Methodes d'ensemble et forets aleatoires

Yannig Goude

Bagging

Introduit dans Breiman (1996)

- deux ingredients clefs : bootstrap et aggregation
- l'agregation de methode de prevision initiales independantes (base learners) mene à une reduction importante de l'erreur de prevision.
- il faut donc oeuvrer dans le but d'obtenir des methodes initiales aussi independantes que possible.
- Idee naive : entrainer nos "base learners" (ex : CART) sur des sous-ensembles d'observations disjoints de l'ensemble d'entrainement.
- Probleme : le nombre d'observations de l'ensemble d'entrainement n'est pas infini ! les "base learners" auront trop peu de donnees et de mauvaises performances.

Intuition

Soit (X, Y) de loi P, un échantillon $\mathcal{L} = (x_i, y_i)_{1 \leq i \leq n}$ et un predicteur individuel $\widehat{y} = \phi(x, \mathcal{L})$.

Le prédicteur baggé associé est, en supposant qu'on effectue un grand nombre de tirage aléatoire:

$$\phi_a(x, P) = E_{\mathcal{L}}(\phi(x, \mathcal{L}))$$

Le risque quadratique associé à chaque prédicteur est:

$$E_{\mathcal{L}} E_{X,Y}(Y - \phi(x, \mathcal{L})^2$$

Le risque quadratique associé au predicteur baggé est

$$E_{X,Y}(Y - \phi_a(x, P))^2$$

Par l'inegalité de Jensen $[E(Z)]^2 \leq E(Z^2)$:

$$E_{X,Y}(Y - \phi_a(x, P))^2 \leq E_{\mathcal{L}} E_{X,Y}(Y - \phi(x, \mathcal{L})^2$$

Intuition

Le risque du prédicteur baggé est donc inférieur à celui des prédicteurs individuels. De combien dépend de l'inegalite:

$$E_{\mathcal{L}}(\phi(x,\mathcal{L})^2) - [E_{\mathcal{L}}(\phi(x,\mathcal{L}))]^2 \geq 0$$

d'autant plus vrai que les predicteurs individuels sont instables (forte variance en fonction de $\mathcal{L}$).

Bagging

Le bagging crée des sous-ensembles d'entrainement à l'aide d'echantillonnage bootstrap R. J. Tibshirani and Efron (1993).

Pour créer un nouveau "base learner":

- on tire aleatoirement avec remise n observations de l'ensemble d'entrainement.
- on entraine notre methode (ex : CART) sur cet ensemble d'observations
- chaque **base learner** contient un sous ensemble des observations de l'ensemble d'entrainement.
- la performance d'un "base learner" est obtenu par l'erreur **out-of-bag**.

Les forêts aléatoires

Méthode introduite dans Breiman (2001), succède et unifie des idées plus anciennes :Breiman (1996), arbres de décisions CART Breiman et al. (1984)

Preuves de convergences récentes, voir Biau and Scornet (2016) et Genuer and Poggi (2017) pour un survey récent, un site web utile : https://www.stat.berkeley.edu/users/breiman/

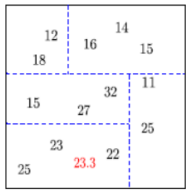
Comme expliqué dans Genuer and Poggi (2017): il n'y a pas de résultats théoriques solides disponibles pour les RF proposés par Breiman mais seulement pour des variantes simplifiées:

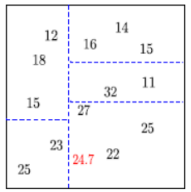
- Pure Forest: forêts purement aléatoires lorsque les partitions associées aux arbres sont choisies aléatoirement, indépendamment des observations, voir Arlot and Genuer (2014).
- Extra-Trees (Extremly Randomized Trees): tirage de m variable aléatoirement puis tirage des seuils de coupures aléatoirement sur le support de chaque variable. La coupure se fait sur le "meilleur" découpage parmi les m obtenus. voir Geurts, Ernst, and Wehenkel (2006).
- Random subspace: tirage d'un sous-ensemble de variables fixe pour tout l'arbre, Liu, Ting, and Fan (2005)

Les forêts aléatoires

Les forêts aléatoires consistent à faire tourner en parallèle un grand nombre (plusieurs centaines) d'arbres de décisions construits aléatoirement, avant de les moyenner.

