Apprentissage supervisé avec données manquantes

Nicolas Prost [1*†⋆], Julie Josse [2**], Erwan Scornet [3**] & Gael Varoquaux [4†⋆]

[1]*nicolas.prost@inria.fr,* [2]*julie.josse@polytechnique.edu,*
[3]*erwan.scornet@polytechnique.edu,* [4]*gael.varoquaux@inria.fr.*
[*]*CMAP (Centre de Mathématiques Appliquées),* [†]*Inria, équipe Parietal.*
[⋆] *Bâtiment Turing, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau.*

Résumé. Dans de nombreuses applications, les données sont affectées par des valeurs manquantes, qui perturbent l'analyse statistique. Une littérature abondante traite des données manquantes dans un cadre d'inférence, où l'objectif est d'estimer des paramètres et leurs variances à partir de tableaux incomplets. Ici, nous considérons un cadre d'apprentissage supervisé où l'objectif est de prédire au mieux une variable cible lorsque des données manquantes apparaissent à la fois dans le jeu d'apprentissage et de validation.

Nous montrons la consistance de deux approches pour estimer la fonction de régression. Nous prouvons en particulier que l'imputation par la moyenne en amont de la phase d'apprentissage, technique très utilisée en pratique, est consistante lorsque les données manquantes ne sont pas informatives. Ceci contraste avec le contexte inférentiel où l'imputation par la moyenne est connue pour ses sérieux inconvénients en termes de déformation des lois marginales et jointe des données. Le fait qu'une approche si simple soit consistante a d'importantes conséquences en pratique.

Ce résultat est valable asymptotiquement, pour un algorithme d'apprentissage dont le risque, en l'absence de données manquantes, tend vers zéro. Nous apportons des analyses supplémentaires sur les arbres de décision, car ils sont naturellement adaptés à la minimisation du risque empirique avec données manquantes. Cela est dû à leur capacité à prendre en compte la nature semi-discrète des variables avec données manquantes. Après avoir comparé théoriquement et empiriquement différentes stratégies pour prendre en compte les données manquantes dans la construction des arbres, nous recommandons d'utiliser la méthode "missing incorporated in attribute" car elle peut gérer des données manquantes informatives et non informatives.

Mots-clés. Imputation, arbres de décision, algorithme EM

Abstract. In many application settings, the data are plagued with missing features. These hinder data analysis. An abundant literature addresses missing values in an inferential framework, where the aim is to estimate parameters and their variance from incomplete tables. Here, we consider supervised-learning settings where the objective is to best predict a target when missing values appear in both training and test sets. We analyze which missing-values strategies lead to good prediction.

We show the consistency of two approaches to estimating the prediction function. The most striking one shows that the widely-used mean imputation prior to learning method

is consistent when missing values are not informative. This is in contrast with inferential settings as mean imputation is known to have serious drawbacks in terms of deformation of the joint and marginal distribution of the data. That such a simple approach can be consistent has important consequences in practice.

This result holds asymptotically when the learning algorithm is consistent in itself. We contribute additional analysis on decision trees as they can naturally tackle empirical risk minimization with missing values. This is due to their ability to handle the half-discrete nature of variables with missing values. After comparing theoretically and empirically different missing-values strategies in trees, we recommend using the missing incorporated in attributes method as it can handle both non-informative and informative missing values.

Keywords. Imputation, decision trees, expectation maximization

1 Contexte

Les données s'accumulent et se diversifient. Cela vient à la fois d'une prise de conscience de leur importance dans les prises de décision, et de progrès techniques. Cette hétérogénéité pose de nombreux problèmes de pré-traitement. Les données manquantes sont un de ces problèmes : pour certains échantillons, seulement une partie des covariables est observée.

D'une part, les données manquantes ont été largement étudiées depuis les années 1970, avec en tête de file Rubin (1976). Il y a eu essentiellement deux axes d'étude, détaillés par Little et Rubin (1987, 2002). Le premier est la qualité d'estimation dans des modèles paramétriques, où l'objectif est d'approcher les paramètres sans données manquantes, notamment grâce à un algorithme EM (Dempster et al., 1977). Le deuxième axe est l'imputation, qui consiste à estimer du mieux possible les valeurs non disponibles (imputation simple) ou à conserver la distribution des données (imputation multiple). On répertorie classiquement les données manquantes selon trois mécanismes : MCAR *(missing completely at random)* lorsque l'absence d'une variable est indépendante de toutes les variables, MAR *(missing at random)* lorsque elle peut dépendre d'autres variables observées, et MNAR *(missing not at random)* autrement (Little et Rubin, 1987, 2002).

