer of concept class H. Therefore,
Let $d = \sum_{t=1}^{T} \text{VCdim}_t(m) + \text{VCdim}_{\text{linear}}(m) = dT + T + 1 < (T + 1)(d + 1)$ and there are T+1 nodes in this set up.

Using theorem 1 from Baum and Haussler[1], where $d = (d + 1)(T + 1)$, $N = T + 1$

$\Pi_{\mathcal{F}}(m) \leq [\frac{(T+1)em}{(d+1)(T+1)}]^{(d+1)(T+1)} < 2^m$, the last relationship is based on the assumption of finite VC-dimension for $\mathcal{F} \Rightarrow \text{VCdim}(\mathcal{F}) \leq \min\{m : \Pi_{\mathcal{F}}(m) < 2^m\}$

$[\frac{(T+1)em}{(d+1)(T+1)}]^{(d+1)(T+1)} < 2^m$

$\Leftrightarrow (d + 1)(T + 1) \log_2[\frac{(T+1)em}{(d+1)(T+1)}] < m$

Using the hint from 2014, setting $x = (d + 1)(T + 1)$ and $y = (T + 1)e/[(d + 1)(T + 1)]$

$\Leftrightarrow m = 2(d + 1)(T + 1) \log_2[(T + 1)e] \geq 1,\ xy = (T + 1)e > 4,\ x, y > 0,$ and $m \geq 1$ and the inequality is satisfied.

$\Rightarrow \text{VCdim}(\mathcal{F}) \leq m = 2(d + 1)(T + 1) \log_2[(T + 1)e] \quad \square$

Reference:

1. E. B. Baum and D. Haussler, What size net gives valid generalization?, Adv. Neural Inform. Process. Systems I, pp. 8190, Morgan Kaufmann,1989.
2. Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119-139, 1997.

B. Problem B

B.1. $\{X^+ \cup \{x_{m+1}\}, X^-\}$ **and** $\{X^+, X^- \cup \{x_{m+1}\}\}$ **are both linearly separable dichotomies by hyperplanes going through the origin (A) iff** $\{X^+, X^-\}$ **is linearly separable by a hyperplane going through the origin and** x_{m+1} **(B).**

First, to show (A)$\Rightarrow$(B)

(A)$\Rightarrow$ $\exists \mathbf{w}_1$ at origin $s.t.$ $\mathbf{w}_1 \cdot \mathbf{x}^+ > 0\ \forall \mathbf{x}^+ \in X^+ \cup \{x_{m+1}\},\ \mathbf{w}_1 \cdot \mathbf{x}^- < 0\ \forall \mathbf{x}^- \in X^-$

and

$\exists \mathbf{w}_2$ at origin $s.t.$ $\mathbf{w}_2 \cdot \mathbf{x}^+ > 0\ \forall \mathbf{x}^+ \in X^+,\ \mathbf{w}_2 \cdot \mathbf{x}^- < 0\ \forall \mathbf{x}^- \in X^- \cup \{x_{m+1}\}$

Define the hyperplane defined by $\mathbf{w}_1$ as U_1 and by $\mathbf{w}_2$ as U_2

Consider $f(\alpha) = [\alpha \mathbf{w}_1 + (1-\alpha)\mathbf{w}_2] \cdot \mathbf{x}_{m+1},\ \alpha \in [0,1]$

$f(\alpha)$ is a linear function of α therefore is continuous in $\alpha \in [0,1]$ and $f(0) < 0,\ f(1) > 0$, by intermediate value theorem $\exists \alpha' \in (0,1)$ $s.t.$ $f(\alpha') = 0$

$\Rightarrow \mathbf{y} = \alpha' \mathbf{w}_1 + (1-\alpha')\mathbf{w}_2$ $s.t.$ $\mathbf{y} \cdot \mathbf{x}_{m+1} = 0$

$\Rightarrow \mathbf{y}$ is a normal vector that defines a hyperplane U containing $\mathbf{x}_{m+1}$

$\because \alpha', (1-\alpha') > 0$

$\mathbf{y} \cdot \mathbf{x}^+ = \alpha' \mathbf{w}_1 \cdot \mathbf{x}^+ + (1-\alpha')\mathbf{w}_2 \cdot \mathbf{x}^+ > 0, \forall \mathbf{x}^+ \in X^+$

$\mathbf{y} \cdot \mathbf{x}^- = \alpha' \mathbf{w}_1 \cdot \mathbf{x}^- + (1-\alpha')\mathbf{w}_2 \cdot \mathbf{x}^- < 0, \forall \mathbf{x}^- \in X^-$

$\Rightarrow \{X^+, X^-\}$ is linearly separable by the hyperplane defined by $\mathbf{y}$

