

Fiche d'auto-évaluation 05

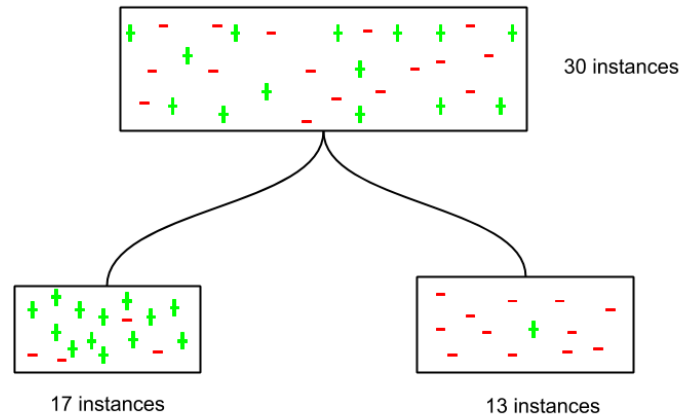
January 27, 2022

Caractéristique = Feature

1. Considérez l'entraînement d'un réseau de neurones. Dessinez un graphe hypothétique qui représente les courbes de la fonction de perte sur les données d'entraînement et de validation en fonction des époques d'entraînement, dans un cas où l'entraînement termine trop tôt et il aurait été mieux de continuer avec plus d'époques.
2. Dessinez un graphe comme le précédent, en assumant cette fois que le modèle obtenu souffre de surapprentissage.
3. Dessinez encore une fois un graphe comme le précédent, en assumant cette fois que le modèle obtenu souffre de sous-apprentissage.
4. Y a-t-il des cas, dans le contexte de l'apprentissage automatique, où on a la garantie absolue que la descente de gradient converge?
5. Jugez si la façon suivante de mettre à l'échelle (Min-Max scaling dans ce cas) est correct ou pas? Si elle n'est pas correcte, corrigez-la.
 - (a) On divise l'ensemble de données \mathcal{D} en données d'entraînement $\mathcal{D}^{\text{train}}$ et de test $\mathcal{D}^{\text{test}}$.
 - (b) On met les données de test à côté.
 - (c) On calcule la valeur minimale \min_j^{train} , maximale \max_j^{train} et moyenne μ_j^{train} de chaque colonne j sur les données d'entraînement.
 - (d) On transforme chaque valeur originale $x_j^{(i)}$ en $\frac{x_j^{(i)} - \dots}{\dots - \dots}$ (Completez la formule).
 - (e) On entraîne notre modèle sur l'ensemble $\mathcal{D}^{\text{train}}$ transformé.
 - (f) On prend les données de test $\mathcal{D}^{\text{test}}$, on calcule la valeur minimale \min_j^{test} , maximale \max_j^{test} et moyenne μ_j^{test} de chaque colonne j sur les données de test $\mathcal{D}^{\text{test}}$.
 - (g) On transforme $\mathcal{D}^{\text{test}}$ avec une formule similaire à avant, en utilisant cette fois $\min_j^{\text{test}}, \max_j^{\text{test}}, \mu_j^{\text{test}}$

- (h) On utilise le modèle entraîné pour faire les prédictions sur $\mathcal{D}^{\text{test}}$ transformé.
6. Si on fait du 7-fold cross-validation, combien de modèles il faut entraîner?
 7. Parlons de Sélection de Modèle (Model Selection): si on a un modèle et on veut trouver la meilleure combinaison de hyper-paramètres, comment peut-on utiliser la recherche par quadrillage (grid search) et cross-validation ensemble? Expliquez la procédure, pas à pas. (c'est la procédure implémentée par `sklearn.model_selection.GridSearchCV`)
 8. Dans une forêt d'arbres décisionnels, y a-t-il des cas où en augmentant le nombre d'arbre empire la justesse de la forêt? Si oui, quels cas?
 9. Quelle est la différence entre une forêt aléatoire, bagging trees et extra-trees?
 10. Si on a un ensemble d'arbres décisionnels, est-il préférable qu'ils se ressemblent ou qu'il soient différents?
 11. Explique l'algorithme CART pour entraîner un arbre décisionnel, pas à pas. Après, explique comment on intègre cela dans une forêt aléatoire, en écrivant un pseudo-code.
 12. Si on augmente le nombre d'arbres dans une forêt aléatoire, la variance de notre modèle diminue ou augmente?
 13. Un arbre décisionnel, est-il un classificateur linéaire ou non? Et une forêt aléatoire?
 14. Dans quel sens une prédiction d'une forêt aléatoire est interprétable? (Répond avec au moins deux éléments)
 15. Comment peut-on calculer l'importance d'une colonne en se basant sur un ensemble d'arbres?
 16. Est il nécessaire ou utile de mettre notre ensemble de données à l'échelle avant d'utiliser un arbre décisionnel ou un ensemble d'arbres?
 17. Que veut dire "classificateur linéaire"?
 18. Écrivez la formule de l'index d'impureté de Gini et de l'entropie dans le cas où on a 3 classes.
 19. Écrivez la formule du gain d'information.

20. Considérez la coupe (split) suivante¹



Calculez le gain d'information de la coupe, en considérant l'entropie et aussi l'index d'impureté de Gini.

21. Le gain d'information, peut-il être négatif? Si oui, dans quel cas?
22. Dans quelle intervalle de valeurs peut se trouver l'index de Gini, quand on a $K=2$ classes? Et pour $K=3$? Et pour K générique?
23. Répondez à la question précédente, en considérant cette fois-ci l'entropie.
24. Est-il convenable d'avoir un arbre de taille complète (fully grown)? Pourquoi?
25. Pourquoi le pruning est important? Est-ce qu'on le fait pendant l'entraînement ou le test? Quelles sont les critères qu'on peut adopter pour réduire (to prune) un arbre de décision?
26. Comment introduit-on l'aléatoire dans Bagging, forêts d'arbres décisionnels et Extra-trees?
27. Le concept de bagging est-il limité à l'apprentissage basé sur les arbres (tree-based learning) ou peut-on l'appliquer à d'autres types de modèles?
28. Supposer d'avoir un modèle de classification constitué par un ensemble d'arbres de décision. Si on augmente la diversité entre les arbres, la variance du modèle augmente ou diminue? Et si on augmente le nombre d'arbres?
29. Expliquez comment on entraîne une forêt d'arbres de décision pour faire de la régression.
30. Écrivez la formule de l'importance d'une caractéristique (feature importance).

¹Picture from <https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/>

31. Supposez de faire de l'entraînement d'une forêt d'arbres de décision, en essayant plusieurs configurations, en ajoutant de plus en plus des arbres. Y a-t-il un point auquel c'est mieux de s'arrêter et ne plus ajouter d'arbres? Quand?