D'autre part, l'apprentissage supervisé consiste à approcher une fonction de lien entre des *covariables*, que l'on représente par un vecteur $\mathbf{X}$, et une *réponse* Y, avec l'objectif de minimiser une fonction de perte entre la valeur prédite et la vraie réponse (Vapnik 1999). Afin de mesurer la capacité de généralisation d'un prédicteur, on distingue généralement un jeu d'apprentissage, et un jeu de validation sur lequel on mesure l'erreur.

En apprentissage supervisé, les données manquantes apparaissent naturellement dans $\mathbf{X}$, autant dans le jeu d'apprentissage que dans le jeu de validation, d'où la nécessité d'une méthode de traitement adaptée à cet objectif particulier. En pratique, les méthodes les plus souvent utilisées dans la communauté du machine learning sont l'imputation de

chaque colonne par sa moyenne, et l'ajout d'une variable supplémentaire indiquant la présence ou non de données manquantes dans la colonne en question. Pourtant, ces deux méthodes sont connues pour leurs mauvaises propriétés : l'imputation par la moyenne perturbe la distribution des covariables (Little et Rubin, 1987, 2002), et l'ajout de l'indicatrice biaise la régression dans le modèle linéaire (Jones 1996). Nous étudions leurs propriétés en prédiction.

Nous nous attachons tout d'abord à formaliser le problème d'apprentissage supervisé avec données manquantes, en faisant le lien entre les deux littératures. Dans ce cadre, notre seconde contribution est d'étudier la consistance de deux approches. Premièrement, étant donné un prédicteur optimal pour des données complètement observées, l'imputation multiple permet d'établir une procédure consistante sur un jeu de validation. Deuxièmement, l'imputation par la moyenne préalable à l'apprentissage est consistante pour l'apprentissage supervisé. Ce dernier résultat a d'importantes conséquences en pratique puisqu'il justifie cette approche communément utilisée. Nous nous concentrons enfin sur les arbres de décisions, car ils sont adaptés à la prise en compte des données manquantes, comme d'autres types de variables non standard. Analytiquement et empiriquement, nous montrons l'intérêt d'une méthode appelée "Missing Incorporated in Attribute" (MIA, Twala et al., 2008) pour prendre en compte les données manquantes dans l'apprentissage. Nous montrons également l'intérêt d'ajouter la présence de données manquantes comme autant de variables explicatives supplémentaires afin d'aider les méthodes plus classiques dans des cas de données manquantes non homogènes.

2 Résultats théoriques

Notre premier résultat concerne la consistance de l'imputation multiple pour prédire sur un jeu de validation avec un prédicteur parfait. Pour tout vecteur $\widetilde{\mathbf{x}} \in (\mathbb{R} \cup \{\texttt{NA}\})^d$, soient $\mathbf{m}$ l'indicateur de manque, $\mathbf{x}_o = o(\mathbf{x}, \mathbf{m})$ les données observées et $\mathbf{x}_m$ les données manquantes. On tire les données manquantes $\mathbf{X}_m$ dans leur distribution conditionnelle à $\mathbf{X}_o = \mathbf{x}_o$ et on calcule la fonction de régression sur ces observations complétées. L'imputation multiple correspondante est donnée par

$$f^{\star}_{mult\ imput}(\widetilde{\mathbf{X}}) = \mathbb{E}_{\mathbf{X}_m|\mathbf{X}_o=\mathbf{x}_o}[f(\mathbf{X}_m, \mathbf{x}_o)]. \tag{1}$$

Théorème 2.1. *Considérons le modèle de régression*

$$Y = f(\mathbf{X}) + \varepsilon,$$

où $\mathbf{X} = (X_1, \ldots, X_d)$ *est à valeur dans* $\mathbb{R}^d$*, en supposant que pour tout sous-ensemble* $\mathcal{S} \subset \{1, \ldots, d\}$*,* $(M_j, j \in \mathcal{S}) \perp\!\!\!\perp (X_j, j \in \mathcal{S}) | (X_k, k \in \mathcal{S}^c)$ *(mécanisme MAR) et que* $\varepsilon \perp\!\!\!\perp (M_1, X_1, \ldots, M_d, X_d)$ *est un bruit centré. Alors l'imputation multiple décrite ci-dessus*

est consistante, c'est-à-dire, pour tout $\widetilde{\mathbf{x}} \in (\mathbb{R} \cup \texttt{NA})^d$,

$$f^\star_{mult\ imput}(\widetilde{\mathbf{x}}) = \mathbb{E}[Y|\widetilde{\mathbf{X}} = \widetilde{\mathbf{x}}].$$

Notre deuxième résultat porte sur la consistance de l'imputation par une constante, lorsque celle-ci est faite avant l'apprentissage, et par la même valeur dans les jeux d'apprentissage et de validation sous l'hypothèse MAR.