En termes statistiques, si les arbres sont décorrélés, cela permet de reduire la variance des prévisions.





prévoit $\frac{24.7+23.3}{2} = 24$

Les forêts aléatoires: intuition

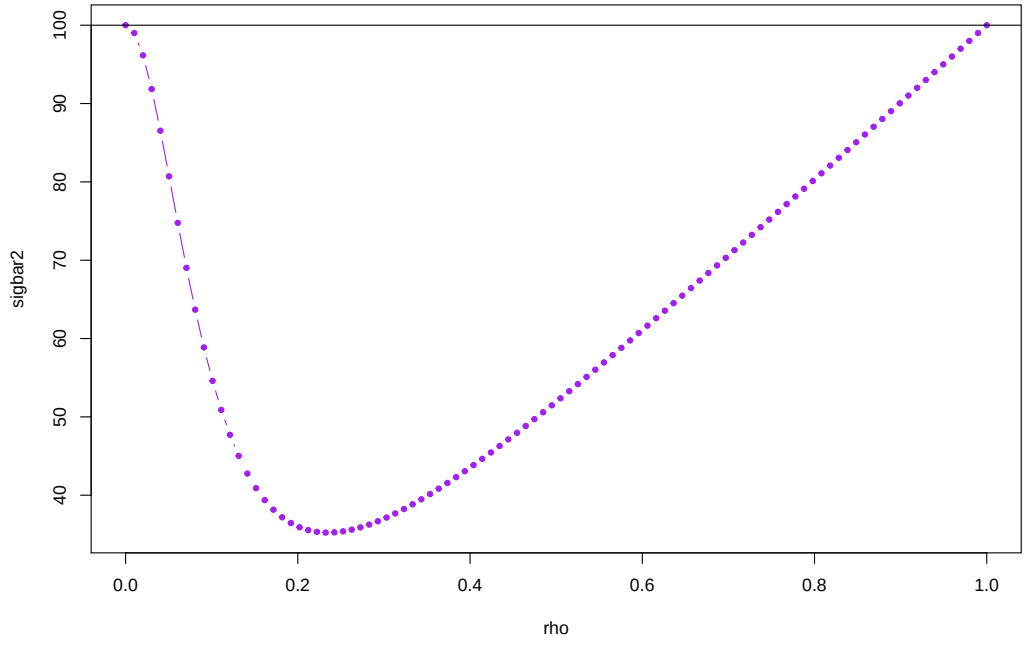
si K arbres Y_i sont identiquement distribués, de variance σ^2, avec un coefficient de corrélation deux à deux ρ la variance de leur moyenne $\frac{1}{K}\sum_{i=1}^{K} Y_i$) est alors:

$$\bar{\sigma} = \frac{1}{K^2}(K\sigma^2 + K(K-1)\rho\sigma^2)$$

$$\bar{\sigma} = \rho\sigma^2 + \frac{\sigma^2}{K}(1-\rho)$$

Les forêts aléatoires: intuition

```
sigma<-10
rho<-seq(0,1,length=100)
K<-1+100*rho^2
sigbar2<-rho*sigma^2+sigma^2*(1-rho)/K
plot(rho,sigbar2,type='b',pch=20,col='purple')
abline(h=sigma^2)
```



Construire des arbres peu corrélés

- Bootstrapping: Plutôt qu'utiliser toutes les données pour construire les arbres, on choisit aléatoirement pour chaque arbre un sous-ensemble (avec répétition possibles) des données.
- Choix aléatoire de la variable explicative à couper. Contrairement à CART pas d'élagage

Construire des arbres peu corrélés

q : paramêtre contrôlant l'aléatoire

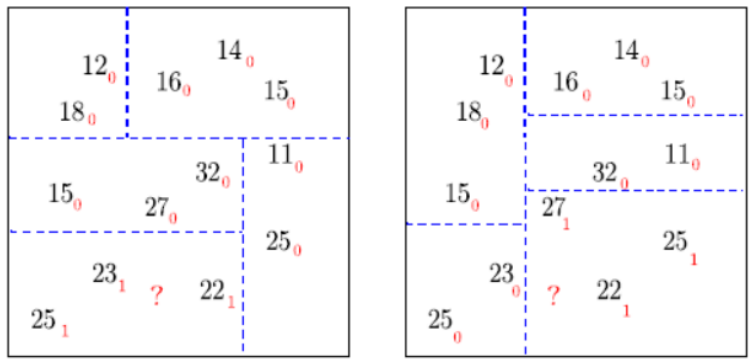
Pour couper un noeud :

- on choisit aléatoirement un sous-ensemble de q variables explicatives potentielles parmis les p disponibles
- on choisit la variable à couper et le seuil de coupe en minimisant un critère de variabilité (cf CART: Variance, Entropie, Gini) parmis ce sous-emsemble.
- si q = p : pas d'aleatoire. On retrouve le choix déterministe de CART
- si q = 1 : aléatoire total dans le choix de la variable (mais pas dans le seuil de la coupure).

En pratique, dans le package randomForest il est proposé $q = \sqrt{p}$ pour la classification et $q = p/3$ pour la régression.

Notion de proximite

Intuition : tomber souvent dans les mêmes feuilles des arbres signifie expliquer la sortie Y de façon similaire.



$$prox(X_t, X_s) = \frac{1}{K} \sum_{k=1}^{K} 1(X_t, X_s \in \text{meme feuille de l'arbre } k)$$

On predit ensuite Y_t par:

$$\frac{1}{C} \sum_{i=1}^{n} prox(X_t, X_i) Y_i$$

Erreur OOB

L'utilisation du bagging fait qu'on a laissé une partie des données de côté pour l'apprentissage de chaque arbre. Ces données non-utilisées sont appelées les données out of bag (oob).

