Gaspard Palay / XgBoost

Gaspard Palay

29/01/2021

1. Critères d'évaluation

1. Comportement du Rmd lors de son exécution
2. Qualité de la rédaction du dossier
3. Didactisme et pertinence du dossier
4. Qualité et lisibilité du Rmarkdown
5. Qualité des explications du modèle mathématique

2. Lien vers le document commenté

En cliquant **ici**, vous trouverez le lien menant au GitHub de Chaymae GASMI hébergeant le fruit de sa réalisation.

3. Auteurs du document commenté

Le document évalué dans le cadre de ce rendu a été produit par Zakaria RIDADARAJAT, Chaymae GASMI, Hakim DAIF étudiants en MSc Data Management à Paris School of Business.

4. Synthèse du document

Le document est une production RMD et PDF présentant une définition du package XgBoost ainsi que :

- Son modèle et ses paramètres
- La fonction objectif
- L'optimizing Tree stucture
- L'addiditive training
- L'implémentation XgBoost
- Le developpement de Taylor
- le retour sur les objectifs
- Les paramètres Xgboost

Xg Boost est un package bibliothèque de renformcement de gradient distribuée. Il s'exécute dans le cadre de Gradient Boosting. C'est une méthode d'apprentissage d'ensemble. C'est un modèle d'apprentissage supervisé. L'auteur fait un état de ce qu'est l'apprentissage supervisé avant d'exposer l'algorithme XgBoost.

Il rappekl aussi ce qu'est une fonction d'objectif : composée d'une fonction de perte et d'un terme de régularisation : $\boldsymbol{Obj(\cdot) = L(\cdot) + K(\cdot)}$ ou Obj : fonction objectif composée par L fonction de perte et K terme de régularisation. la fonction de perte mesure la qualité de la prédicition du modèle sur les données. Elle est l'erreur quadratique moyenne : $L(\theta) = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$

la régularisation contrôle la compléxité d'un modèle et évite l'overfitting.

L'auteur expose ensuite qu'XgBoost est composé d'agrégation d'arbre c'est à dire un modèle composé d'arbres de régression ou de classification.

5. Extrait commenté des parties de code

1. L'optimizing Tree stucture

$$\min \leftarrow \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{b=1}^{B} \Omega\left(f_b\right)$$

$\Omega\left(f_b\right) = \gamma|T| + \frac{1}{2}\lambda \sum_{j=1}^{|T|} w_j^2$, et $|T| = d + 1$ le nombre de feuille de l'arbre de décision

w_j les poids de régression d'une feuille (région) R_j

2. L'addiditive training

Le boosting est une prédiction constante ou l'on ajoute à chaque fois une nouvelle fonction :

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1\left(x_i\right) = \hat{y}^{(0)} + f_1\left(x_i\right) \\
\hat{y}_i^{(2)} &= f_1\left(x_i\right) + f_2\left(x_i\right) = \hat{y}^{(1)} + f_2\left(x_i\right) \\
&\cdots \\
\hat{y}_i^{(t)} &= \sum_{b=1}^{t} f_b\left(x_i\right) = \hat{y}^{(t-1)} + f_t\left(x_i\right)
\end{aligned}$$

3. L'implémentation XgBoost

A la lecture du document, je comprends que : pour ajouter le boosting, il faut définir le fractionnement c'est à dire le rang auquel on va omplémenter XgBoost. On le définit grâce à la fonction coût :

$$\begin{aligned}
\mathrm{Obj}^{(t)} &= \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_t\left(x_i\right)\right)\right)^2 + \Omega\left(f_t\right) + C_1 \\
&= \sum_{i=1}^{n} \left[2\left(\hat{y}_i^{(t-1)} - y_i\right) f_t\left(x_i\right) + f_t\left(x_i\right)^2\right] + \Omega\left(f_t\right) + C_2
\end{aligned}$$