By construction, since $\alpha' \in (0,1)$, U must be between U_1 and U_2

By squeeze theorem U_1 and U_2 go through the origin$\Rightarrow$ U also goes through the origin

$\Rightarrow \{X^+, X^-\}$ is a dichotomy that is linearly separable by a hyperplane that goes through the origin and $\mathbf{x}_{m+1}$. This completes the proof for (A)$\Rightarrow$(B)

Now, to show (B)$\Rightarrow$(A)

Consider the family of planes defined by the set of vectors $\{\mathbf{y}(\epsilon) = (1-\epsilon)\mathbf{w} + \epsilon \mathbf{x}_{m+1} : \epsilon \in (0,1]\}$

Using this definition,

$\mathbf{y} \cdot \mathbf{x}_{m+1} = \epsilon ||\mathbf{x}_{m+1}||^2 > 0$

We'd want to find an ϵ where $\mathbf{y} \cdot \mathbf{x} > 0,\ \forall \mathbf{x} \in X^+$ and $\mathbf{y} \cdot \mathbf{x} < 0,\ \forall \mathbf{x} \in X^-$

Denote any point in X^+ as $\mathbf{x}^+$ and in X^- as $\mathbf{x}^-$:

For points in X^+ to be correctly classified,

$\mathbf{y} \cdot \mathbf{x}^+ = (1-\epsilon)\mathbf{w} \cdot \mathbf{x}^+ + \epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}^+ > 0$

$\Leftrightarrow (1-\epsilon)\mathbf{w} \cdot \mathbf{x}^+ > |\epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}^+|$

$\Leftrightarrow (1-\epsilon)/\epsilon > |\mathbf{x}_{m+1} \cdot \mathbf{x}^+|/\mathbf{w} \cdot \mathbf{x}^+ \qquad \because \epsilon > 0,\ \mathbf{w} \cdot \mathbf{x}^+ > 0$

$\Leftrightarrow 1/\epsilon > |\mathbf{x}_{m+1} \cdot \mathbf{x}^+|/\mathbf{w} \cdot \mathbf{x}^+ + 1$

$\Leftrightarrow \epsilon < [|\mathbf{x}_{m+1} \cdot \mathbf{x}^+|/\mathbf{w} \cdot \mathbf{x}^+ + 1]^{-1} = \delta_1 > 0$

Choose $\delta_1' = \min_{\mathbf{x}^+} \delta_1$

$\therefore$ if $\epsilon < \delta_1'$ then all the points in X^+ will still be correctly classified

Similarly, for points in X^- to be correctly classified,

$\mathbf{y} \cdot \mathbf{x}^- = (1-\epsilon)\mathbf{w} \cdot \mathbf{x}^- + \epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}^- < 0$
$\Leftrightarrow (1-\epsilon)\mathbf{w} \cdot \mathbf{x}^- < -|\epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}|$
$\Leftrightarrow (1-\epsilon)/\epsilon > -|\epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}|/\mathbf{w} \cdot \mathbf{x}^-$ $\quad \because \mathbf{w} \cdot \mathbf{x}^- < 0$
$\Leftrightarrow \epsilon < [-|\mathbf{x}_{m+1} \cdot \mathbf{x}^+|/\mathbf{w} \cdot \mathbf{x}^- + 1]^{-1} = \delta_2 > 0$
Choose $\delta_2' = \min_{\mathbf{x}^-} \delta_2$
$\therefore$ if $\epsilon < \delta_2'$ then all the points in X^- will still be correctly classified
So if we choose some $\epsilon' < \min(\delta_1', \delta_2') \Rightarrow$
$\mathbf{y}(\epsilon') \cdot \mathbf{x} > 0 \ \forall \mathbf{x} \in X^+$
$\mathbf{y}(\epsilon') \cdot \mathbf{x} < 0 \ \forall \mathbf{x} \in X^-$
$\mathbf{y}(\epsilon') \cdot \mathbf{x}_{m+1} > 0$
Therefore the dichotomy $\{X^+ \cup \{\mathbf{x}_{m+1}\}, X^-\}$ is linearly separable by $\mathbf{y}(\epsilon')$

