1 Méthode des couples (1h)

Soit un réseau de neurones RN et un arbre de décision AD utilisés pour faire de la classification automatique de textes (utile pour la gestion automatique de mails selon le contenu par exemple). Ces modèles ont été appris à partir d'une même base d'apprentissage B et l'on désire savoir lequel des deux modèles est le plus performant pour classifier des documents qui n'ont jamais été vus au cours de l'apprentissage (on veut donc mesurer le pouvoir de généralisation de ces deux modèles). Pour cela, on a mis en place un processus de validation croisée sur 10 ensembles : on a divisé la base d'apprentissage en 10 sous-bases disjointes $B_1, \ldots, B_{10}$ de même taille (si B contient ℓ exemples, $B_1, \ldots, B_{10}$ contiennent chacune $\ell/10$ exemples) et on a réalisé le processus qui suit. On a appris RN et AD en utilisant tous les exemples contenus dans $B_2, \ldots, B_{10}$ puis mesuré la performance de classification de RN et AD sur les exemples de B_1, puis reporté la valeur dans la première ligne du tableau ci-dessous. On a effectué le même processus en apprenant sur $B_1, B_3, \ldots, B_{10}$ et en testant sur B_2, la performance obtenue étant reportée dans le deuxième ligne du tableau ci-dessous[1]. On réitère cette procédure jusqu'à B_{10}.

Base	RN	AD
B_1	88.4	88.2
B_2	88.3	87.9
B_3	89.0	89.4
B_4	89.1	88.9
B_5	89.3	88.3
B_6	88.8	88.6
B_7	89.4	88.3
B_8	89.9	89.2
B_9	89.2	88.7
B_{10}	89.3	89.4

Tab. 1 – Performance du réseau de neurones (RN) et de l'arbre de décision sur chacune des bases en utilisant le processus de validation croisée décrit dans l'énoncé.

Dans la suite, on supposera que $X_1, \ldots, X_{n_x}$ sont des variables aléatoires i.i.d de même loi de moyenne μ_x et $Y_1, \ldots, Y_{n_y}$ des variables aléatoires i.i.d de même loi de moyenne μ_y ; on suppose que la performance mesurée pour RN sur la base B_i est une réalisation de la variable aléatoire X_i et que, de la même manière, la performance mesurée pour AD sur la base B_i est une réalisation de la variable aléatoire Y_i (on notera que pour les applications numériques $n_x = n_y = n = 10$). S_x^2 et S_y^2 désignent respectivement les variances empiriques des échantillons $X_1, \ldots, X_{n_x}$ et $Y_1, \ldots, Y_{n_y}$ et $\bar{X}$ et $\bar{Y}$ les moyennes empiriques de ces échantillons.

On s'intéresse au test d'hypothèse $H_0 = \{\mu_x = \mu_y\}$ contre $H_1 = \{\mu_x > \mu_y\}$.

1. Calculer les moyennes et variances empiriques des performances reportées dans le tableau 1 pour RN et AD.
2. Supposons que $X_i \sim \mathcal{N}(\mu_x, \sigma^2), \forall i$, $Y_i \sim \mathcal{N}(\mu_y, \sigma^2), \forall i$.
   (a) Quelles sont les lois de $\bar{X}$ et $\bar{Y}$?
   (b) En déduire la loi de $\bar{X} - \bar{Y}$. Montrer que sous l'hypothèse H_0 on a
   $$\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \sigma^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\right).$$
   (c) Soit T la variable aléatoire définie par
   $$T = \frac{\bar{X} - \bar{Y}}{\sigma^2\left(\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}\right)}.$$

[1] Pour information, la mesure de performance utilisée est la *F-mesure* ; plus la valeur de la F-mesure est élevée plus la qualité du classifieur appris est grande. La validation croisée permet d'estimer la performance réelle en généralisation de chacun des modèles (en fait, elle permet d'obtenir un intervalle de confiance sur la performance en généralisation).

Quelle est la loi de T sous l'hypothèse H_0 ? Justifier.

(d) Il est possible d'exploiter la loi suivie par T pour construire un test statistique unilatéral appelé « t test » de niveau α permettant de décider la conservation ou le rejet de H_0 lorsque σ^2 est inconnue (lorsque σ^2 est connue, le test à effectuer est le test de comparaison de la moyenne d'une loi normale à une valeur de référence). Donner et justifier la règle de décision associée à ce test (cette règle de décision fait intervenir, entre autres quantités, un quantile de la loi de Student à déterminer et la variance empirique S^2 des deux échantillons définie par $S^2 = \frac{(n_x-1)S_x^2-(n_y-1)S_y^2}{n_x+n_y-2}$).

(e) Calculer la variance empirique s^2 (réalisation de S^2) des deux échantillons pour les valeurs du tableau 1.

(f) Calculer pour les données du tableau 1 la valeur t réalisation de T. En déduire un encadrement de la valeur de la probabilité critique pour la statistique de test T. Peut-on conclure au risque $\alpha = 0.05$ que RN donne de meilleures performances que AD ?

