Machine Learning from Data - IDC
HW6 – Theory

Instructions: This assignment must be submitted individually, submitting in pairs is not allowed.

1 VC-Dimension

Compute the VC-dimension of the following hypothesis classes:

1. (10 pt.) Assume the instance space $\mathcal{X}$ satisfies $|\mathcal{X}| = \infty$. The space of binary hypotheses, which given a training set, returns the target y of $\mathbf{x}$ if the pair $(\mathbf{x}, y)$ was observed in the training set, and $+1$ otherwise. Formally, compute $VC(\mathcal{H})$ of

$$\mathcal{H} = \{h : \mathcal{X} \to \{-1, +1\} : h \text{ equals } -1 \text{ on a finite subset of } \mathcal{X} \text{ and } +1 \text{ elsewhere}\}\,.$$

Solution: For any set of size m, $A = (z_1, \ldots, z_m)$, and any dichotomy, choose $h \in \mathcal{H}$ such that $h(z_i) = -1 \iff y_i = -1$. This classifier is in $\mathcal{H}$ since there are only finitely many examples with label -1 in A, and everywhere else h returns $+1$. Therefore $\mathrm{VCdim}(\mathcal{H}) = \infty$.

2. (15 pt.) n-Interval classifiers of length ≥ 2. Let $\mathcal{X} = \mathbb{R}$,

$$\mathcal{H} = \{x \mapsto +1 \iff x \in [a_1, b_1] \cup [a_2, b_2] \cup \cdots \cup [a_n, b_n] : a_1 + 2 \leq b_1, \ldots, a_n + 2 \leq b_n\}\,.$$

Solution: Since a single interval shatters any 2 instances (see the VC of interval predictors proof), we can use n intervals to shatter the set $A = \{0, 1, \ldots, 2n\}$ by assigning each interval to a pair of adjacent instances, and using the interval to obtain their two labels, effectively obtaining all possible dichotomies on A. Note that this is possible since that each interval does not affect the prediction obtained on any other instances but its assigned pair.

For any set of size $2n+1$, $A = \{z_1, z_2, \ldots, z_{2n+1}\}$, assume w.l.o.g. that it is ordered $z_1 < z_2 < \ldots < z_{2n+1}$. Then the dichotomy $(+1, -1, +1, -1, \ldots, -1, +1)$ cannot be attained as we're trying to positively classify $n+1$ instances with only n intervals, which by the pigeonhole principle implies that there are two positive instances classified by the same interval, which in turn implies that the negative instance between them must also be positively labeled. Thus $\mathrm{VCdim}(\mathcal{H}) < 2n + 1$, hence $\mathrm{VCdim}(\mathcal{H}) = 2n$.

3. (20 pt.) Linear classifiers in the plain. Let $\mathcal{X} = \mathbb{R}^2$,

$$\mathcal{H} = \left\{ (x_1, x_2) \mapsto \begin{cases} +1 & w_1x_1 + w_2x_2 + b > 0 \\ -1 & w_1x_1 + w_2x_2 + b \leq 0 \end{cases} : w_1, w_2, b \in \mathbb{R} \right\}.$$

Show that $\text{VCdim}(\mathcal{H}) = 3$:

(a) Find a set of size 3 that $\mathcal{H}$ shatters.

Solution: Choose $A = \{(0,0), (0,1), (1,0)\}$, then solving the linear system

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

for each of the 8 dichotomies $y_1, y_2, y_3 \in \{-1, +1\}$ produces a linear classifier for it (note the matrix is invertible therefore a solution always exists).

(b) Show that no set of size 4, $A = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4)$, $\mathbf{z}_i \in \mathbb{R}^2$ can be shattered by $\mathcal{H}$.
Guidance: First prove the following lemma:

Lemma 1. *Suppose a linear classifier h obtains prediction $y \in \{-1, +1\}$ on a set of points $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^2$ $(h(\mathbf{z}) = h(\mathbf{z}') = y)$. Then it also obtains the same prediction on any intermediate point. Namely,*

$$\forall \alpha \in [0,1] \quad h((1-\alpha)\mathbf{z} + \alpha\mathbf{z}') = y.$$

Solution:

Proof of the lemma. Suppose w.l.o.g. that $h(\mathbf{z}) = h(\mathbf{z}') = +1$ (the proof is the same for -1). Then

$$\begin{aligned} h((1-\alpha)\mathbf{z} + \alpha\mathbf{z}') &= \text{sign}(\langle \mathbf{w}, (1-\alpha)\mathbf{z} + \alpha\mathbf{z}' \rangle + b) \\ &= \text{sign}((1-\alpha)\langle \mathbf{w}, \mathbf{z} \rangle + \alpha \langle \mathbf{w}, \mathbf{z}' \rangle + (1-\alpha)b + \alpha b) \\ &= \text{sign}((1-\alpha)\underbrace{(\langle \mathbf{w}, \mathbf{z} \rangle + b)}_{>0} + \alpha\underbrace{(\langle \mathbf{w}, \mathbf{z}' \rangle + b)}_{>0}) = +1. \end{aligned}$$

□

And use it in each of the following 3 possible cases:

- The convex hull of A forms a line.
  Suppose $\mathbf{z}_1, \mathbf{z}_2$ are on the edges of the line, then classifying $\mathbf{z}_1, \mathbf{z}_2$ as $+1$ implies from the Lemma that $\mathbf{z}_3, \mathbf{z}_4$ are also classified as $+1$, hence the dichotomy $(+1, +1, -1, -1)$ is not possible.
- The convex hull of A forms a triangle.
  Suppose $\mathbf{z}_4$ is in the interior of the triangle formed by $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$. Then the dichotomy $(+1, +1, +1, -1)$ is not possible since from the Lemma the boundary of the triangle is also classified as $+1$ and therefore also its interior.

