Apprentissage

Master 2 SDS, Université Paris-Saclay 16/01/2025

Généralités et compréhension de cours

Si vous pensez qu'une question est ambiguë, donnez la réponse qui vous semble meilleure. Justifier votre réponse quand c'est nécessaire.

1. Pour de très grands jeux de données d'apprentissage, lequel des modèles suivants aura généralement le temps d'apprentissage (entraînement) le plus court ?
   - **a.** Régression logistique
   - **b.** Réseau de neurones
   - **c.** Forêt aléatoire
   - **d.** Gradient boosting
   - **e.** CART
   - **f.** K-NN
2. Pour de très grands jeux de données d'apprentissage, lequel des algorithmes suivants produira un modèle nécessitant le plus petit nombre de paramètres ?
   - **a.** Réseau de neurones
   - **b.** Forêt aléatoire
   - **c.** Gradient boosting
   - **d.** CART
   - **e.** K-NN
3. (*Vrai ou Faux ?*) Soit un modèle d'arbre de décision estimé sur un jeu de données qui n'est ni centré ni réduit. Nous proposons de centrer et réduire le jeu de données test (toutes les variables explicatives ont une moyenne de zéro et un écart-type de 1). La forme d'un arbre est invariante par normalisation (centrage et réduction). Est-il de même pour la performance du modèle sur le jeu de données test ?
4. Une règle de classification K-NN donnera des précisions très proches de
   - **a.** Réseau de neurones
   - **b.** Forêt aléatoire
   - **c.** Gradient boosting
   - **d.** Un arbre profond

e. Un arbre peu profond

f. k-means (chaque classe est étiquetée avec la classe majoritaire des éléments qu'elle contient)

5. (*Vrai ou Faux ?*) Une perte quadratique est moins sensibles aux valeurs aberrantes qu'une perte absolue.

6. Pour un perceptron :

a. La précision du test sera, en général, plus élevée avec les variables explicatives d'origine.

b. La précision du test sera, en général, plus élevée en ajoutant comme variables explicatives, le double de chaque variable explicative d'origine.

c. La précision du test sera toujours la même avec l'un ou l'autre ensemble de variables explicatives.

7. Dans les réseaux de neurones, quel est l'avantage de la fonction d'activation ReLU par rapport à la fonction d'activation Sigmoïd ?

a. Les ReLU permettent au modèle d'apprendre des frontières de décision non-linéaires.

b. Les ReLU permettent de calculer plus rapidement la rétropropogation du gradient.

c. La fonction d'activation ReLUs peut être utilisée dans les couches de sortie alors qu'une fonction d'activation sigmoïdale ne peut pas l'être.

d. Toutes les réponses précédentes sont vraies.

8. Dans laquelle des architectures neuronales suivantes, certains poids sont-ils réutilisés plus d'une fois à chaque pass-forward ?

a. Une couche classique d'un réseau de neurones multi-couches.

b. Une couche CNN.

c. Un auto-encodeur

d. Toutes les réponses précédentes sont vraies.

9. Dans une itération t de l'algorithme Adaboost, on calcule le poids α_t de la règle faible h_t à l'aide de la formule $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ où ε_t est l'erreur de classification de la règle faible h_t sur le jeu de données d'apprentissage. Ici, nos règles faibles sont des arbres de décision de profondeur 1 qui produisent des frontières de décision verticales ou horizontales en demi-plan. Si nous effectuons deux itérations de Adaboost sur le jeu de données de la figure 1, est ce que

a. $\alpha_1 > \alpha_2$

b. $\alpha_1 < \alpha_2$

c. $\alpha_1 = \alpha_2$

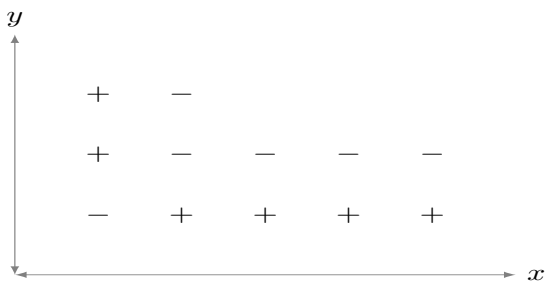

Figure 1: Jeux de données pour 2 itérations adaboost.

d. Informations insuffisantes.

