ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Spring 2026 — Assignment date: March 30, 2026, 08:15

Learning Theory — Due date: March 30, 2026, 10:00

Midterm Exam – CS 526 – Room INM 200

Use scratch paper if needed to figure out the solution. This exam is open-book (lecture notes, exercises, course materials). You can use the uploaded material on your computer but **switch off the wifi**. Good luck!

SCIPER No.: _____

Problem 1	/ 20
Problem 2	/ 20
Problem 3	/ 20
Total	/60

Problem 1 (20 pts)**.** *Short questions*

Justify all your answers.

1. Let $\mathcal{H} = \{h_\theta\}_{\theta \in \Theta}$ be a hypothesis class parameterized by a single variable θ and suppose that $\text{VCdim}(\mathcal{H}) = +\infty$. Then Θ, the set of all possible values that the parameter θ can take:

   (a) is finite

   (b) can be countable

   (c) can be uncountable

   (d) can be finite, countable or uncountable

   Several answers can be possible.

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 $\mathcal{H}$?

Solution to Problem 1 (total 20 pts)**:**

1. (2 pt) **B and C**

   The set Θ parametrizing the hypothesis class must be infinite: if $\mathcal{H}$ has finite cardinality then $\text{VCdim}(\mathcal{H}) \leq \log|\mathcal{H}|$. In the second graded homework, we studied the hypothesis class $\mathcal{H} = \{\lceil \sin(\theta\pi.)\rceil\}_{\theta\in\Theta}$ and proved that it has an infinite VC dimension if $\Theta = \{2n\}_{n\in\mathbb{N}}$ (and by extension $\theta = \mathbb{R}$). Therefore B and C are correct.

2. (8 pt)

   (a) False. If $\mathcal{H}$ has finite VC dimension then it is PAC learnable due to the Fundamental theorem of Statistical learning.

   (b) True. According to the Fundamental theorem of Statistical learning.

   (c) False. We saw in the homework that there are hypotheses classes with infinite VC dimension that are specified by a single parameter.

   (d) True. If $\mathcal{H}_1, \mathcal{H}_2$ have finite VC dimension then the VC dimension of their union is also finite and therefore $\mathcal{H}$ is also PAC learnable.

3. (5 pt) The VC dimension is 2: A set of size 2 can be shattered by $\mathcal{H}$, but for a set of size 3 with elements $x_1 < x_2 < x_3$ the labeling $(0, 1, 0)$ cannot be obtained by any $h_{a,b} \in \mathcal{H}$. Therefore, the VC dimension is 2.

4. (5 pt)The VC dimension is 4: A set of size 4 can be shattered, but a set of size 5 with elements $x_1 < \ldots < x_5$ with labels $(1, 0, 1, 0, 1)$ cannot be obtained by any $h_{a,b,c,d} \in \mathcal{H}$. Therefore, the VC dimension is 4.

Problem 2 (20 pts)**.** *(k+1) folds validation*

Let $\mathcal{H}$ be a infinite hypothesis class and let A be a learning algorithm with the following guarantee:

There exists a constant $\delta_0 \in (0,1)$ and a function $m_{\mathcal{H}} : (0,1) \to \mathbb{N}$ such that for every $\epsilon \in (0,1)$, if $m \geq m_{\mathcal{H}}(\epsilon)$, then for every distribution $\mathcal{D}$,

$$\Pr_S \left(L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right) \geq 1 - \delta_0.$$

We want to construct a new learning algorithm B that uses A as a subroutine and learns $\mathcal{H}$ in the usual agnostic PAC learning model with a better confidence while keeping the error guarantee ϵ. We want B to achieve confidence $1 - \delta$ for any $\delta \leq \delta_0$, with sample complexity $m \geq m_{\mathcal{H}}(\epsilon, \delta)$.

Learning algorithm B:

- Define $k = \left\lceil \frac{\log(\delta/2)}{\log(\delta_0)} \right\rceil$.
- Given a training set S, partition it into $k+1$ disjoint subsets:
  - $S_1, \ldots, S_k$, each of size $m_{\mathcal{H}}(\epsilon/2)$,
  - S_{k+1} of size m'.
- For each $i = 1, \ldots, k$, run A on S_i to obtain hypotheses $h_i = A(S_i)$.
- Use the validation set S_{k+1} to select $\hat{h} = \arg\min_{h_i} L_{S_{k+1}}(h_i)$.