4. le retour sur les objectifs

Pour un retour su objectif, on doit alors supprimer les constantes et les regrouper par feuilles avec :

$$\begin{aligned}
\mathrm{Obj}^{(t)} &\approx \sum_{i=1}^{n} \left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right)\right] + \Omega\left(f_t\right) \\
&= \sum_{i=1}^{n} \left[g_i w_{J(x_i)} + \frac{1}{2} h_i w_{J(x_i)}^2\right] + \gamma|T| + \frac{1}{2}\lambda \sum_{j=1}^{|T|} w_j^2 \\
&= \sum_{j=1}^{|T|} \left[\left(\sum_{i:x_i \in R_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i:x_i \in R_j} h_i + \lambda\right) w_j^2\right] + \gamma|T|
\end{aligned}$$

Si la structure de l'arbre $\left(R_j\right)_{j=1}^{|T|}$ fixée,

le poids optimal de chaque feuille résultante sera :

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

La valeur objective résultante sera :

$$\text{Obj}^{(t)} \approx -\frac{1}{2}\sum_{j=1}^{|T|}\frac{G_j^2}{H_j + \lambda} + \gamma \mid T$$

5. Les paramètres Xgboost

Ici l'auteur définit les différents paramètres d'XgBoost, j'en résume certains ci dessous : - Maxdepth [par défaut = 6] Profondeur maximale d'un arbre. Augmenter cette valeur rendra le modèle plus complexe et plus susceptible de dépasser t. 0 indique non limite.

- GAMMA [par défaut = 0, alias: minsplitloss] Réduction minimale des pertes requise pour effectuer une nouvelle partition sur unnœud feuille de l'arbre. Plus le gamma est grand, plus il est conservateur l'algorithme sera.
- ETA [par défaut = 0,3, alias: taux d'apprentissage] Réduction de la taille des pas utilisée dans la mise à jour pour éviter le surajustement. Après à chaque étape de boost, nous pouvons directement obtenir le poids des nouvelles fonctionnalités,et eta réduit les poids des fonctionnalités pour rendre le processus de renforcement plus conservateur.

- Le sous-échantillonnage se produira une fois dans chaque itération d'amplification.

6. Evaluation du travail suivant les 5 critères précités

1. Comportement du Rmd à l'exécution

L'auteur du trvaail fournit aussi bien un RMD et un PDF. La qualité du RMD est excellente, il s'exécute très bien et les fonctions sont bien paramétrées.

2. Qualité de la rédaction du dossier

La rédaction de ce document est de très bonne qualité. Il est bien structuré, et les étapes de progression son respectées. L'auteur est très pragmatique. Les termes employés sont bons. Il manquerait peut être un peu de détails d'explications des termes pour les lecteurs les plus novices en modèle statistiques (comme moi…)

3. Didactisme et pertinence du dossier

La lecture de ce dossier est aisée et accessible. Il est didactique. Il ets pertinent puisqu'il fait partie de notre programme et ce modèle sera étudié au prochain semestre du MSc. Le choix du sujet est très bon. Il reprend toutes les étapes de l'algorithme de programmation XgBoost.

4. Qualité et lisibilité du RMD

Le fichier est lisible et accessible sur son GitHub. Le fichier est compilé en PDF à partir du RMD. Il est lisible et aéré. Les titres sont mis en forme.

5. Qualité des explications du modèle mathématique

Les applications sont simples. L'auteur utilise de multiples expréssions Latex complexes pour illustrer ses propos. J'aurai peut être aimé lire plus détails, de commentaires et de développement sur les expréssions mathématique, elles ne sont pas faciles à comprendre pour un lecteur non aguérri.

7. Conclusion

Selon moi, il s'agit globalement d'un très bon travail.

L'auteur synthétise et définit le modèle XgBoost. Il s'appuie sur plusieurs expréssions mathématique pour démontrer la chronologie de l'algorithme et son fonctionnement. La rédaction du dossier est claire et limpide. Très bon travail qui me resservira dans la poursuite de ma formation.

Vous retrouvez ce document sur mon **GitHub**.