Pour une observation i de l'échantillon d'apprentissage on calcul l'erreur obtenu par l'agrégation de tous les arbres n'ayant pas été appris avec cette observation. L'erreur oob est la somme des erreurs calculé ainsi sur l'ensemùble de l'échantillon d'apprentissage.

L'erreur oob est une approximation de l'erreur de généralisation de la forêt, elle n'utilise jamais les prédictions de la forêt elle-même, mais plutôt celles de prédicteurs qui sont des agrégations d'arbres de cette forêt.

Importance des variables

Les forêts aléatoires permettent de classer les variables explicatives par ordre d'importance (VI) dans la prévision.

Tout d'abord, on construit la forêt aléatoire, on calcule l'erreur E "out-of-bag" de la foret

Le score $VI(X^j)$ d'une variable explicative X^j est calculé comme suit:

- on permute aléatoirement les valeurs de la variable explicative parmi les observations de l'ensemble d'entrainement.
- on calcule à nouveau l'erreur out-of-bag et on fait la différence avec E.
- on renormalise les scores.

Avantages et inconvénients des random-forests

Avantages

- pas de sur-apprentissage
- en géneral : meilleure performance que les arbres de décision, calcul de l'erreur "Out-of-Bag" direct
- effet 2 en 1: validation croisée non necessaire grâce aux echantillons oob
- paramêtres faciles à calibrer
- parallélisation possible
- souvent utilisées comme benchmark dans les compétition de machine learning

Inconvénients

- boite noire : difficilement interpretable, difficilement améliorable
- entrainement plus lent

les random forest fonctionnent tout le temps bien mais excellent plus rarement

Sélection de variable

même si les forêts sont robustes à un nombre de variable important, en grande dimension ($n << p$) il est nécessaire d'effectuer de la sélection de variable.

- utilisation du score d'importance puis sélection forward, backward ou stepwise
- dans Genuer, Poggi, and Tuleau-Malot (2010): sélection automatique (notamment choix du nombre m de variables à conserver) des variables en 2 étapes:
  - 1. classement des variables par importance décroissante, détermination de m en exploitant la variance des VI pour plusieurs forêts (monte carlo)
  - 2. *interpretation*: pour $k = 1, ..., m$ construire les forêts incluant les k premières variables et choisir la forêt avec la plus faible erreure oob. *prediction*: à partir de l'ensemble de variable sélectionné pour l'interprétation, enrichir progressivement la forêt avec des variables faisant baisser significativement l'erreur oob.

Extensions

De nombreuses variantes existent associées aux différentes variantes des arbres CART, de plus on trouve des extensions liées à des problèmes statistiques variés, entre autres:

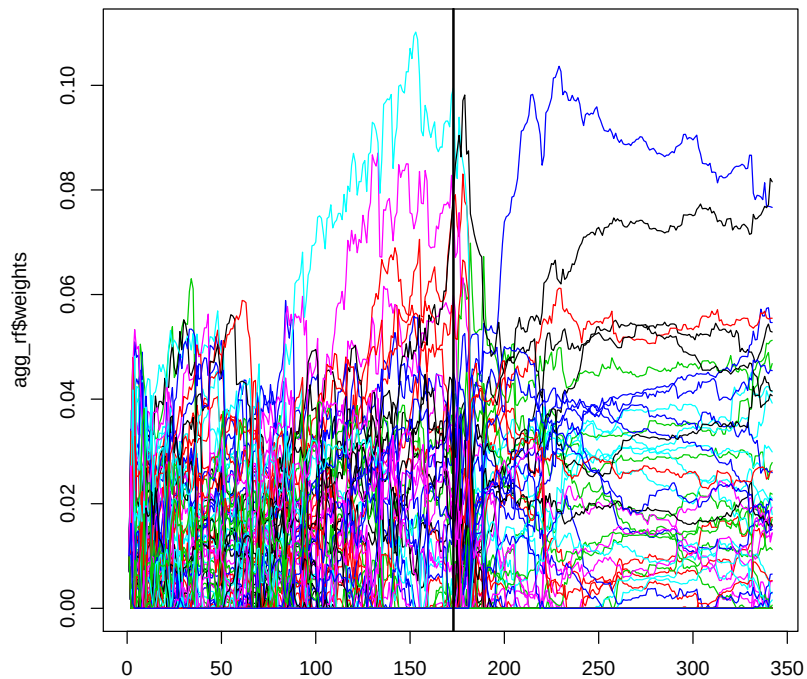
- ranking: Clémençon, Depecker, and Vayatis (2013)
- données de survie: Hothorn et al. (2006)
- régression quantile : Meinshausen (2006)
- modélisation de loi β: Weinhold et al. (2019)

des travaux concernent l'amélioration des forêts:

- élagage de forêt: sélection d'un sous-ensemble d'arbre "les plus divers" Fawagreh, Gaber, and Elyan (2015), gain d'interprétabilité et souvent de performance
- couplage forêts et deep learning: Biau, Scornet, and Welbl (2019)
- spatial forest: geographical random forest, Georganos et al. (2019)

Extensions

- forêt généraliséees: non-parametric quantile regression, conditional average partial effect estimation, heterogeneous treatment effect estimation: Athey et al. (2019)
- boosting de forêts: Ghosal and Hooker (2020)
- forêts en ligne (données (X_t, Y_t) observées séquentiellement): agrégation en ligne avec des poids mis à jours au cours du temps Zhong et al. (2020), Mourtada, Gaïffas, and Scornet (2019)



Interprétation des RF

Nous avons déjà vu que la notion d'importance peut permettre d'interpréter les modèles boites noires de type $Y = f(X) + \varepsilon$, par exemple une forêt.

On peut également vouloir visualiser les effets univariés $f_k(x_k)$ ou les effets bivariés $f_q(x_k, x_l)$ afin de rendre **intelligible** le comportement du modèle à des variations de X comme cela est possible par construction dans un modèle GAM. Pour simplifier et sans perte de généralité on considère dans la suite que $X = (X_1, X_2)$.

Interprétation des RF: Partial depandance plots (PDP)

Proposé par J. H. Friedman (2001)}:

$$f_{1,PD}(x_1) = E(f(x_1, x_2)) = \int f(x_1, x_2) p_2(x_2) \, \mathrm{d}x_2$$

ou $p_2(x_2)$ est la loi marginale de X_2. Cela se traduit par l'estimateur empirique suivant:

$$\widehat{f}_{1,PD}(x_1) = \frac{1}{n} \sum_{i=1}^{n} f(x_1, x_{i,2})$$

- les variables sont supposées indépendantes, en pratique lorsque x_1 est fixé les autres variables ne peuvent pas prendre n'importe quelles valeurs.
- l'intégrale se fait "loin" des données ou f est estimé, il y a donc une extrapolation sous-jacente ce qui peut être dangereux avec les méthodes non-paramétriques.

Interprétation des RF: Marginal plots (MP)

pour éviter ce pb d'extrapolation on peut définir des PDP "locaux":

$$f_{1,PD}(x_1) = E(f(x_1, x_2)/X_1 = x_1) = \int f(x_1, x_2) p_{2/1}(x_2/x_1)\, \mathrm{d}x_2$$

$$\widehat{f}_{1,M}(x_1) = \frac{1}{n(x_1)} \sum_{i \in V(x_1)}^{n} f(x_1, x_{i,2})$$

ou $p_{2/1}(x_2/x_1)$ est la densité conditionnelle de X_2/X_1, $V(x_1)$ est un voisinage de x_1 à définir par l'utilisateur.

- on ne sépare par les effets communs de X_1 et X_2 dans f
- revient à faire une régression de Y sur X_1 en ignorant X_2

Interprétation des RF: Accumulated Local Effects plots (ALE)

