ASI - APPC - Examen final - **3 h**

Exercise 1 — L'épreuve du réel — 10 points

Le but de cet exercice est d'écrire la fonction en python, R ou matlab permettant de prédire l'étiquette (0 où 1) une entrée future x_f, un vecteur de 64 composants. Pour ce faire, nous disposons des données présentes dans le fichier `digits` disponible dans l'API sklearn qui contient :

— les entrées X
— les étiquettes y

Pour accéder à ces données, vous pouvez utiliser les instructions suivantes :

```
from sklearn.datasets import load_digits

X,y = load_digits(n_class=10, return_X_y=True)
```

```
whos

Variable       Type         Data/Info
-----------------------------------
X              ndarray      1797x64: 115008 elems, type 'float64', 920064 bytes (898.5 kb)
load_digits    function     <function load_digits at 0x120f98048>
y              ndarray      1797: 1797 elems, type 'int64', 14376 bytes
```

1. Traitement des données en mode non supervisé (en ignorant les étiquettes y)
   a) Visualisez l'ensemble des données.
   b) Proposez un découpage non supervisée des données en groupes homogènes.
2. Traitement des données en mode supervisé (en utilisant les étiquettes y). Donner un programme (en python, R ou matlab) permettant de prédire si une entrée future x_f, un vecteur de 64 composants représente un chiffre pair ou impair. On insistera sur la méthodologie mise en œuvre et sur la mesure des performances.

Exercise 2 — bis repetita placent — 4 points

Vous venez d'arriver en stage, votre patron vous demande de l'aider à définir la stratégie de l'entreprise par rapport à l'apprentissage statistique

1. Qu'est-ce que l'apprentissage statistique et à quoi ça sert ?
2. Quelles sont les applications principales de l'apprentissage ?
3. Quelles sont les méthodes (algorithmes d'apprentissage) les plus importantes et pourquoi ?
4. Comment faire pour mettre en œuvre un projet avec de l'apprentissage

---

Exercise 3 — Cerise sur le gâteaux — 6 points

L'inégalité de Hoeffding's stipule que, pour toute suite de variables aléatoires $X_1, \ldots, X_i \ldots, X_n$ réelles à valeur chacune dans un intervalle $[a_i, b_i]$ et indépendantes, $\forall t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]\right| > t\right) \leqslant 2\exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) .$$

Supposons que (X, Y) est un couple de variables aléatoire à valeur dans $\mathbb{R}^n \times \{-1, 1\}$. L'un des objectifs de la classification supervisée est de définir une fonction $h : \mathbb{R}^n \; to\{-1, 1\}$, appelée *classifier*, telle que $h(X)$ soit la meilleure prédiction de Y dans un contexte donné. Dans ce contexte, la probabilité de mauvaise classification de h est :

$$L(h) = \mathbb{P}\left(Y \neq h(X)\right) .$$

Supposons qu'il existe une fonction $\eta : \mathbb{R}^n \to [-1, 1]$ telle que $\mathbb{E}[Y|X] = \eta(X)$.

1. Montrer que la règle de décision $h_\star$, définie $\forall x \in \mathbb{R}^n$, par

$$h_\star(x) = \begin{cases} 1 & \text{si } \eta(x) > 0\,, \\ -1 & \text{sinon }, \end{cases}$$

est tel que

$$h_\star = \underset{h:\mathbb{R}^n \to \{-1,1\}}{\operatorname{argmin}} \; L(h)\,.$$

2. En pratique, la minimisation de L s'effectue par rapport à un ensemble spécifique $\mathcal{H}$ de classificateurs (souvent appelé em dictionnaire, qui peut ne pas contenir le classificateur Bayes. De plus, la probabilité d'erreur L ne peut être ni calculé ni minimisé. On la remplace par un risque de classification empirique défini par

$$\widehat{L}^n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \neq h(X_i)\}}\,,$$

où $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ sont des observations indépendantes ayant la même distribution que (X, Y) et $\mathbb{1}_{\{A\}}$ la fonction indicatrice de l'expression A, $\mathbb{1}_{\{A\}} = 1$, si A est vraie et zéro sinon. Le problème de classification se résume alors à la résolution de :

$$\widehat{h}^n_{\mathcal{H}} \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \widehat{L}^n(h)\,.$$

Montrer que pour tout ensemble $\mathcal{H}$ de classifieur et tout $n \geqslant 1$,

$$L(\widehat{h}^n_{\mathcal{H}}) - L(\widetilde{h}) \leqslant 2 \sup_{h \in \mathcal{H}} \left| \widehat{L}^n(h) - L\,(h) \right|\,,$$

où

$$\widetilde{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; L(h)\,.$$

3. En utilisant l'inégalité de Hoeffding montrer que pour $\mathcal{H} = \{h_1, \ldots, h_M\}$ pour $M \geqslant 1$ donné, on a $\forall \delta > 0$,

$$\mathbb{P}\left(L(\widehat{h}^n_{\mathcal{H}}) \leqslant \min_{1 \leqslant j \leqslant M} L(h_j) + \sqrt{\frac{2}{n} \log\left(\frac{2M}{\delta}\right)} \right) \geqslant 1 - \delta\,.$$

4. Que concluez vous lorsque la taille du dictionnaire et celui de l'échantillon augmentent ?