Université Paris-Saclay-Faculté de médecine
Master : sciences des données de santé
Apprentissage

Session du 23 janvier 2023 de 14h à 15h30.

I. Compréhension de cours

A. Lequel des points suivants empêche les réseaux neuronaux d'être plus largement utilisés ?

- **(a)** Ils ne sont pas facilement interprétables.
- **(b)** Ils ne sont pas performants en général.
- **(c)** Ils ne sont pas performants pour les tâches basées sur l'image.
- **(d)** Ils ne sont pas performants pour les tâches basées sur les séquences ou les données textuelles.

B. Laquelle des affirmations suivantes est vraie concernant le modèle des k plus proches voisins ?

- **(a)** Il est toujours préférable de choisir un k plus petit.
- **(b)** Un modèle à k voisins les plus proches a exactement k paramètres entraînables.
- **(c)** La précision d'apprentissage d'un modèle des 3 plus proches voisins est généralement supérieure à celle d'un modèle de 1 plu proche voisin.
- **(d)** La frontière de décision d'un modèle de k plus proches voisins est linéaire.
- **(e)** Plus k augmente, plus le biais (au sens de la décomposition biais-variance) augmente.

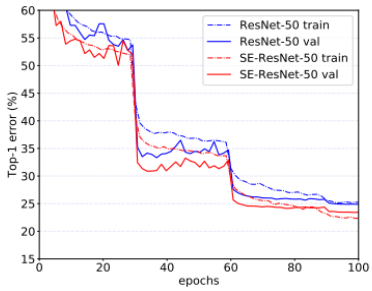
C. Laquelle des affirmations suivantes est vraie concernant la régression logistique ?

- **(a)** Si toutes observations d'apprentissage sont bien classées, alors l'erreur d'apprentissage est nulle.
- **(b)** Si toutes les observations d'apprentissage sont mal classées, alors l'erreur d'apprentissage est infinie
- **(c)** L'erreur d'apprentissage est toujours positive.
- **(d)** L'erreur d'apprentissage est toujours négative.
- **(e)** Aucune des réponses ci-dessus n'est vraie.

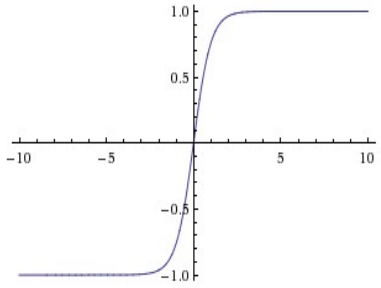
D. La descente de gradient et la rétropropagation sont utilisés pour mettre à jour les éléments suivants :

- **(a)** Les activations seulement.
- **(b)** Les couches et les activations.
- **(c)** Les couches seulement.
- **(d)** Les paramètres et les activations.
- **(e)** Les paramètres seulement.

E. Lequel des éléments suivants ne permet **pas** d'éviter le surapprentissage ?



(a) courbe d'apprentissage



(b) fonction d'activation

FIGURE 1 – 2 Figures des question **F** et **G**

(a) En diminuant le nombre d'unités cachées dans un perceptron multicouche.
(b) Utilisation d'un ensemble d'apprentissage plus grand.
(c) Utiliser une taille de batch size plus petite.
(d) Utiliser la descente de gradient stochastique avec momentum (une version particulière).
(e) Les deux réponses **(c)** et **(d)**.

F. La courbe d'apprentissage de la figure 1 indique le nombre d'époques sur l'axe des x, et le taux d'erreur (100% moins le taux de bon classement) sur l'axe des y. Quel est le changement de paramètre le plus probable dans les époques 25 et 60 qui aurait pu produire cette forme de la courbe d'apprentissage ?

(a) Le taux d'apprentissage a été diminué.
(b) Le taux d'apprentissage a été augmenté.
(c) Le batch size a été augmenté.
(d) Le batch size a été réduite.
(e) La décroissance progressive du taux d'apprentissage a été introduite.

G. Quelle est la fonction d'activation de la figure 1 ?

(a) ReLU
(b) tanh
(c) sigmoïde
(d) softmax
(e) Aucune des réponses précédentes.

II. Compromis biais-variance - comprendre le concept

A. Vous avez accès à une base de données européenne de 1000000 arbres de différents types. Chaque arbre (ligne de la base de données) est décrit par les variables suivantes
— Type d'arbre (bouleau, pin, tremble, etc.). Au total, 98 espèces.
— Âge
— Taille

— Circonférence (à 1 mètre de hauteur)
— Coordonnées géographiques de la position de l'arbre
— Type de végétation (forêt ouverte, forêt mixte, forêt de montagne, conifères humides, etc.)

Toutes les régions d'Europe sont bien représentées dans la base de données. Considérons un problème de régression où l'on veut modéliser l'âge d'un arbre à partir de sa hauteur et de sa circonférence. Nous utilisons un modèle de régression linéaire avec deux variables

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \varepsilon,$$

où les variables explicatives représentent la taille et la circonférence, et la variable réponse est l'âge.