3. Une autre version du « t test » est réalisable pour le problème traité dans cet exercice : il s'agit du « paired t-test » ou *méthode des couples pour variables appariées*. Pour cela, on remarque que les variables X_i et Y_i sont liées par le fait qu'elle correspondent aux performances de RN et AD sur la même base B_i et X_i et Y_i sont dites *appariées*. Dans ces conditions, il est commun de faire l'hypothèse que $D_i = X_i - Y_i \sim \mathcal{N}(\mu_x - \mu_y, \sigma^2)$ pour tout i, les D_i étant i.i.d. Par ailleurs, puisque les variables sont appariées, on a nécessairement $n_x = n_y = n$.

   (a) Quelle est la loi de $\bar{D}$?

   (b) Définir et justifier une règle de décision au niveau α permettant de rejeter ou conserver H_0. Cette règle de décision doit faire intervenir $\bar{D}$, S_D, n et un quantile de la loi de Student à déterminer.

   (c) Appliquer cette règle de décision (ou bien, calculer la probabilité critique) pour les performances du tableau 1 afin de donner une conclusion au niveau $\alpha = 0.05$ sur l'acceptation ou le rejet de H_0. Commenter.

4. Pourquoi la puissance d'un test est-elle un bon critère pour comparer deux tests ?

5. Dans cette question, on suppose que σ^2 est connue et que, par ailleurs, $\mu_x = \mu_y + \delta$ avec $\delta > 0$. Cela signifie que l'hypothèse H_0 est fausse (et donc que l'utilisation de RN est préférable à celle de AD pour la tâche de classification de textes envisagée). On suppose également que $n_x = n_y = n$.

   (a) Donner une expression de la puissance du « t test » au niveau α. Cette expression fait intervenir la fonction de répartition F et un quantile de la loi normale centrée réduite, δ, n et σ.

   (b) Donner une expression de la puissance du « paired t-test » au niveau α. Cette expression fait intervenir la fonction de répartition F et un quantile de la loi normale centrée réduite, δ, n et σ.

   (c) En comparant ces deux expressions, dire quel est le test qui doit être choisi lorsque ce choix est possible.

2 Arbre de décision (1h)

Soit l'ensemble d'instances suivant :

Id	classe	att_1	att_2	att_3	att_4	att_5
1	1	A	E	H	K	M
2	1	B	E	I	L	M
3	1	A	G	I	L	N
4	1	B	G	H	K	M
5	1	A	G	I	L	M
6	2	B	F	I	L	M
7	2	B	F	J	L	N
8	2	B	E	I	L	N
9	2	C	G	J	K	N
10	2	C	G	J	L	M
11	2	D	G	J	K	M
12	2	B	F	I	L	M
13	3	D	E	H	K	N
14	3	A	E	H	K	N
15	3	D	E	H	L	N
16	3	D	F	J	L	N
17	3	A	F	H	K	N
18	3	D	E	J	L	M
19	3	C	F	J	L	M
20	3	D	F	H	L	M

où l'attribut Id est simplement un identifiant de l'instance concernée.

1. Quelle est la particularité des instances 6 et 12 ? Comment gérer ce genre de situation lors de l'apprentissage ? Justifier (on peut par exemple imaginer la situation où un grand nombre d'instances est concerné par la particularité des instances 6 et 12).
2. En utilisant le critère d'entropie, et en considérant que les pré-traitements nécessaires induits par la question précédente ont été effectués, construire 3 arbres de décision T_1, T_2 et T_3. Le premier (T_1) permettra de faire la classification entre des instances de la classe 1 et de la classe 2, le second (T_2) entre des instances des classes 2 et 3 et le troisième (T_3) entre les instances des classes 1 et 3. Justifier les calculs.
3. Comment utiliser ces trois arbres de décision conjointement pour classifier une instance dans l'une des classes 1, 2 ou 3 ? Est-il possible d'attribuer un poids particulier à chacun des arbres appris, et si oui, comment ? En particulier quelle est la classe de l'instance suivante : $[B, G, J, K, N]$?
4. Souvent, dans le cas de données réelles, il arrive que des attributs soient manquants. Dans le cas de cet exercice, cela signifie qu'au lieu d'être décrit par 5 attributs, un exemple peut n'être décrit que par 1, 2, 3 ou 4 attributs. Répondre aux questions suivantes :
   (a) quelle classe attribuer à l'exemple $[A, \emptyset, \emptyset, \emptyset, \emptyset]$?
   (b) quelle classe attribuer à l'exemple $[B, \emptyset, \emptyset, K, \emptyset]$? Justifier.
   (c) quelle classe attribuer à l'exemple $[B, E, \emptyset, \emptyset, \emptyset]$?
5. Proposer une instance étiquetée qui, si introduite dans l'ensemble d'apprentissage, empêche l'arbre T_2 de n'avoir que des feuilles pures. Expliquer.
6. Considérons uniquement l'arbre T_1. Pour mesurer l'erreur de généralisation de ce classifieur, une stratégie est d'utiliser m nouveaux exemples de la distribution qui a permis de générer les exemples d'apprentissage et de mesurer le nombre d'erreurs que T_1 fait sur ces m exemples. Plus m est grand plus la proportion d'erreurs mesurées est proche de l'erreur de généralisaton, c'est-à-dire la probabilité p de se tromper sur un exemple nouveau. Si on considère que le fait que T_1 fasse une erreur sur un exemple tiré au hasard est la réalisation d'une variable de Bernoulli de paramètre p, combien faut-il d'exemples m pour estimer avec une précision ϵ et une confiance $1 - \alpha$ l'erreur de généralisation de T_1, c'est-à-dire p. Ce nombre d'exemples s'écrira en fonction de p, ϵ et α.

