Additional Exercises on PAC-Learning and VC-Dimension
(Problems are from previous years' exams)
CS-526 Learning Theory

Short problems

1. [Several correct answers are possible.] Let $\mathcal{H} = \{h_\theta\}_{\theta\in\Theta}$ be a hypothesis class such that $\text{VCdim}(\mathcal{H}) = +\infty$. Then the set of parameters Θ:
   (a) is finite
   (b) can be countable
   (c) can be uncountable
   (d) can be finite, countable or uncountable
2. Consider some hypothesis class $\mathcal{H}$. Which of the following is true? Why or why not?
   (a) If $|\mathcal{H}|$ is infinite, it is not PAC learnable.
   (b) If $\mathcal{H}$ is PAC learnable, it has finite VC dimension.
   (c) If $\mathcal{H}$ is specified by a finite number of parameters, it has finite VC dimension.
   (d) If $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$, where $\mathcal{H}_1$ and $\mathcal{H}_2$ are some hypothesis classes that are PAC learnable, then $\mathcal{H}$ is also PAC learnable.
3. Let $\mathcal{H}$ be the class of indicator functions defined by the intervals over $\mathbb{R}$, $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ where $h_{a,b}(x) = \mathbb{1}_{[x\notin(a,b)]}$. What is the VC dimension of $\mathcal{H}$?
4. Let $\mathcal{H}$ be the class of indicator functions defined by the intervals over $\mathbb{R}$, $\mathcal{H} = \{h_{a,b,c,d} : a, b, c, d \in \mathbb{R}, a < b, c < d\}$ where $h_{a,b,c,d}(x) = \mathbb{1}_{[x\in(a,b) \text{ OR } x\in(c,d)]}$. What is the VC dimension of $\mathcal{H}$?

VC dimension of unbiased neurons
Let $\mathcal{H} = \{h_{\alpha_1,\alpha_2}(\mathbf{x}) : \alpha_1, \alpha_2 \in \mathbb{R}\}$ with

$$h_{\alpha_1,\alpha_2}(\mathbf{x}) = \mathbb{I}\big(\tanh(\alpha_1 x_1 + \alpha_2 x_2) > 0\big)$$

1. What is $\text{VCdim}(\mathcal{H})$? Call your answer d.
2. Show that $\text{VCdim}(\mathcal{H}) \geq d$?
3. Show that $\text{VCdim}(\mathcal{H}) \leq d$?

VC dimension of union
Let $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_r$ be hypothesis classes over some fixed domain set $\mathcal{X}$. Let $d = \max_i \text{VCdim}(\mathcal{H}_i)$ and assume that $d > 2$.
Prove that:

1. $\text{VCdim}(\bigcup_{i=1}^{r} \mathcal{H}_i) \leq \frac{4d}{\log(2)} \log\left(\frac{2d}{\log(2)}\right) + \frac{2\log(r)}{\log(2)}$.
   Hint: Use Sauer's lemma for bounding the growth function and the inequality
   "Let $a \geq 1$ and $b > 0$. If $x \leq a\log(x) + b$ then $x \leq 4a\log(2a) + 2b$."

2. For $r = 2$ the bound can be strengthen to $\text{VCdim}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq 2d + 1$.
   Hint: $\sum_{i=0}^{k} \binom{k}{i} = 2^k$

Stability implies Generalization

Let $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ be a training dataset composed of n i.i.d. samples drawn from $\mathcal{D}$. As usual, we denote $L_{\mathcal{D}}(h) = E_{(x,y)\sim\mathcal{D}}[l(h(x), y)]$ and $L_S(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)$ the true and empirical risks of a hypothesis h, respectively. For simplicity, let us denote by h_S the output of a learning algorithm when trained with dataset S.

An important property of learning algorithms is their ability to generalize, i.e., the true and empirical risks of the output hypothesis should be close in expectation. Formally, we say that a learning algorithm $\mathcal{A}$ ϵ-generalizes in expectation if

$$|E_S[L_S(h_S) - L_{\mathcal{D}}(h_S)]| < \epsilon \,. \tag{1}$$

An interesting connection arises when we investigate the *stability* of a learning algorithm. Formally, we call a learning algorithm ϵ-*uniformly stable* if $\forall S, S'$ datasets of size n that differ in at most one example we have

$$\sup_{(x,y)} l(h_S(x), y) - l(h_{S'}(x), y) < \epsilon \,. \tag{2}$$

Notations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), (\widetilde{x}_1, \widetilde{y}_1), \ldots, (\widetilde{x}_n, \widetilde{y}_n)$ are $2n$ independently sampled training examples. We define $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $\widetilde{S} = \{(\widetilde{x}_1, \widetilde{y}_1), \ldots, (\widetilde{x}_n, \widetilde{y}_n)\}$ and $S^{(i)} = \{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (\widetilde{x}_i, \widetilde{y}_i), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}$.
Prove that:

1. $L_{\mathcal{D}}(h_S) = E_{\widetilde{S}}[\frac{1}{n}\sum_{i=1}^{n} l(h_S(\widetilde{x}_i), \widetilde{y}_i)]$.