- The convex hull of A forms a quadrilateral.
  Suppose the quadrilateral is ordered by $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4$, then the dichotomy $(-1, +1, -1, +1)$ is not possible since from the lemma the diagonal formed by $\mathbf{z}_1, \mathbf{z}_3$ is negatively classified, and the diagonal formed by $\mathbf{z}_2, \mathbf{z}_4$ is positively classified, which results in the contradiction of their intersection point being classified by both -1 and $+1$.

2 Learning Conjunctions of Literals

(30 pt.) Let $\mathcal{X} = \{0,1\}^n$ (all boolean strings of length n), let $C = \mathcal{H} =$ the set of all conjunctions on $\mathcal{X}$ (e.g. $x_1 \wedge \neg x_3 \wedge x_n$ is in C and $\mathcal{H}$). Define an algorithm L so that C is PAC-learnable by L using $\mathcal{H}$. Prove all your steps.

Solution: We start with a CNF containing all $2n$ literals, which rejects all inputs, and we gradually remove literals from it by going over the positively labeled instances in the data. Given a positive instance, we remove all literals that receive a value of 0 once assigned by the instance (i.e. if $x_i = 0$ we remove x_i from our CNF, if x_i=1 we remove $\neg x_i$ from our CNF).

Clearly, any positive instance we go over will receive a positive classification once the CNF is updated, and since our algorithm only removes literals which causes the CNF to accept more inputs, all the positive instances will be accepted by the resulting CNF returned by the algorithm.

As for the negatively labeled instances, since these are generated by some concept CNF, and since that our algorithm only removes literals that are mandatory in order to accept the positive instances, we are left with the maximal (in terms of the number of literals it has) CNF consistent with all positive examples, and since adding literals causes the CNF to accept less inputs, we are left with a CNF which is also consistent with the negative instances.

Clearly, our algorithm is time efficient. Moreover, it produces a consistent hypothesis. Plugging in the size of the hypotheses space the algorithm uses (the class of all CNFs on n literals) which is upper bounded by 3^n to the formula seen in class we obtain sample complexity

$$m \geq \frac{1}{\epsilon}\left(n \ln 3 + \ln\left(\frac{1}{\delta}\right)\right),$$

which is polynomial in all required parameters, thus establishing PAC-learnability.

3 (Almost) PAC-learnability

(25 pt.) Let C denote the class of all possible target concepts defined over a set of instances $\mathcal{X}$. Suppose that $\mathcal{H}$ is a space of binary hypotheses containing the constant concept c_1 defined by $c_1(x) = +1$ for all $x \in \mathcal{X}$, and having the property that $C \setminus \{c_1\}$ is PAC-learnable by an algorithm L using $\mathcal{H}$ with sample complexity $m(\delta, \epsilon)$.

Provide a learning algorithm L' so that C (including c_1) is PAC-learnable by L' using $\mathcal{H}$ with sample complexity $\max\{m(\delta,\epsilon), \lceil\frac{\log(1/\delta)}{\epsilon}\rceil\}$. Prove all your steps.

Solution: Our algorithm will check the training set S for instances labeled with -1. If none exist, it returns $h = c_1$, otherwise it runs L to obtain hypothesis $h = L(S)$ and return it.

Claim: $\mathcal{H}$ is PAC-learnable with sample complexity $m'(\delta,\epsilon) = \max\{m(\delta,\epsilon), \lceil\frac{\log(1/\delta)}{\epsilon}\rceil\}$.

Proof of the Claim. Consider 3 possible cases:

- If the target concept we're trying to learn is c_1, then clearly all the instances in the training set will be labeled with $+1$, and the algorithm will return $h = c_1$ to obtain 0 generalization error.

- Otherwise, $p = \Pr_{x\sim\mathcal{D}}[c(x) = -1] > 0$, and if the training set contains an instance labeled -1 then the algorithm will return $L(S)$ which by the above property will also work whenever the sample size is $\geq m(\delta,\epsilon)$, since the concept we're trying to learn is not c_1.

- The only problem that might arise is when $p > 0$ and we still get a training set labeled with only $+1$, in which case the constant concept c_1 is returned which obtains generalization error
$$\mathrm{error}_{\mathcal{D}}(c_1) = p.$$
If $p \leq \epsilon$ we are fine. Otherwise $p > \epsilon$, and the probability of obtaining the all $+1$ dichotomy with m examples is
$$(1-p)^m < (1-\epsilon)^m \leq \left(1 - \frac{\log(1/\delta)}{m}\right)^m \leq \exp\left(-m\frac{\log(1/\delta)}{m}\right) = \delta,$$
where the second inequality is due to
$$\epsilon \geq \frac{\log(1/\delta)}{m} \iff m \geq \frac{\log(1/\delta)}{\epsilon},$$
which holds by our assumed sample complexity. We conclude that the probability of drawing at least a single -1 label is $\geq 1-\delta$, concluding the proof of the Claim.

□