3 Perceptron modifié (1h)

Dans cet exercice, les exemples appartiennent à l'ensemble $\mathbb{R}^n$, où n est un entier fixé et on suppose qu'ils sont tous de **norme 1**. Soit l'algorithme d'apprentissage suivant, prenant un paramètre $\sigma > 0$ et un ensemble d'apprentissage $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$ (avec $y_i \in \{-1, +1\}$) en entrée :

`Perceptron modifié :`

1. Choisir aléatoirement (de façon uniforme) un vecteur unitaire $\mathbf{w}$ de $\mathbb{R}^n$
2. **Si** tous les points mal classés (i.e. ceux qui vérifient $y_i \mathbf{w} \cdot \mathbf{x}_i < 0$) de $\mathcal{S}$ vérifient $|w \cdot \mathbf{x}| \leq \sigma \|w\|$ **alors**
   – arrêter
3. **sinon**
   – choisir le point $\mathbf{x} \in \mathcal{S}$ qui maximise la quantité $\frac{\mathbf{w} \cdot \mathbf{x}}{\|\mathbf{w}\|}$
   – mettre $\mathbf{w}$ à jour selon :
   $$\mathbf{w} \leftarrow \mathbf{w} - (\mathbf{w} \cdot \mathbf{x})\mathbf{x}$$
4. **Si** moins de $(1/\sigma^2) \ln n$ mises à jour de $\mathbf{w}$ ont été faites, **alors**
   – retourner au point 2
5. **sinon**
   – retourner au point 1

Nous allons étudier le temps de convergence de cet algorithme. Pour cela, on suppose qu'il existe un hyperplan cible défini par le vecteur unitaire $\mathbf{w}^*$ permettant la séparation de $\mathcal{S}$.

1. Supposons que le vecteur $\mathbf{w}$ unitaire choisi aléatoirement au point 1 de l'algorithme vérifie $\mathbf{w}^* \cdot \mathbf{w} \geq 1/\sqrt{n}$:
   (a) Montrer qu'à chaque mise à jour $\mathbf{w}'$ de $\mathbf{w}$ (cf. point 3) on a :
   $$\mathbf{w}' \cdot \mathbf{w}^* \geq \mathbf{w} \cdot \mathbf{w}^*$$
   (b) Montrer que, de plus,
   $$\|\mathbf{w}'\|^2 \leq \|\mathbf{w}\|^2 (1 - \sigma^2)$$
   (c) En déduire qu'au bout de t mises à jour, le vecteur $\mathbf{w}$ courant vérifie :
   $$\|\mathbf{w}\| \leq (1 - \sigma^2)^{t/2}$$
   (d) En déduire que le nombre d'itérations t vérifie nécessairement :
   $$t \leq \frac{1}{\sigma^2} \ln n$$

   Il vient d'être montré que si jamais le $\mathbf{w}$ inital vérifie bien $\mathbf{w}^* \cdot \mathbf{w} \geq 1/\sqrt{n}$ alors l'algorithme produit un vecteur qui vérifie la condition 2 en un nombre de mises à jour inférieur à $\frac{1}{\sigma^2} \ln n$.

   Note : tous les calculs mis en jeu dans cette question son très proches de ceux vus en cours pour la convergence du perceptron usuel.
2. La probabilité pour qu'un vecteur $\mathbf{w}$ choisi uniformément dans la boule unité vérifie $\mathbf{w}^* \cdot \mathbf{w} \geq 1/\sqrt{n}$ est supérieur à $\frac{1}{8}$.
   – Quelle est la probabilité pour que, en faisant deux tirages aléatoires de vecteurs unitaires, aucun d'eux ne vérifie $\mathbf{w}^* \cdot \mathbf{w} \geq 1/\sqrt{n}$?
   – Quelle est la probabilité pour que, en faisant m tirages de vecteurs unitaires, aucun d'eux ne vérifie cette même condition ?
   – Le paramètre δ étant donné, déduire de la question précédente le nombre de tirages de vecteurs unitaires à faire pour être sûr avec une probabilité $1 - \delta$, que l'un d'eux au moins vérifie la condition voulue.
3. Déduire des questions précédentes, en fonction de δ, σ et n que le temps d'exécution de l'algorithme du perceptron modifié pour produire avec une probabilité $1 - \delta$ un vecteur satisfaisant la condition d'arrêt 2 est de l'ordre de $\frac{1}{\sigma^2} \ln n \ln \frac{1}{\delta}$.