Machine Learning Project

Random Forests & Gaussian Mixture Model

Préparé par:

Manal NHILI

Professeur:

Abdelhak MAHMOUDI

Année universitaire: 2021-2022

Plan

- Random Forests
- Caractéristiques principales du RF
- Algorithme du Random Forest
- Principaux défis du RF
- Gaussian Mixture Model
- Algorithme du GMM
- Expectation Maximization
- Références

Random Forests (RF)

- **Random Forest** est l'un des algorithmes d'apprentissage automatique les plus polyvalents et les plus populaires. Il peut être utilisé pour les problèmes de classification et de régression en ML.
- Un algorithme de forêt aléatoire se compose de plusieurs arbres de décision. La "forêt" générée par l'algorithme de la forêt aléatoire est entraînée par "**bagging"** ou **"bootstrap aggregating".**
- Le bagging est un méta-algorithme d'ensemble qui améliore la précision des algorithmes d'apprentissage automatique.
- La forêt aléatoire est un moyen populaire d'utiliser les algorithmes d'arbres pour obtenir une bonne précision et surmonter le problème de surajustement rencontré dans l'algorithme DT (arbre de décision) unique.

Caractéristiques principales du RF:

- Il utilise une méthode appelée **bagging** pour créer différents sous-ensembles des données d'entraînement originales.
- Il sectionne **aléatoirement** différents sous-ensembles de caractéristiques/attributs et construit l'arbre de décision sur la base de ce sous-ensemble.
- En créant de nombreux arbres de décision différents, basés sur différents sous-ensembles de données d'apprentissage et différents sous-ensembles de caractéristiques, on augmente la probabilité de capturer toutes les façons possibles de modéliser les données.

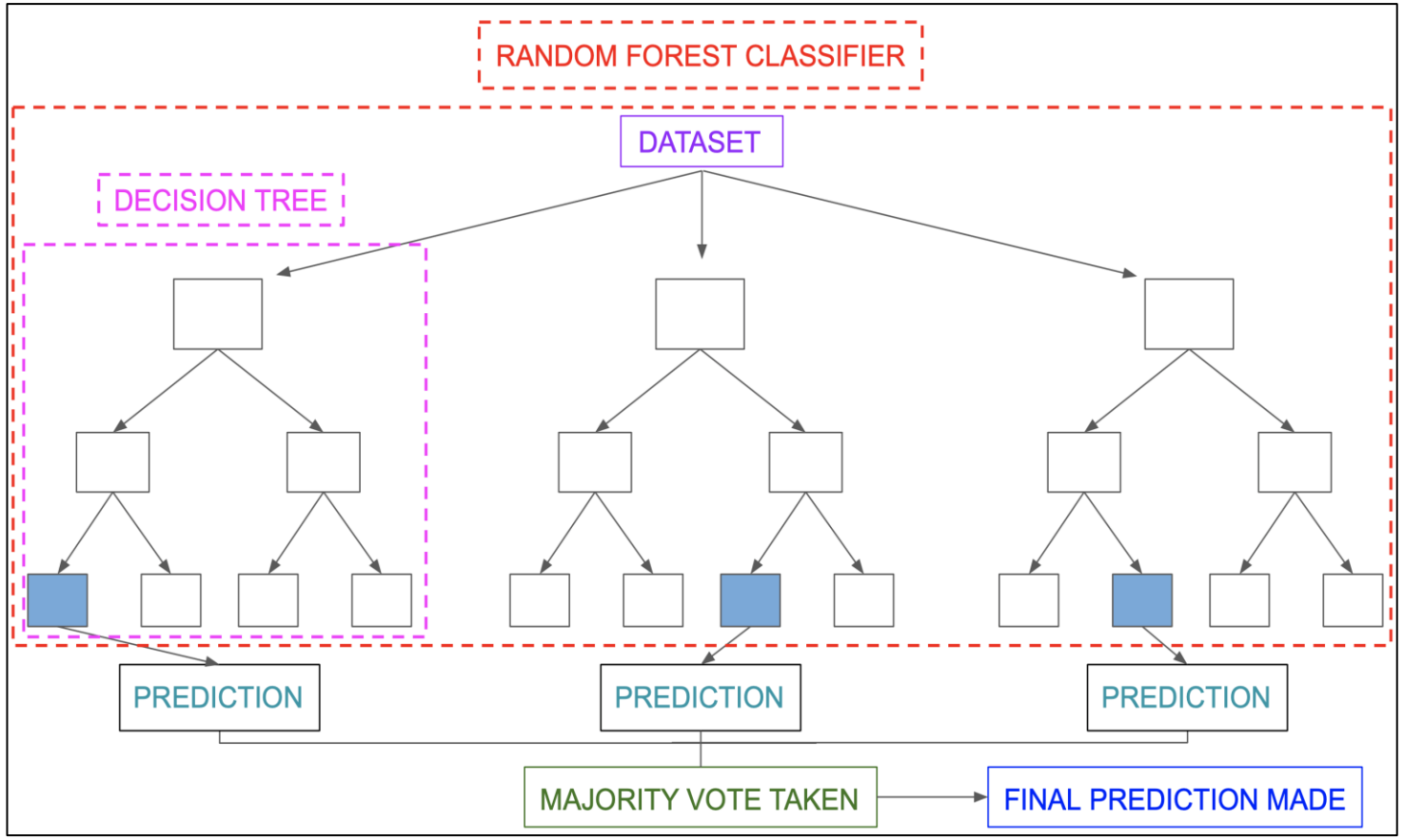

Figure 1 : Algorithme de Random Forest

Comme on peut le voir dans la figure, il y a plusieurs arbres de décision, et chacun d'eux travaille indépendamment pour former sa propre sortie et donner une prédiction.