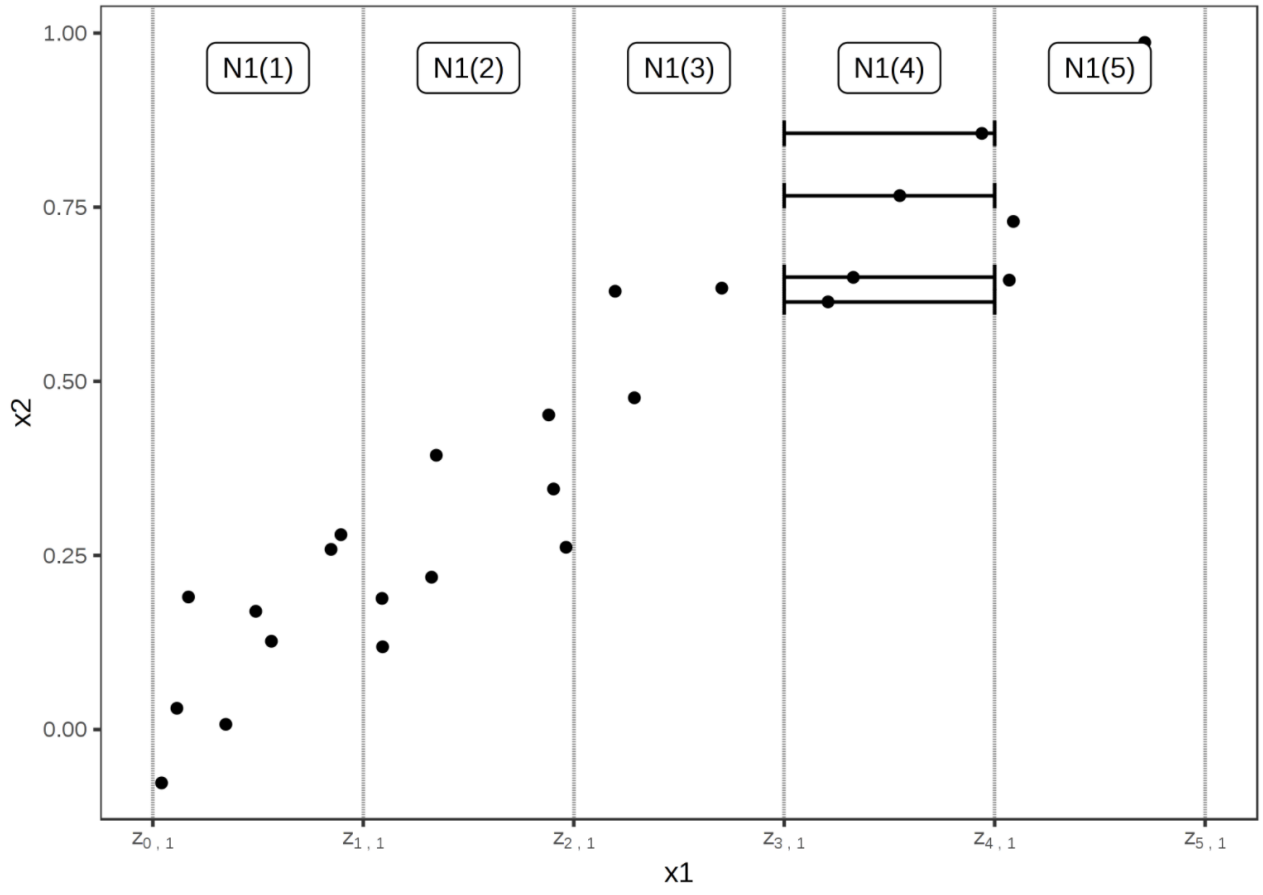
présenté dans Apley and Zhu (2016)

$$f_{1,ALE}(x_1) = \int_{x_{min,1}}^{x_1} E(f^1(x_1, x_2)/X1)$$

$$f_{1,ALE}(x_1) = \int_{x_{min,1}}^{x_1} \int p_{2/1}(x_2/z_1) f^1(z_1, x_2) \mathrm{d}x_2 \mathrm{d}z_1 - c$$

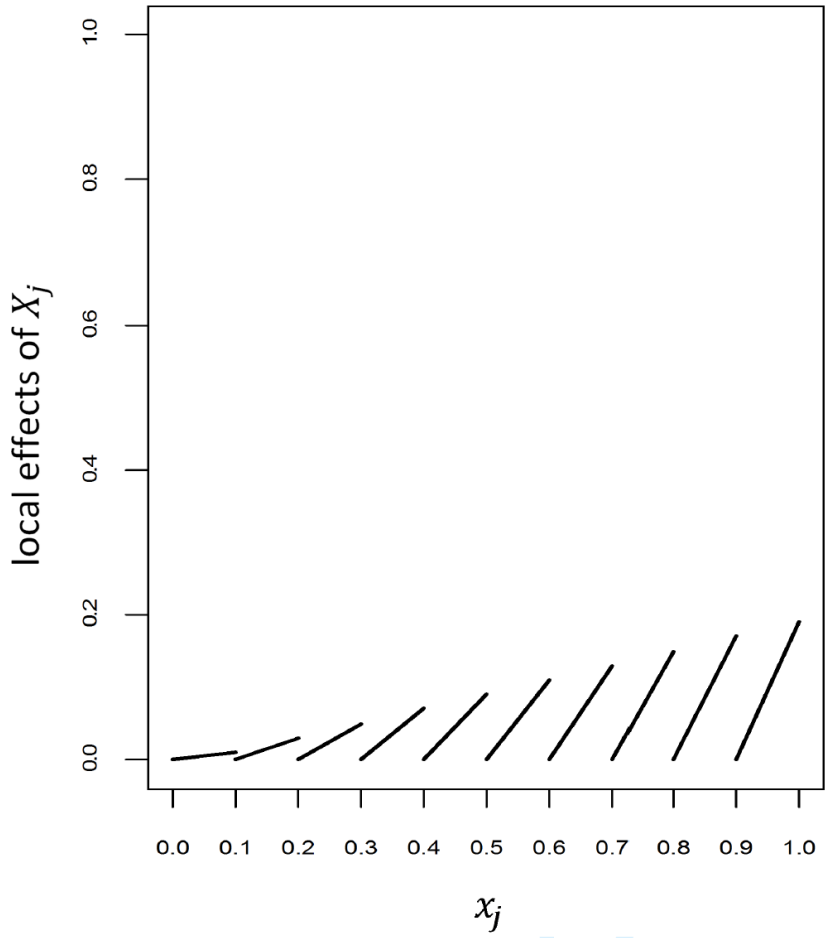
ou $p_{2/1}(x_2/x_1)$ est la densité conditionnelle de X_2/X_1, $f^1 = \partial f(x_1, x_2)/\partial x_1$. x_min est la valeur min du support de p_1, c une constante pour que les ALE soient centrés.

