

Handling Missing Values with BEST

Cédric Beaulac

University of Toronto

(joint work with Jeffrey S. Rosenthal)

June 6, 2018

Introduction

- ▶ Fonctionne lors de l'apprentissage et de la prédiction
- ▶ Works when fitting the parameters and when predicting

- ▶ Ne suppose aucune structure pour les valeurs manquantes
- ▶ Doesn't assume any missingness structure

- ▶ Conserve l'interprétabilité des Arbres de Décisions
- ▶ Keep the interpretability property of Decision Trees

Presentation Plan

Introduction

Decision Tree

 The model

 Fitting the parameters

Branch-Exclusive Splits Trees (BEST)

 Motivation

 The model

 Handling missing values with BEST

Results

Arbre de Décision

Le modèle

- ▶ Pour faire de la classification, nous divisons l'espace des variables explicatives en régions disjointes
- ▶ In k -class classification problems we divide the feature space into k disjoint regions
- ▶ Un arbre de classification représente une division de cet espace
- ▶ A classification tree models the response by a partitioning of the feature space

Le modèle

- ▶ Le classificateur est de la forme suivante :
- ▶ The classifier is of the form :
- ▶ $\hat{f}(x) = \sum_{r=1}^{|R|} c_r I\{x \in R_r\}$
- ▶ où c_r est la valeur prédite pour une observation appartenant à la région r
- ▶ where c_r is the label assigned for an observation that belongs in region r

Le modèle

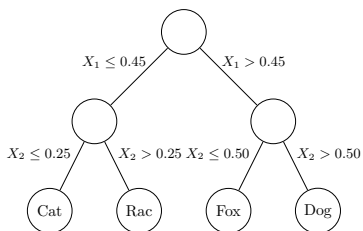
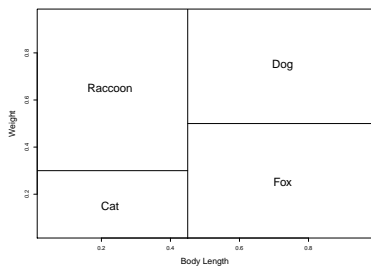


Figure: Decision tree

Le modèle

- ▶ Le classificateur est facilement interpretable
- ▶ The classifier is highly interpretable

- ▶ Il intègre naturellement des interactions de hauts degrés.
- ▶ Allows for high degree of interaction

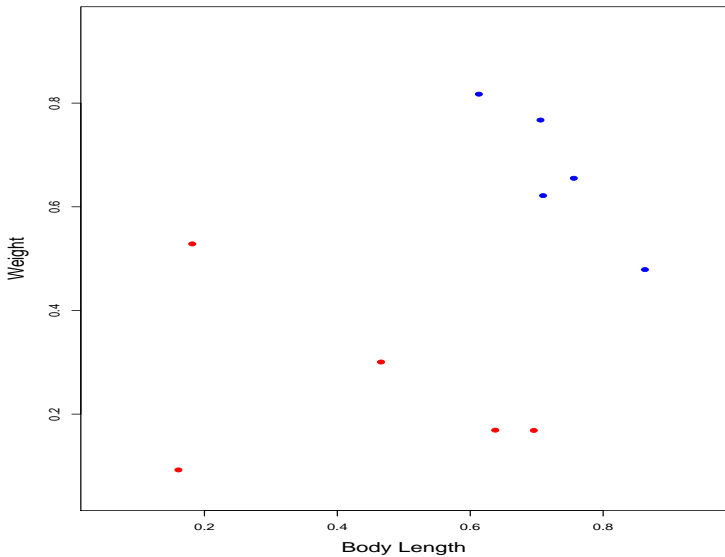
- ▶ Ne présuppose aucune distribution
- ▶ Doesn't assume any distribution