Théorème 2.2. *Supposons que le vecteur* $\mathbf{X} = (X_1, \ldots, X_d)$ *a une densité continue* $g > 0$ *sur* $[0,1]^d$ *, que la réponse*

$$Y = f(\mathbf{X}) + \varepsilon$$

est telle que $\|f\|_\infty$ *existe, que le manque* M_1 *sur la variable* X_1 *vérifie* $M_1 \perp\!\!\!\perp X_1 | X_2, \ldots, X_d$ *et est tel que la fonction*

$$(x_2, \ldots, x_d) \mapsto \mathbb{P}[M_1 = 1 | X_2 = x_2, \ldots, X_d = x_d],$$

soit continue. Supposons de plus que ε *est un bruit centré indépendant de* $(\mathbf{X}, M_1)$*. Les données imputées* $\mathbf{X}' = (X_1', X_2, \ldots, X_d)$ *sont définies par* $X_1' = X_1 \mathbb{1}_{M_1=0} + \mathbb{E}[X_1]\mathbb{1}_{M_1=1}$*. Si la capacité d'approximation de l'algorithme d'apprentissage est suffisamment élevée, il prédira, pour tout* $\mathbf{x}' \in \mathbb{R}^d$*,*

$$\begin{aligned} f^\star_{impute}(\mathbf{x}') = {} & \mathbb{E}[Y|X_2 = x_2, \ldots, X_d = x_d, M_1 = 1]\mathbb{1}_{x_1' = \mathbb{E}[X_1]}\mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]>0} \\ & + \mathbb{E}[Y|\mathbf{X} = \mathbf{x}']\mathbb{1}_{x_1' = \mathbb{E}[X_1]}\mathbb{1}_{\mathbb{P}[M_1=1|X_2=x_2,\ldots,X_d=x_d]=0} \\ & + \mathbb{E}[Y|\mathbf{X} = \mathbf{x}', M_1 = 0]\mathbb{1}_{x_1' \neq \mathbb{E}[X_1]}. \end{aligned}$$

Etant donné

$$\widetilde{\mathbf{X}} = \begin{cases} \mathbf{X}' & \text{if } X_1' \neq \mathbb{E}[X_1] \\ (\texttt{NA}, X_2, \ldots, X_d) & \text{if } X_1' = \mathbb{E}[X_1] \end{cases},$$

la prédiction résultant de l'imputation par la moyenne est égale à la fonction de Bayes presque sûrement, c'est-à-dire

$$f^\star_{impute}(\mathbf{X}') = \widetilde{f}^\star(\widetilde{\mathbf{X}}).$$

3 Résultats expérimentaux

Nous comparons différentes méthodes pour traiter les données manquantes dans les arbres de décision, avec comme métrique le pourcentage de variance expliquée (R^2) – on cherche à obtenir un score élevé. Pour comparer plus précisément les méthodes, le score affiché est le R^2 relatif : pour chacun des 500 tirages, on centre les scores en leur soustrayant leur moyenne. `rpart` et `ctree` sont des arbres qui utilisent par défaut des *surrogate splits*,

une méthode qui ne prend pas en compte le mécanisme de données manquantes. Dans un modèle à trois variables où $Y = X_1^2 + \varepsilon$, la figure 1 montre que dans le cas MNAR, ces méthodes et l'imputation par une loi jointe gaussienne sont moins performantes que les méthodes de codage des données manquantes : imputation par la moyenne, MIA, ajout de l'indicatrice, et propagation par bloc – une méthode proche de MIA utilisée par XGBoost (Chen and Guestrin, 2016). Dans le cas MCAR, ces dernières méthodes sont aussi performantes que les premières. La troisième figure représente un modèle où l'indicatrice est une variable prédictive ($Y = X_1^2 + 3M_1 + \varepsilon$) : dans ce cas il est également avantageux d'encoder les données manquantes. Les expériences ont été menées avec des arbres simples et des forêts aléatoires.