1. Show that with probability at least $1 - \delta/2$, there exists at least one h_i such that:

$$L_{\mathcal{D}}(h_i) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon/2.$$

2. Using results from the class, find a bound on the sample complexity m' that guarantees that with probability at least $1 - \delta/2$, for all i

$$\left| L_{S_{k+1}}(h_i) - L_{\mathcal{D}}(h_i) \right| \leq \epsilon/4.$$

3. Conclude that the learning algorithm B is indeed an agnostic PAC learner with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq k \cdot m_{\mathcal{H}}(\epsilon/2) + m'.$$

Solution to Problem 2:

1. (6 pt)The probability that all k hypotheses are bad is:

$$\Pr\left(\forall i,\, L_{\mathcal{D}}(h_i) > \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\right) \leq \delta_0^k.$$

   By definition of k:

$$\delta_0^k \leq \delta/2.$$

   Thus, with probability at least $1-\delta/2$, there exists at least one h_i such that:

$$L_{\mathcal{D}}(h_i) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon/2.$$

2. (6 pt) Using the sample complexity of the uniform convergence of finite classes, if

$$m' \geq \left\lceil \frac{8\log(4k/\delta)}{\epsilon^2} \right\rceil,$$

   then with probability at least $1-\delta/2$, for all i:

$$\left|L_{S_{k+1}}(h_i) - L_{\mathcal{D}}(h_i)\right| \leq \epsilon/4.$$

3. (8 pt) We want the intersection of the two high-probability events (each holding with probability $\geq 1-\delta/2$):

   (a) There exists h_{i^*} such that

$$L_{\mathcal{D}}(h_{i^*}) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon/2.$$

   (b) The empirical risks approximate true risks well, $\left|L_{S_{k+1}}(h_i) - L_{\mathcal{D}}(h_i)\right| \leq \epsilon/4$.

   Thus, with probability at least $1-\delta$, the ERM selection over $\{h_1,\ldots,h_k\}$ yields $\hat{h}$ such that $L_{S_{k+1}}(\hat{h}) \leq L_{S_{k+1}}(h_{i^*})$ and

$$L_{\mathcal{D}}(\hat{h}) \underset{(b)}{\leq} L_{S_{k+1}}(\hat{h}) + \epsilon/4 \leq L_{S_{k+1}}(h_{i^*}) + \epsilon/4 \underset{(b)}{\leq} L_{\mathcal{D}}(h_{i^*}) + \epsilon/2 \underset{(a)}{\leq} \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

   The total sample size is $m = k \cdot m_{\mathcal{H}}(\epsilon/2) + m'$. Hence:

$$m_{\mathcal{H}}(\epsilon,\delta) \leq k \cdot m_{\mathcal{H}}(\epsilon/2) + \left\lceil \frac{8\log(4k/\delta)}{\epsilon^2} \right\rceil.$$

Problem 3 (20 pts)**.** *Moore Penrose pseudoinverse*

Consider the linear regression problem:

$$\min_{\beta \in \mathbb{R}^d} \|X\beta - Y\|^2$$

where $X \in \mathbb{R}^{m \times d}$ and $Y \in \mathbb{R}^m$. We will generalize the expression of the least-squares minimizer obtained in class, without assuming that either $X^T X$ or XX^T is full rank.

Properties: Let $A \in \mathbb{R}^{m \times d}$. Recall that the Moore–Penrose pseudoinverse $A^\dagger$ is defined as the unique matrix satisfying:

$$(1)\ AA^\dagger A = A, \qquad (2)\ A^\dagger AA^\dagger = A^\dagger, \qquad (3)\ (AA^\dagger)^T = AA^\dagger, \qquad (4)\ (A^\dagger A)^T = A^\dagger A.$$

1. Show that the matrix $P = I - AA^\dagger$ is an orthogonal projector, i.e., prove that $P^2 = P$. Show also that the kernel (or Null space) of P is the image space of A.

2. Using the previous result, prove that

   $$\beta^\star = X^\dagger Y$$

   is a minimizer of this problem.

Solution to Problem 3 (total 20 pts)**:**

1. (10 pt) Let $P = I - AA^\dagger$. Compute:

$$\begin{aligned} P^2 &= (I - AA^\dagger)(I - AA^\dagger) \\ &= I - 2AA^\dagger + AA^\dagger AA^\dagger. \end{aligned}$$

   Using property (2), $A^\dagger AA^\dagger = A^\dagger$: $P^2 = I - 2AA^\dagger + AA^\dagger = I - AA^\dagger = P$. Thus, P is a projector.

   Let x in the kernel of P. Thus, $x = AA^\dagger x$ so $x \in \mathrm{Im}(A)$ giving $\ker(P) \subset \mathrm{Im}(A)$. Now observe that:

$$PA = A - AA^\dagger A = 0.$$

   using property (1). Thus, $\mathrm{Im}(A) \subset \ker(P)$ and $\mathrm{Im}(A) = \ker(P)$.

2. (10 pt) Let $\beta \in \mathbb{R}^d$. We decompose:

$$||X\beta - Y||^2 = ||X\beta + XX^\dagger Y - XX^\dagger Y - Y||^2 = ||X(\beta - X^\dagger Y) - (I - XX^\dagger)Y||^2.$$

   Let $P = I - XX^\dagger$. The two vectors $X(\beta - X^\dagger Y) \in \ker(P)$ and $PY \in \mathrm{Im}(P)$ are orthogonal. Hence:

$$||X\beta - Y||^2 = \|X(\beta - X^\dagger Y)\|^2 + \|PY\|^2.$$

   Since $\|PY\|^2$ does not depend on β, the norm is minimized when:

$$X(\beta - X^\dagger Y) = 0.$$

   This is achieved for $\beta = X^\dagger Y$. Therefore:

$$\beta^\star = X^\dagger Y$$

   is a minimizer of the least-squares problem.