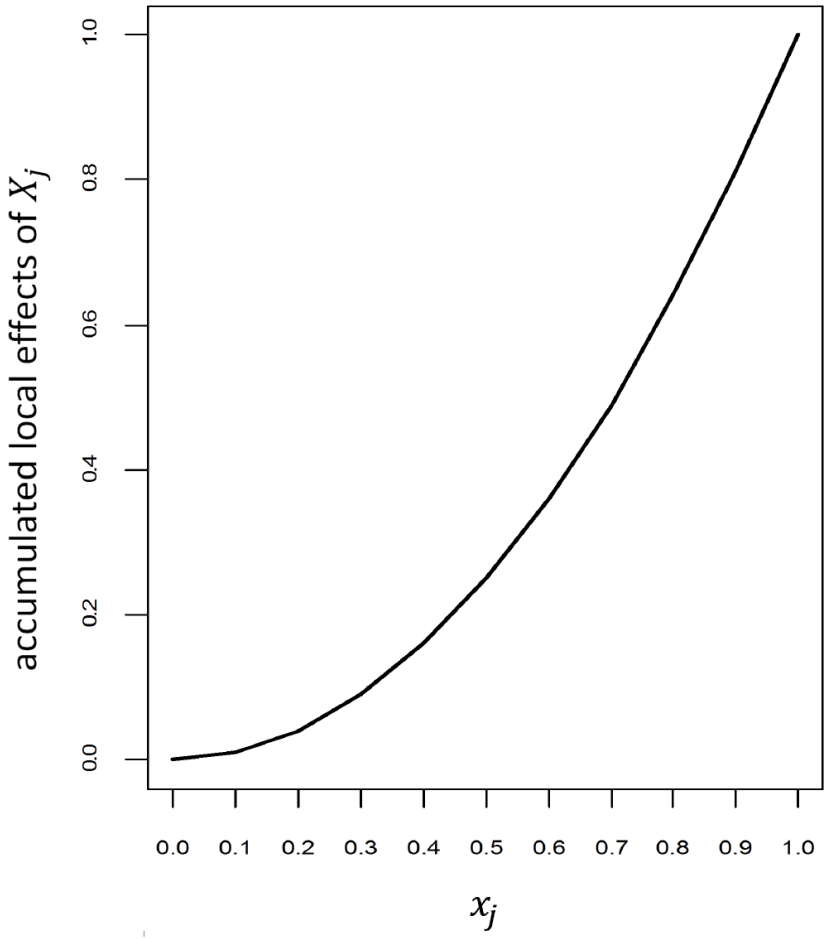
Accumulated Local Effects plots (ALE)



On partitionne les données en K intervalles, pour chaque point d'un intervalle on calcule la différence entre les différences de prévision $f(x_{max}) - f(x_{min})$ puis on somme et on centre. Graphe issu de @molnar2019interpretable.

Accumulated Local Effects plots (ALE)





