Fiche d'auto-évaluation 05

May 6, 2021

Caractéristique = Feature

1. 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 a côté.

   (c) On calcule la valeur minimale $\min_j^{\text{train}}$, maximale $\max_j^{\text{train}}$ 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 $\min_j^{\text{test}}$, maximale $\max_j^{\text{test}}$ 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.

2. Si on fait du 7-fold cross-validation, combien de modèles il faut entraîner?

3. 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)

4. Dans une forêt aléatoire, y-a-t-il des cas ou augmenter le nombre d'arbre empire la justesse de la forêt? Si oui, quels cas?

5. Quelle est la différence entre une forêt aléatoire, bagging trees et extra-trees?

6. Si on a un ensemble d'arbres décisionnels, est-il préférable qu'ils se ressemblent ou qu'il soient différents?

7. 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.

8. Si on augmente le nombre d'arbres dans une forêt aléatoire, la variance de notre modèle diminue ou augmente?

9. Un arbre décisionnel, est-il un classificateur linéaire ou non? Et une forêt aléatoire?

10. Dans quel sens une prédiction d'une forêt aléatoire est interprétable? (Répond avec au moins deux éléments)

11. Comment peut-on calculer l'importance d'une colonne en se basant sur un ensemble d'arbres?

12. 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?