10. Supposons que vous ayez un problème de classification pour lequel vous souhaitez pénaliser les erreurs de classification d'autant plus qu'elles sont éloignées de la frontière de classement. Quelle est la fonction de perte appropriée ?

a. Perte binaire

b. Perte exponentielle $e^{-yg(x)}$

c. Perte Hinge $\max\big(0, 1 - yg(x)\big)$

d. La perte logistique $\log(1 + e^{-yg(x)})$

e. Toutes les pertes précédentes

11. Supposons que vous soyez confronté à un problème de classification pour lequel vous ne souhaitez pas récompenser davantage les classifications correctes. Quel choix de fonction de perte serait approprié ?

a. Perte binaire

b. Perte exponentielle $e^{-yg(x)}$

c. Perte Hinge $\max\big(0, 1 - yg(x)\big)$

d. La perte logistique $\log(1 + e^{-yg(x)})$

e. Toutes les pertes précédentes

12. (*Vrai ou Faux ?*) Un perceptron est assuré d'apprendre une frontière de classement parfaite en un nombre fini d'itérations pour des données linéairement séparables.

13. (*Vrai ou Faux ?*) Lorsque l'on utilise le gradient boosting, l'utilisation d'arbres complexes (à grande profondeur) permet généralement d'améliorer la précision du modèle sur le jeu de données test.

14. (*Vrai ou Faux ?*) La méthode des moindres carrés revient à faire une descente de gradient pour minimiser - log-vraisemblance.

15. (*Vrai ou Faux ?*) Les poids d'un réseau de neurones sont mis à jour lors de la propagation vers l'avant.

16. (*Vrai ou Faux ?*) Pour effectuer une tâche de classification binaire, la sortie finale d'un réseau de neurones passe généralement par une activation ReLu avant d'être comparée à l'étiquette.

17. Vous trouverez ci-dessous les courbes de perte pour 3 petits réseaux de neurones identiques formés avec les algorithmes d'optimisation suivants : A. Descente du gradient **B.** Descente du gradient stochastique avec des lots de données de grande taille. **C.** Descente du gradient stochastique avec des lots de données de petite taille. Nous utilisons

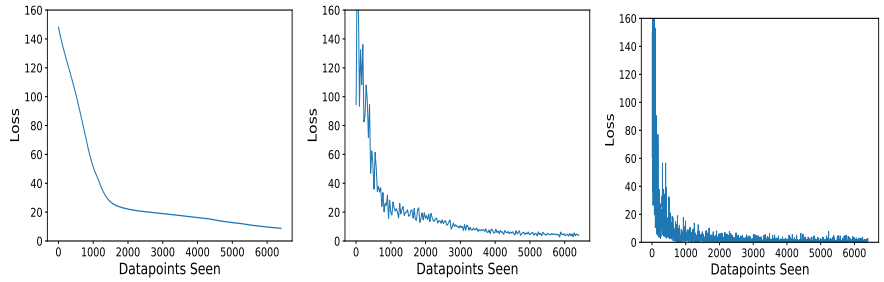

Figure 2: 3 courbes de pertes sur le même réseau.

le même taux d'apprentissage constant dans les trois cas. L'axe des abscisses correspond à la taille du lot multipliée par le nombre d'itérations (stochastiques) de descente de gradient. Associez les images (de gauche à droite) à la méthode d'optimisation utilisée.

Propositions : (A, B, C), (A, C, B), (B, A, C), (B, C, A), (C, A, B) ou (C, B, A)

18. La figure 3 présente un jeu de données d'apprentissage et la fonction de régression de trois réseaux de neurones différents. Le jeu de données se compose de 100 couples d'entrées-sorties scalaires. Les 3 réseaux comportent une couche cachée de 20 unités, mais diffèrent par le choix de la fonction d'activation : sigmoïd, ReLu et la fonction d'identité. Faites correspondre la fonction de régression estimée avec la fonction d'activation utilisée dans le réseau de neurones correspondant.