Pour chaque arbre de décision produit, l'algorithme utilisera une mesure, telle que l'indice de Gini, pour sélectionner les attributs à répartir à chaque nœud de l'arbre de décision.

La forêt aléatoire prend ensuite la prédiction de chaque arbre et sélectionne la majorité de la classe prédite par chaque arbre comme la vraie classe prédite de l'ensemble de données.

Principaux défis

- **Processus fastidieux :** Comme les algorithmes de forêt aléatoire peuvent traiter de grands ensembles de données, ils peuvent fournir des prédictions plus précises, mais ils peuvent être lents à traiter les données car ils calculent les données pour chaque arbre de décision individuel.

- **Nécessite plus de ressources :** Puisque les forêts aléatoires traitent de plus grands ensembles de données, elles auront besoin de plus de ressources pour stocker ces données.

- **Plus complexe :** la prédiction d'un seul arbre de décision est plus facile à interpréter que celle d'une forêt d'arbres.

Gaussian Mixture Model (aka. GMM)

- Les modèles de **mélange gaussien sont un modèle probabiliste** permettant de représenter des sous-populations normalement distribuées au sein d'une population globale.

One-dimensional Model :

$$p(x) = \sum_{i=1}^{K} \phi_i \mathcal{N}(x \mid \mu_i, \sigma_i)$$

$$\mathcal{N}(x \mid \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

$$\sum_{i=1}^{K} \phi_i = 1$$

Multi-dimensional Model:

$$p(\vec{x}) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i)$$

$$\mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_i|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^{\mathrm{T}} {\Sigma_i}^{-1} (\vec{x}-\vec{\mu}_i)\right)$$

$$\sum_{i=1}^{K} \phi_i = 1$$

Algorithme du GMM

- La première étape va être de spécifier à l'algorithme le nombre de clusters à trouver.
- L'algorithme initialise ensuite de façon aléatoire pour chaque cluster à trouver une moyenne et un écart type qui sont les principaux paramètres de la distribution gaussienne.
- Ensuite, pour chaque observation l'algorithme va calculer sa probabilité d'appartenance à une distribution gaussienne (cluster).
- L'algorithme s'arrête une fois que toutes les observations ont été positionnées dans un cluster.

Expectation Maximization (aka. EM)

- L'algorithme d'espérance-maximisation (EM) est une série d'étapes permettant de trouver de bonnes estimations des paramètres lorsqu'il existe des variables latentes.
- **Étapes de l'EM :**
  - **Etape E** : consiste à calculer l'espérance des affectations des composantes pour chaque point de données compte tenu des paramètres du modèle.
  - **Etape M** : consiste à maximiser les attentes calculées dans l'étape E par rapport aux paramètres du modèle. Cette étape consiste à mettre à jour les valeurs des paramètres du modèle.
  - Une fois que l'algorithme EM a été exécuté jusqu'au bout, le modèle ajusté peut être utilisé pour effectuer diverses formes d'inférence. Les deux formes d'inférence les plus courantes effectuées sur les GMM sont l'estimation de la densité et le clustering.

Merci

Références

- 1. Education, I. C. (2021, January 26). Random Forest. IBM. Retrieved from https://www.ibm.com/cloud/learn/random-forest
- 2. RandomForest Machine Learning – Oracle Machine Learning (OML). (2020, June 24). Oralytics. Retrieved from https://oralytics.com/2020/06/24/randomforest-machine-learning-oracle-machine-learning-oml/
- 3. Ampadu, H. (2021, May 10). Random Forests Understanding. Ai-Pool. https://ai-pool.com/a/s/random-forests-understanding
- 4. Gaussian Mixture Model | Brilliant Math & Science Wiki. (n.d.). Brilliant.Org. https://brilliant.org/wiki/gaussian-mixture-model/
- 5. Ghassany, M. (2021, October 18). 8 Gaussian Mixture Models & EM | Machine Learning. Mghassny.Com. https://www.mghassany.com/MLcourse/gaussian-mixture-models-em.html#the-gaussian-distribution
- 6. Team, D. (2021, March 8). Gaussian Mixture Model with Case Study – A Survival Guide for Beginners. DataFlair. https://data-flair.training/blogs/gaussian-mixture-model/
- 7. Gupta, P., & Sehgal, N. K. (2021). Introduction to Machine Learning in the Cloud with Python: Concepts and Practices (1st ed. 2021 ed.). Springer. https://doi.org/10.1007/978-3-030-71270-9
- 8. Vannieuwenh, A. (2019). Intelligence artificielle vulgarisée - Le Machine Learning et le Deep Learning par la pratique (French Edition) [E-book]. ENI.