2. $E_{S,\widetilde{S}}[l(h_S(\widetilde{x}_i), \widetilde{y}_i)] = E_{S,S^{(i)}}[l(h_{S^{(i)}}(x_i), y_i)]$.

3. An ϵ-uniformly stable learning algorithm ϵ-generalizes in expectation.

VC dimension of decision trees with binary features

In this problem, we consider the class $\mathcal{H}_{btree}$ of decision trees with binary features and binary labels. We have a set of samples $x^{(1)}, \ldots, x^{(m)}$, where $x^{(i)} \in \{0, 1\}^d$. A decision tree is a classifier that returns the binary label y for a sample x after performing a series of tests of the type "$x_i = 0$?" for $0 \leq i < d$, which are organized in a binary tree-like manner. Nodes of this tree correspond to the tests and leaves to the returned label values. Note that it is allowed to return the same label value from both branches.

1. Consider the subclass $\mathcal{H}_1$ of trees with a single decision node (see Fig. 1). Show that

$$\text{VCdim}\mathcal{H}_1 \leq \lfloor \log_2(d+1) \rfloor + 1.$$

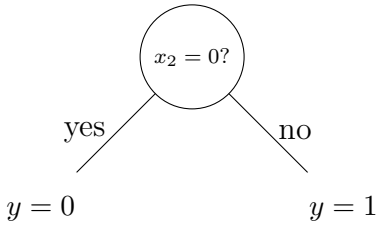

Figure 1: Example of single-node decision tree

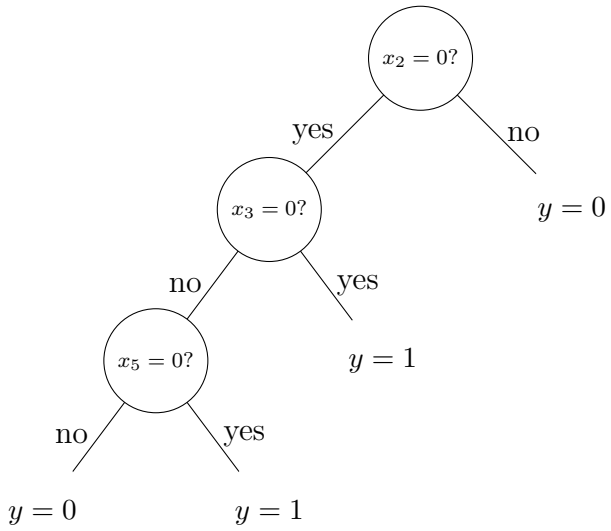

Figure 2: Example of degenerate tree with $N = 3$ nodes.

2. Show that
$$\text{VCdim}\mathcal{H}_1 \geq \lfloor \log_2(d+1) \rfloor + 1.$$

3. Consider the subclass $\mathcal{H}_{deg,N}$ of degenerate trees. Now the tree has N decision nodes but each node except the bottom one has a single child node (see Fig. 2). Prove that
$$\text{VCdim}\mathcal{H}_{deg,N} \geq \lfloor \log_2(d-N+2) \rfloor + N.$$

Hint: Start from the case $N = 1$. What changes when we add another node to the tree?

Expectation Learnability

Assume that the realizability assumption holds throughout the problem.

A hypothesis class $\mathcal{H}$ is Expectation learnable (E learnable) if there exists a function $m_{\mathcal{H}}^{(\text{E})} : (0,1) \to \mathbb{N}$ and a learning algorithm with the following property: For every $\gamma \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, when running the learning algorithm on a set S of $m \geq m_{\mathcal{H}}^{(\text{E})}(\gamma)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h (which depends on S) such that

$\mathbb{E}\big[L_{(\mathcal{D},f)}(h)\big] \leq \gamma$ (where the expectation is taken over the training set S). Recall that the error of a prediction is defined to be

$$L_{(\mathcal{D},f)}(h) := \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)].$$

1. Show that if a hypothesis class $\mathcal{H}$ is E learnable, then it is PAC learnable.
2. Show that if a hypothesis class $\mathcal{H}$ is PAC learnable, then it is E learnable.
3. Show that every finite hypothesis class $\mathcal{H}$ is E learnable with sample complexity

$$m_{\mathcal{H}}^{(\mathrm{E})}(\gamma) \leq \left\lceil \frac{2\log\left(\frac{2|\mathcal{H}|}{\gamma}\right)}{\gamma} \right\rceil.$$

Hint: You can use results proved in the course, and the relation between sample complexity of PAC learning and E learning derived in previous parts.