(a) Quelle est la cause du biais du modèle ? Pensez-vous que le biais est élevé ou faible ?

(b) Quelle est la cause de la variance du modèle ? Pensez-vous que la variance est élevée ou faible ?

(c) Quelle est la cause de l'erreur irréductible du modèle ?

(d) Où voyez-vous le plus grand potentiel d'amélioration du modèle (réduction du biais, de la variance ou de l'erreur irréductible) ? et comment vous y prendriez-vous pour l'améliorer ?

B. Considérons maintenant un problème de classification où on modélise le type d'arbre comme variable réponse et les coordonnées géographiques comme variables explicatives. Nous utilisons un modèle à k-plus proches voisins avec $k = 1$.

(a) Quelles sont les causes du biais et de la variance du modèle ? Pensez-vous que le biais et la variance sont respectivement élevé ou faible ?

(b) Comment pourriez-vous échanger le biais contre une certaine variance (ou vice versa selon votre réponse à la question précédente) ?

III. Taux d'erreur

Supposons que nous prenions un ensemble de données, que nous le divisions en ensembles d'apprentissage et de test de tailles égales, puis que nous essayions deux procédures de classification différentes. Nous utilisons d'abord la régression logistique et obtenons un taux d'erreur de 20% sur les données d'apprentissage et de 30% sur les données de test. Ensuite, nous utilisons les k plus proches voisins (c'est-à-dire k-NN avec $k = 1$) et nous obtenons un taux d'erreur moyen (calculé sur les ensembles de données de test et d'apprentissage) de 18%. Sur la base de ces résultats, quelle méthode devrions-nous préférer utiliser pour la classification de nouvelles observations ? Pourquoi ?

IV. Gradient boosting : AdaBoost

Dans le boosting, nous étiquetons (par commodité) les classes comme suit $+1$ et -1 respectivement. On s'intéresse à une règle de classification de la forme $\widehat{y} = \text{sign}\big(C(x)\big)$, i.e., un seuillage d'une fonction réelle $C(x)$ en 0. Supposons qu'on a estimé la fonction $\widehat{C}(x)$ à partir d'un jeu de données d'apprentissage telle qu'on peut prédire $\widehat{Y} = \widehat{G}(x) = \text{sign}\Big(\widehat{C}(x)\Big)$.

$\widehat{C}(x_\star)$	$\widehat{y}_\star$	Exponential loss $\exp(-y_\star \widehat{C}(x_\star))$	Misclassification loss $I(y_\star \neq \widehat{y}_\star)$	$y_\star$
0.3				-1
-0.2				-1
1.5				1
-4.3				1

FIGURE 2 – Table à remplir à partir des informations données à la première et la dernière colonne.

A. Remplir les colonnes vides du tableau ci-dessous et expliquer dans quel sens la perte exponentielle est-elle plus informative que la perte par le taux de mauvais classement, et comment cette information peut-elle être utilisée lors de l'apprentissage d'une règle de classification ?

B. Déterminer les poids AdaBoost : Une règle de classification obtenue par AdaBoost peut être écrite comme $\widehat{y}_{\text{AdaBoost}}(x) = \text{sign}\big(C^B(x)\big)$ où les fonctions $C^1(x), \ldots, C^B(x)$ sont construites de manière séquentielle comme suit

$$C^b(x) = C^{b-1} + \alpha_b \widehat{y}^b(x),$$

initialisées par $C^0(x) \equiv 0$. La b ième règle de classification $\widehat{y}^b(x)$ est obtenue en appliquant une règle de classification faible sur la version pondérée du jeu de données d'apprentissage. Une fois que la règle $\widehat{y}^b(x)$ est ajustée, nous devons également calculer le coefficient de confiance (poids de la règle obtenue à la b ième itération) α_b. Cela se fait en minimisant la perte exponentielle pondérée calculée sur les données d'apprentissage,

$$\alpha_b = \underset{\alpha}{\text{argmin}} \left\{ \sum_{i=1}^{n} w_i^b \exp\Big(- \alpha y_i \widehat{y}^b(x_i) \Big) \right\}, \tag{1}$$

où $w_i^b = \exp\Big(- y_i C^{b-1}(x_i) \Big)$.

Montrer que la solution optimale est donnée par

$$\alpha_b = \frac{1}{2} \log\left(\frac{1 - \text{err}_b}{\text{err}_b} \right)$$

où $\text{err}_b = \sum_{i=1}^{n} \frac{w_i^b}{\sum_{j=1}^{n} w_j^b} \mathbf{1}\big\{ y_i \neq \widehat{y}^b(x_i) \big\}$.

Indication *Nous avons $y_i \in \{-1, 1\}$ et $\widehat{y}^b(x_i) \in \{-1, 1\}$. Utiliser ce fait pour décomposer la somme dans l'équation* (1) *en somme sur l'ensemble des points d'apprentissage correctement classés et une somme sur l'ensemble des points d'apprentissage mal classés.*