TD12 : Apprentissage

1 Apprendre les rectangles alignés sur les axes de base

On considère la classe d'hypothèses des rectangles de dimension 2 alignés sur les axes de base : $\mathcal{H}^2_{rec} = \{h_{(a_1,b_1,a_2,b_2)} \mid a_1 \leq b_1 \text{ et } a_2 \leq b_2\}$ où :

$$h_{(a_1,b_1,a_2,b_2)}(x_1,x_2) = \begin{cases} 1 & \text{si } a_1 \leq x_1 \leq b_1 \text{ et } a_2 \leq x_2 \leq b_2 \\ 0 & \text{sinon} \end{cases}$$

Nous voulons montrer que cette classe peut être apprise par un algorithme respectant la régle ERM. Nous supposerons que la distribution $\mathcal{D}$ sur les données est continue et que la classe $\mathcal{H}^2_{rec}$ est réaliste.

Question 1 : Proposez un algorithme A d'apprentissage de $\mathcal{H}^2_{rec}$ qui respecte la règle ERM.

Question 2 : Montrez que si A reçoit une donnée d'apprentissage de taille $m \geq \frac{4\log(4/\delta)}{\epsilon}$ et si H_m est la sortie correspondante de A, on a $\mathbf{Pr}(L(H_m) > \epsilon) \leq \delta$.

Question 3 : Que pouvez-vous en conclure pour la classe des rectangles alignés sur les axes de dimension $d \geq 2$.

Question 4 : Quelle est la complexité en temps de l'algorithme A ?

Question 5 : Que vaut $\mathsf{VCdim}(\mathcal{H}^2_{rec})$?

Question 6 : Que vaut $\mathsf{VCdim}(\mathcal{H}^d_{rec})$ pour $d \geq 2$?

2 Apprendre la classe des singletons

Soit $\mathcal{X}$ un domaine discret. On considère la classe d'hypothèses $\mathcal{H}_{sing} = \{h_z \mid z \in \mathcal{X}\} \cup \{h^-\}$ où pour tout $x \in \mathcal{X}$, on a $h_z(x) = 1$ si $x = z$ et $h_z(x) = 0$ sinon et $h^-(x) = 0$. On suppose que la classe $\mathcal{H}_{sing}$ est réaliste.

Question 1 : Proposez un algorithme A d'apprentissage de $\mathcal{H}_{sing}$ qui respecte la règle ERM.

Question 2 : Montrez que $\mathcal{H}_{sing}$ peut être apprise et donnez une borne supérieure sur la taille des données d'apprentissage.

3 VC dimensions de conjonctions booléennes

On considère $\mathcal{H}^d_{con}$ la classe des conjonctions booléennes sur les variables $x_1, x_2, \ldots, x_d$ avec $d \geq 2$. Nous savons déjà que cette classe est finie. Nous allons calculer ici sa VC dimension.

Question 1 : Montrez que $|\mathcal{H}^d_{con}| \leq 3^d + 1$.

Question 2 : En déduire une borne sur $\mathsf{VCdim}(\mathcal{H}^d_{con})$.

Question 3 : Montrez que $\mathcal{H}^d_{con}$ brise l'ensemble des vecteurs de l'unité $\{\mathbf{e}_i \mid 1 \leq i \leq d\}$ (on assimile ici 1 à vrai et 0 et faux comme cela est fait habituellement).

Question 4 : Montrez que $\mathsf{VCdim}(\mathcal{H}^d_{con}) \leq d$.

Question 5 : On considère la classe $\mathcal{H}^d_{mcon}$ des conjonctions booléennes monotones sur les variables $x_1, x_2, \ldots, x_d$ avec $d \geq 2$. La monotonicité signifiant dans ce contexte que les conjonctions ne contiennent pas de négation. La conjunction vide est supposée égale à vrai. Nous ajoutons à $\mathcal{H}^d_{mcon}$ l'hypothèse h^- qui est toujours fausse. Montrez que $\mathsf{VCdim}(\mathcal{H}^d_{con}) = d$.

4 Quelques questions en vrac

Question 1 : VC dimension d'unions. Soient $\mathcal{H}_1, \ldots, \mathcal{H}_r$ des classes d'hypothèse sur un ensemble fixe d'objets $\mathcal{X}$. Soit $d = \mathsf{max}_{1 \leq i \leq r} \mathsf{VCdim}(\mathcal{H}_i)$ et supposons que $d \geq 3$.

1. Soit $a > 0$. Montrez que si $x \geq 2a\mathsf{log}(a)$ alors $x \geq a\mathsf{log}(x)$.
2. Soient $a \geq 1$ and $b > 0$. Si $x \geq 4a\mathsf{log}(2a) + 2b$ alors $x \geq a\mathsf{log}(x) + b$.
3. Montrez que $\mathsf{VCdim}(\bigcup_{i=1}^{r} \mathcal{H}_i) \leq 4d\mathsf{log}(2d) + 2\mathsf{log}(r)$. (On pourra utiliser le Lemme 16 du cours).
4. Montrez que, dans le cas où $r = 2$, on a $\mathsf{VCdim}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq 2d + 1$.

Question 2 : Soient $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$ et soit $\mathcal{H}$ la classe des disques concentriques dans le plan, c'est-à-dire $\mathcal{H} = \{h_r \mid r \in \mathbb{R}_+\}$ où $h_r(x) = \mathbb{1}_{\{||x|| \leq r\}}$. On suppose la distribution $\mathcal{D}$ continue et la classe $\mathcal{H}$ réaliste. Montrez que $\mathcal{H}$ peut être apprise et donnez une borne supérieure sur la taille de la donnée d'apprentissage.

Question 3 : Démontrez la proposition 27 du cours dans le cas d'une distribution quelconque (c'est à dire que l'on ne suppose plus que $\epsilon_0 > 0$ ni que la distribution est continue.