Similar, to show that the dichotomy $\{X^+, X^- \cup \{\mathbf{x}_{m+1}\}$ is linearly separable, we use the family of planes defined by vectors in $\{\mathbf{z}(\epsilon) = (1-\epsilon)\mathbf{w} - \epsilon \mathbf{x}_{m+1} \ \epsilon \in (0,1]\}$
$\Rightarrow \mathbf{z} \cdot \mathbf{x}_{m+1} = -\epsilon||\mathbf{x}_{m+1}||^2 < 0$
The conditions:
$\mathbf{z} \cdot \mathbf{x}^+ = (1-\epsilon)\mathbf{w} \cdot \mathbf{x}^+ - \epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}^+ > 0$
and
$\mathbf{z} \cdot \mathbf{x}^- = (1-\epsilon)\mathbf{w} \cdot \mathbf{x}^- - \epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}^- < 0$
simplify to the same forms:
$(1-\epsilon)\mathbf{w} \cdot \mathbf{x}^+ > |\epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}^+|$
and
$(1-\epsilon)\mathbf{w} \cdot \mathbf{x}^- < -|\epsilon \mathbf{x}_{m+1} \cdot \mathbf{x}|$
So we can choose the same ϵ' as defined in the previous part and the following conditions will hold:
$\mathbf{z}(\epsilon') \cdot \mathbf{x} > 0 \ \forall \mathbf{x} \in X^+$
$\mathbf{z}(\epsilon') \cdot \mathbf{x} < 0 \ \forall \mathbf{x} \in X^-$
$\mathbf{z}(\epsilon') \cdot \mathbf{x}_{m+1} < 0$
Therefore the dichotomy $\{X^+ \cup \{\mathbf{x}_{m+1}\}, X^-\}$ is linearly separable by $\mathbf{z}(\epsilon')$, and this completes the proof for (B)$\Rightarrow$(A) $\square$

B.2. **Let $\mathbf{X}=\{\mathbf{x}_1, ..., \mathbf{x}_m\} \subset \mathbb{R}^d$ *s.t.* any d-element subset of X is linearly independent. Then the number of linearly separable labelings of X is $C(m,d) = 2\sum_{k=0}^{d-1}\binom{m-1}{k}$.**
The proof is by complete induction on $m+d$ and follow very closely to that of Sauer's lemma.

Base case:
For any m, $d=1$
$C(m,1) = 2\binom{m-1}{0} = 2$, which is to say the data points in $\mathbb{R}$ are degenerate and can only be labeled all +1 or all -1
For and d, $m=1$

$C(1,d) = 2\binom{0}{0} = 2$, for any dimension d, if there is only 1 point there can only be 2 ways of labeling it.

Inductive case: Assume all $m' + d' < m + d$ true.

Let the set of all linearly separable labelings of $X = \{\mathbf{x}_1, ..., \mathbf{x}_{m-1}\}$ be G and $|G| = C(m,d)$. Let $T = \{\mathbf{x}_1, ..., \mathbf{x}_{m-1}\}$ and denote the set of linearly separable labelings of this as G_T.

Construct set $G_2 = \{g' \subseteq T : (g' \in G) \land (g' \cup \{\mathbf{x}_m\} \in G)\}$. This is the set where each labeling is in the overall set G but not in G_T. Specifically, if $U \subseteq T$ where $|U| = d-1$ such that $U \cup \{\mathbf{x}_m\}$ admits only 1 possible labeling for $\mathbf{x}_m$ then this particular g' is in G_T, otherwise if both labeling are possible, then g' will be in both G_T and G_2. Using this definition, $|G| = |G_T| + |G_2|$.

Now we need to find $|G_T|$ and $|G_2|$. For any labeling of $X = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$, the labeling for any $m-1$ elements must be in G_T, and the maximum number of linearly separable points in G_T is still $d \Rightarrow |G_T| = C(m-1,d)$

And for any labeling of U, a $d-1$ element subset of X, to be in G_2, both possible labeling for $\mathbf{x}_m \in T \cup \{\mathbf{x}_m\}$ must be in $G \Rightarrow$ the maximum number of linearly separable points in G_2 must be $d-1 \Rightarrow |G_2| = C(m-1,d-1)$

Using the following properties:

$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}$

Using the induction hypothesis:

$C(m,d) = C(m-1,d) + C(m-1,d-1) = 2\sum_{k=0}^{d-1}\binom{m-2}{k} + 2\sum_{k=0}^{d-2}\binom{m-2}{k}$

$= 2\sum_{k=0}^{d-1}[\binom{m-2}{k} + \binom{m-2}{k-1}]$

$= 2\sum_{k=0}^{d-1}\binom{m-1}{k}$

which completes the proof $\square$

B.3. **Growth function of linear combination of linearly independent variable.**

$\mathcal{F} = \{x \mapsto \text{sgn}(\sum_{k=1}^{p} a_k f_k(x)) : a_1, ..., a_p \in \mathbb{R}\}$

Using results from the previous part, let $U = \{\Phi(x_1), ..., \Phi(x_m)\}$ where every p-subset is linearly independent. Then the total number of linearly separable labeling of U is $2\sum_{i=1}^{p-1}\binom{m-1}{i} \Rightarrow \Pi_{\mathcal{F}}(m) \leq 2\sum_{i=1}^{p-1}\binom{m-1}{i}$. We need to show that this relationship is strictly equal