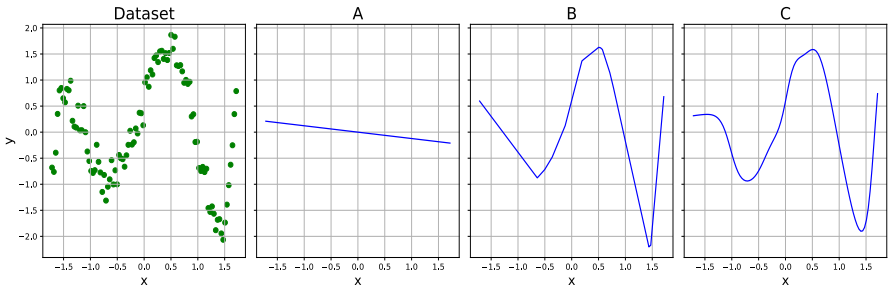

Figure 3: Les 3 réseaux de neurones.

19. Lequel de ces modèles donne une solution globalement optimale de minimisation de la fonction de perte ?

a. Gradient boosting

b. Réseau de neurones

c. Arbre de décision

d. Régression logistique

e. Random Forest

20. Le jeu de données d'apprentissage de la figure 4 se compose de points étiquetés de manière binaire. Vous souhaitez former un modèle de réseau neurones pour prédire les deux classes. Si vous n'utilisez qu'une seule couche cachée avec des unités d'activation ReLU, quel est le plus petit nombre d'unités d'activation requis pour séparer parfaitement les deux classes ? (justifier votre réponse)

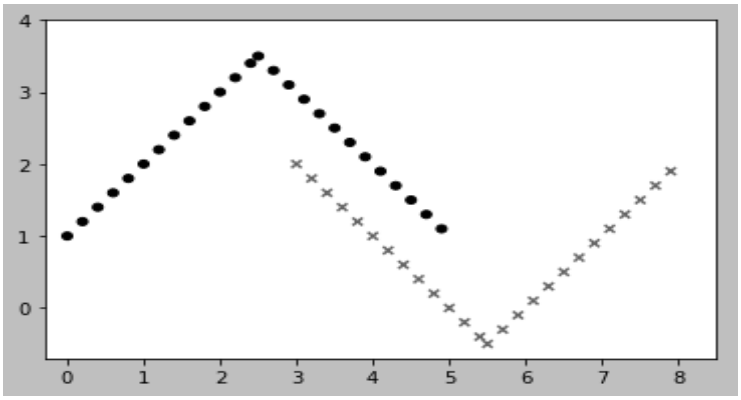

Figure 4: Classification binaire.

21. La plupart des méthodes de mesure de l'importance des variables (par exemple, comme nous l'avons vu pour les forêts aléatoires) sont conçues :

a. Pour évaluer approximativement l'importance de de l'effet de la variable explicative sur la variable réponse si cette variable varie dans le monde réel.

b. Pour estimer approximativement l'importance de l'effet de la variable explicative sur la variable réponse si la variable explicative était supprimée du modèle.

c. Les deux réponses ci-dessus sont identiques, elles sont donc toutes les deux correctes.

23. Vous souhaitez entraîner un réseau de neurones de grande taille (100 couches) à une tâche de classification binaire, en utilisant une activation sigmoïd dans la dernière couche et un mélange d'activations tanh et ReLU pour toutes les autres couches. Vous remarquez que les poids d'un sous-ensemble de vos couches cessent d'être mis à jour après le premier cycle d'apprentissage, alors que votre réseau n'a pas encore convergé. Une analyse approfondie révèle que les gradients de ces couches sont complètement, ou presque

complètement, nuls dès le début de l'apprentissage. Parmi les solutions suivantes, laquelle pourrait vous aider ? (Vous remarquez également que votre perte reste dans un ordre de grandeur raisonnable).

a. Augmenter la taille du jeu de données d'apprentissage

b. Mettre des fonctions d'activation leaky ReLU à la place des ReLU

c. Augmenter le pas de la descente du gradient

24. Vous évaluez les temps de calcul des couches couramment utilisées dans les CNN. Parmi les couches suivantes, laquelle devrait être la plus rapide (en termes d'opérations en virgule flottante) ?

a. Couche de convolution

b. Max pooling

c. Average pooling

d. Batch Normalization

25 Vous mettez en place un modèle d'apprentissage profond pour diagnostiquer le cancer du poumon à partir d'images radiographiques. Selon vous, quelle serait la mesure d'évaluation la plus appropriée et pourquoi ? Accuracy, Precision, Recall ou F1-score.

– Precision = TP/TP+FP

– Recall = TP/TP+FN

– F1-score = 2*(Recall * Precision) / (Recall + Precision)