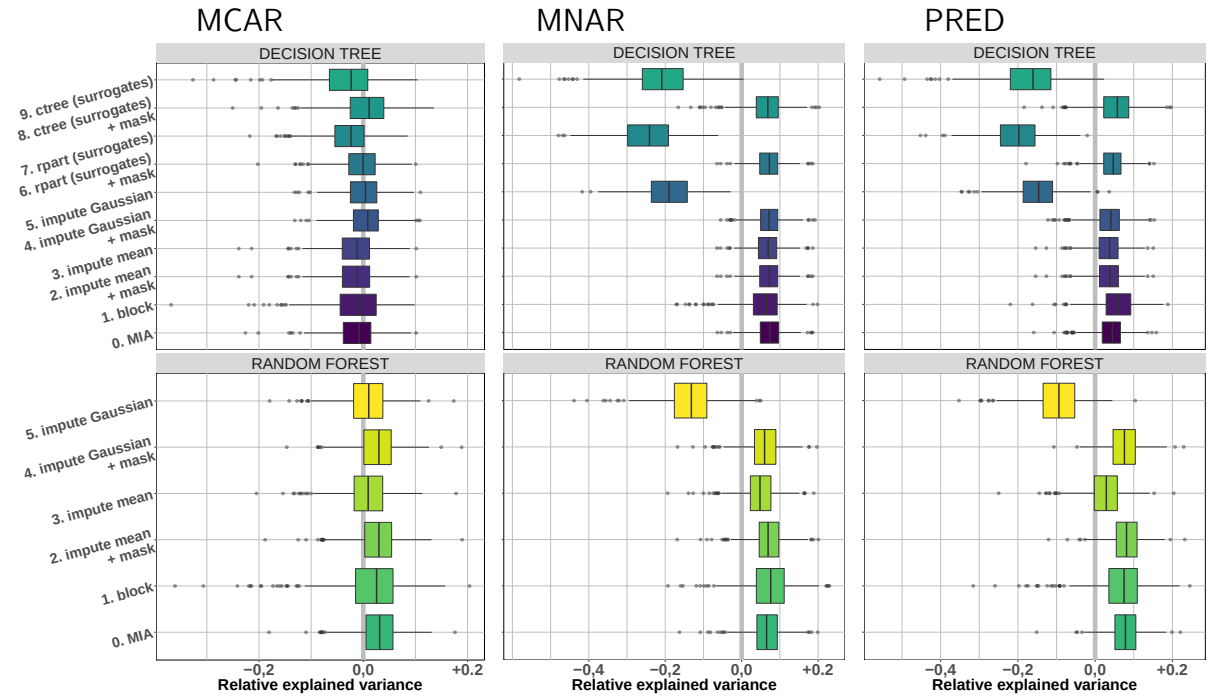

FIGURE 1 – **Scores relatifs** • Pourcentages de variance expliquée, relatifs, pour différents mécanismes avec 20% de valeurs manquantes sur la première variable.

4 Conclusion

Cette étude de l'apprentissage supervisé avec données manquantes nous permet d'émettre plusieurs recommandations. Premièrement, si l'on dispose d'un modèle appris sur données complètes, l'imputation multiple rend possible une bonne prédiction sur un jeu de test. Deuxièmement, le même modèle d'imputation doit être utilisé sur le jeu d'apprentissage et le jeu de test. L'imputation par la moyenne suffit, à condition d'avoir un modèle suf-

fisamment puissant. Troisièmement, une bonne imputation accélère la consistance, mais lorsque le mécanisme de données manquantes est lié à la réponse, cela ne suffit pas. C'est pourquoi nous recommandons la méthode "missing incorporated in attribute" qui permet de prendre en compte le mécanisme en construisant les arbres. Ces résultats sont valables asymptotiquement, il reste à établir des résultats théoriques à régime fini.

Bibliographie

T. Chen et C. Guestrin. Xgboost : A scalable tree boosting system. In *sigkdd international conference on knowledge discovery and data mining*, page 785. ACM, 2016.

A. P. Dempster, N. M. Laird, et D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

M.P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433), 222-230, 1996.

R.J. Little et D.B. Rubin. *Statistical analysis with missing data.* John Wiley & Sons, 1987, 2002.

D. B. Rubin. Multiple Imputation for Nonresponse in Surveys. *Wiley*, 1987.

B. E. T. H. Twala, M. C. Jones, et D. J. Hand. Good methods for coping with missing data in decision trees. *Pattern Recogn. Lett.*, 29(7) :950–956, May 2008. ISSN 0167-8655.

V. N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5) :988–999, 1999.