By definition, since any p-element subset of U is linearly independent, for any such subset, $\exists \mathbf{a} \in \mathbb{R}^p$, choose this $\mathbf{a}$ to plug into the definition of $\mathcal{F}$ and we will be able to attain any combination labeling of U, and therefore all labelings of U are possible using $\mathcal{F}$ and the equality is satisfied. $\square$

C. Problem C

D. Problem D

D.1. **Show** $K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^N \cos^n(x_i^2 - y_i^2)\ \ \forall(\mathbf{x},\mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$ **is PDS.**

Using $\cos(\alpha-\beta) = \cos\alpha\cos\beta + \sin\alpha\sin\beta = \begin{bmatrix}\cos\alpha\\ \sin\alpha\end{bmatrix} \cdot \begin{bmatrix}\cos\beta\\ \sin\beta\end{bmatrix}$

Let $\Phi_i(\mathbf{x}) = \begin{bmatrix}\cos x_i^2\\ \sin x_i^2\end{bmatrix}$ and $K_i(\mathbf{x},\mathbf{y}) = \Phi_i(\mathbf{x}) \cdot \Phi_i(\mathbf{y})$ is PDS

$K(\mathbf{x},\mathbf{y}) = \sum_{i=1}^N [K_i(\mathbf{x},\mathbf{y})]^n \Rightarrow K(\mathbf{x},\mathbf{y})$ is PDS by closure of PDS kernels. $\square$

D.2. **Show** $K(\mathbf{x},\mathbf{y}) = \exp(-||\mathbf{x}-\mathbf{y}||/\sigma)\ \ \forall(\mathbf{x},\mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N$ **is PDS.**

Consider if $\sum_{i=1}^m c_i = 0$,

$\sum_{i,j=1}^m c_i c_j ||\mathbf{x}_i - \mathbf{x}_j|| = \sum_{i,j}^m c_i c_j \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j)}$

Case 1: $(\mathbf{x}_i - \mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j) \le 1$

$\sum_{i,j=1}^m c_i c_j ||\mathbf{x}_i - \mathbf{x}_j|| = \sum_{i,j} c_i c_j \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j)}$
$\quad \le \sum_{i,j} c_i c_j = \sum_i c_i \sum_j c_j = 0$

Case 2: $(\mathbf{x}_i - \mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j) > 1 \Rightarrow \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j)} \le (\mathbf{x}_i - \mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j)$

$\sum_{i,j=1}^m c_i c_j ||\mathbf{x}_i - \mathbf{x}_j|| = \sum_{i,j} c_i c_j \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j)}$
$\quad \le \sum_{i,j} c_i c_j (\mathbf{x}_i - \mathbf{x}_j)\cdot(\mathbf{x}_i - \mathbf{x}_j)$
$\quad = \sum_{i,j} c_i c_j (||\mathbf{x}_i||^2 + ||\mathbf{x}_j||^2 - 2\mathbf{x}_i \cdot \mathbf{x}_j)$
$\quad = \sum_{i,j} c_i c_j (||\mathbf{x}_i||^2 + ||\mathbf{x}_j||^2) - 2\sum_{i,j} c_i c_j (\mathbf{x}_i \cdot \mathbf{x}_j)$
$\quad = \sum_{i,j} c_i c_j (||\mathbf{x}_i||^2 + ||\mathbf{x}_j||^2) - 2\sum_i c_i \mathbf{x}_i \cdot \sum_j c_j \mathbf{x}_j$
$\quad = \sum_{i,j} c_i c_j (||\mathbf{x}_i||^2 + ||\mathbf{x}_j||^2) - 2||\sum_i c_i \mathbf{x}_i||^2$
$\quad \le \sum_{i,j} c_i c_j (||\mathbf{x}_i||^2 + ||\mathbf{x}_j||^2)$
$\quad = \sum_i c_i \sum_j c_j ||\mathbf{x}_j||^2 + \sum_j c_j \sum_i c_i ||\mathbf{x}_i||^2 = 0 \qquad \because \sum_i c_i = 0$

$\Rightarrow \sum_{i,j=1}^m c_i c_j ||\mathbf{x}_i - \mathbf{x}_j|| \le 0,\ \forall(\mathbf{x}_i,\mathbf{x}_j) \in \mathbb{R}^N \times \mathbb{R}^N$

$\Rightarrow$ $K_0(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||$ is an NDS kernel as defined in textbook.

$\Rightarrow$ $K(\mathbf{x},\mathbf{y}) = \exp(-K_0/\sigma)$, and $1/\sigma > 0 \Rightarrow K(\mathbf{x},\mathbf{y})$ is PDS. $\square$