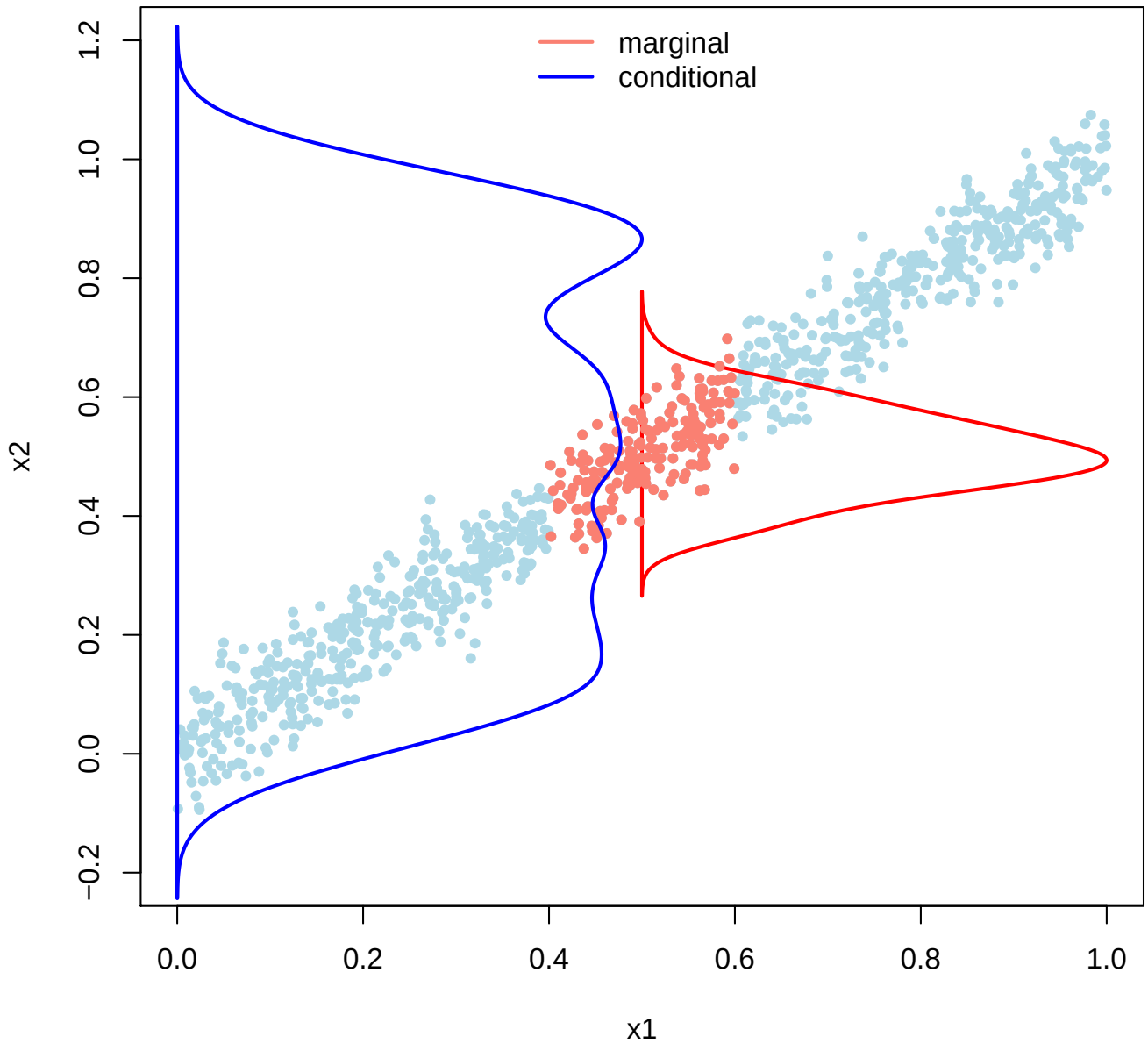
Graphe issu de Apley and Zhu (2016).

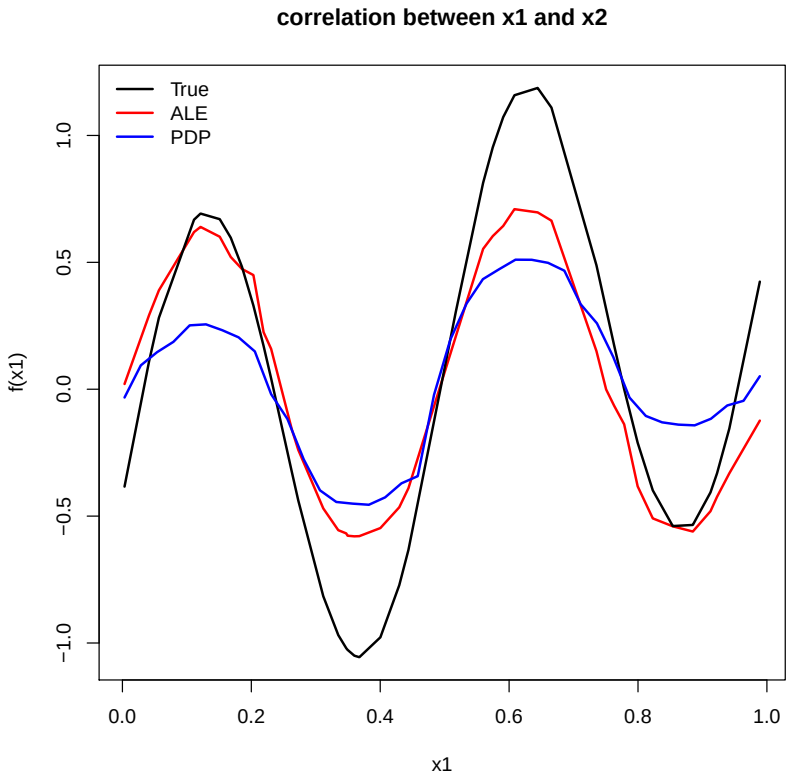
ALE plots: Exemple illustratif

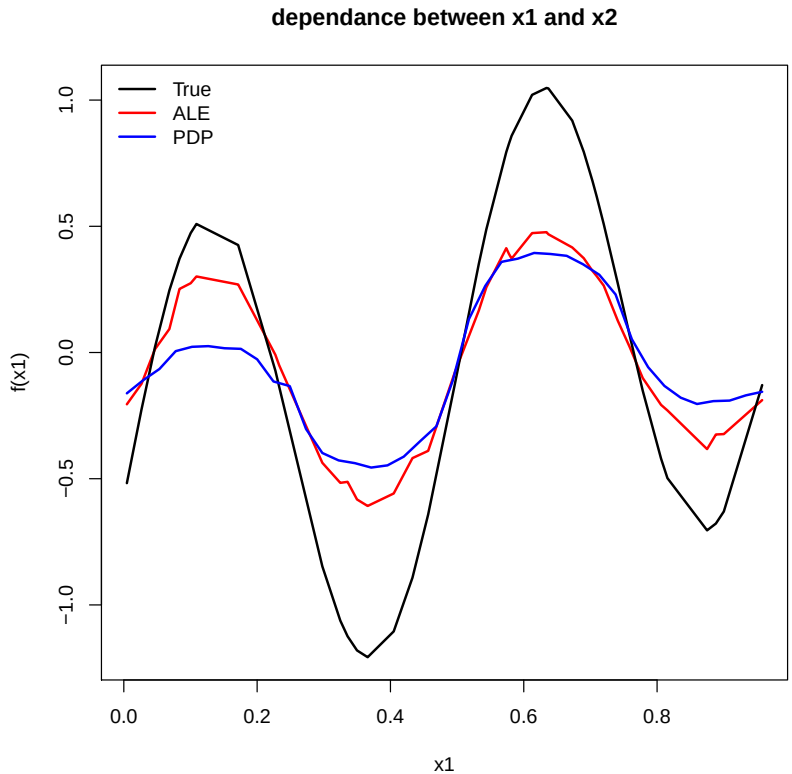
Les données sont simulées selon le protocole suivant:

- U une variable uniforme sur $[0, 1]$
- Cas **corrélé**: $X_1 = U$ et $X_2 = U + Z_1$ ou Z_1 est une variable gaussienne $\mathcal{N}(\mu = 0, \sigma = 0.05)$
- Cas **dépendant**: $X_1 = U$ et $X_2 = X1$
- la réponse Y est définie ainsi (ou $\varepsilon : \mathcal{N}(\mu = 0, \sigma = 0.05)$):

$$Y = sin(4 * \pi * X1) + X_2 + \varepsilon$$


marginal
conditional
x2
x1
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0


correlation between x1 and x2
True
ALE
PDP
f(x1)
x1
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8
1.0


dependance between x1 and x2
True
ALE
PDP
f(x1)
x1
1.0
0.5
0.0
−0.5
−1.0
0.0
0.2
0.4
0.6
0.8

Implementations

en R:

- package originel `randomForest` puis `ranger` faisant maintenant référence
- package `party`, fonction `cforest`
- `VSURF` pour la sélection de variable automatique
- `grf` pour les forêts généralisées
- `ALEPlot` pour l'interprétation par représentation des ALE

References

Apley, Daniel W, and Jingyu Zhu. 2016. "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models." *arXiv Preprint arXiv:1612.08468*.

Arlot, Sylvain, and Robin Genuer. 2014. "Analysis of Purely Random Forests Bias." *arXiv Preprint arXiv:1407.3939*.

Athey, Susan, Julie Tibshirani, Stefan Wager, and others. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2). Institute of Mathematical Statistics: 1148–78.

Biau, Gérard, and Erwan Scornet. 2016. "A Random Forest Guided Tour." *Test* 25 (2). Springer: 197–227.

Biau, Gérard, Erwan Scornet, and Johannes Welbl. 2019. "Neural Random Forests." *Sankhya A* 81 (2). Springer: 347–86.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2). Springer: 123–40.

———. 2001. "Random Forests." *Machine Learning* 45 (1). Springer: 5–32.

Breiman, Leo, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees*. CRC press.

Clémençon, Stéphan, Marine Depecker, and Nicolas Vayatis. 2013. "Ranking Forests." *Journal of Machine Learning Research* 14 (Jan): 39–73.

Fawagreh, Khaled, Mohamad Medhat Gaber, and Eyad Elyan. 2015. "An Outlier Detection-Based Tree Selection Approach to Extreme Pruning of Random Forests." *arXiv Preprint arXiv:1503.05187*.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*. JSTOR, 1189–1232.

Genuer, Robin, and Jean-Michel Poggi. 2017. "Arbres Cart et forêts Aléatoires, Importance et Sélection de Variables."

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. 2010. "Variable Selection Using Random Forests." *Pattern Recognition Letters* 31 (14). Elsevier: 2225–36.

Georganos, Stefanos, Tais Grippa, Assane Niang Gadiaga, Catherine Linard, Moritz Lennert, Sabine Vanhuysse, Nicholus Mboga, Eléonore Wolff, and Stamatis Kalogirou. 2019. "Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling." *Geocarto International* 0 (0). Taylor & Francis: 1–16. doi:10.1080/10106049.2019.1595177.