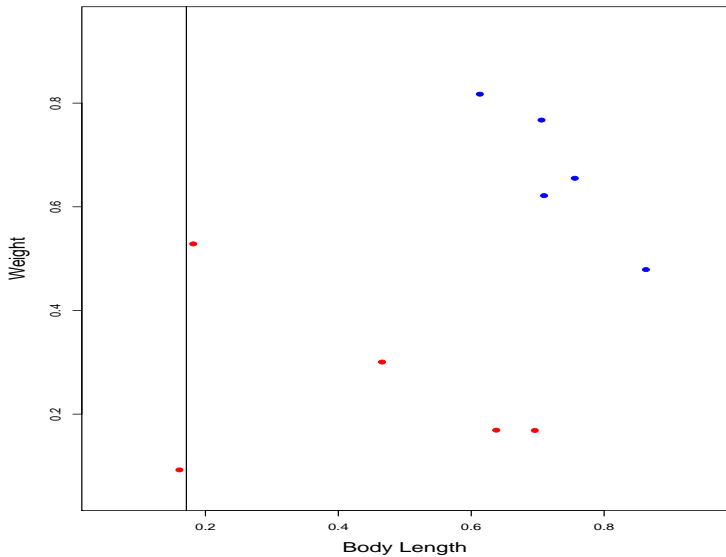
Déterminer les paramètres

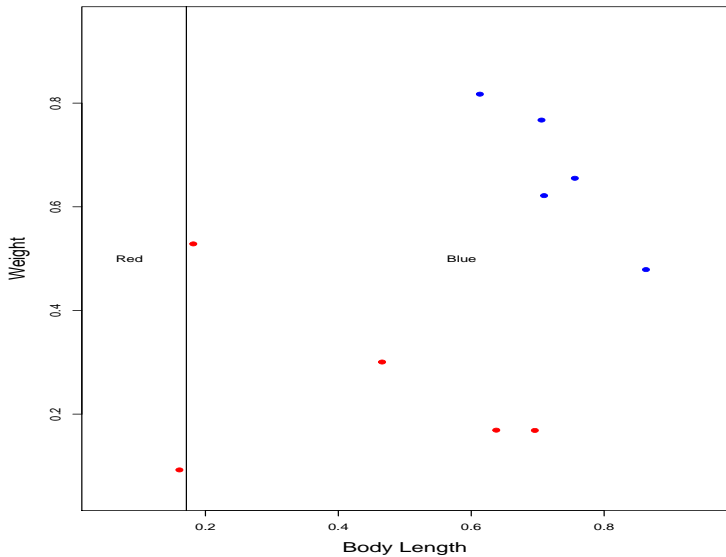
- ▶ L'algorithme examine toutes les partitions possibles des données
- ▶ The algorithm looks through all binary partitioning of the training data
- ▶ L'algorithme choisi la partition qui minimise une mesure d'erreur préalablement définie
- ▶ It selects the partitioning that reduces the most a pre-specified error measure
- ▶ Par exemple : L'erreur de de classification
- ▶ Example : Missclassification error

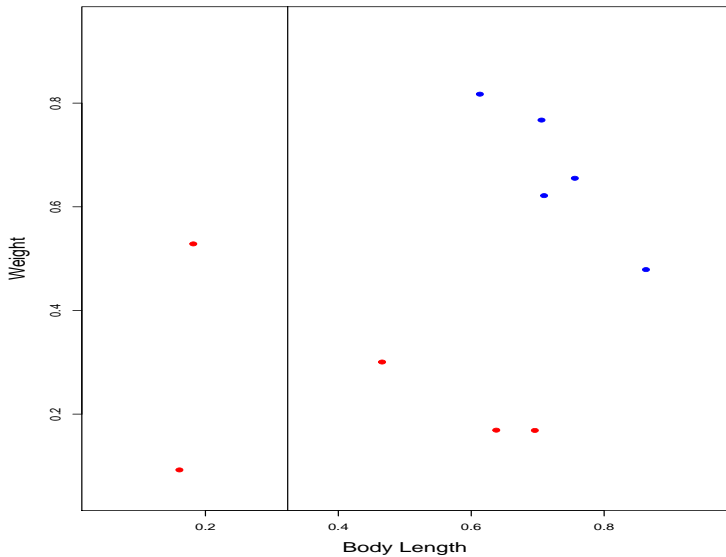
Déterminer les paramètres

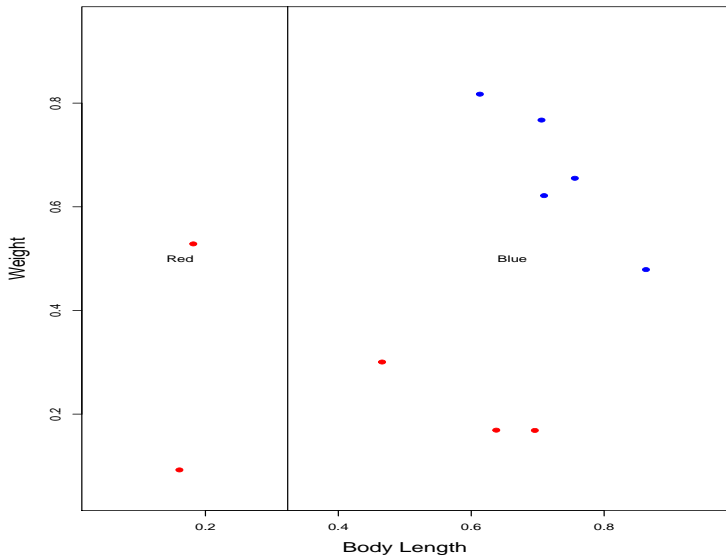
- ▶ Le processus de division des données arrête selon une certaine condition d'arrêt
- ▶ The partitioning process stops when a stopping rule is applied
- ▶ Traditionnellement, le processus s'arrête lorsque le nombre d'observation dans une région est inférieur à un certain seuil
- ▶ Usually when the number of observations in a region is lower than a threshold

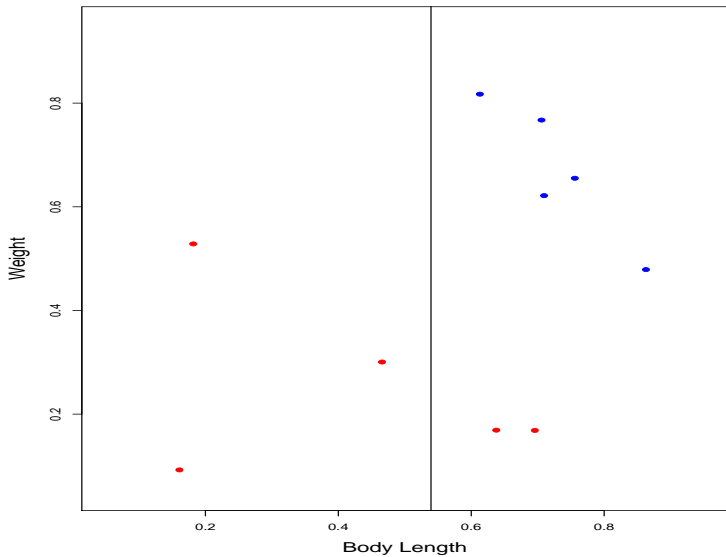


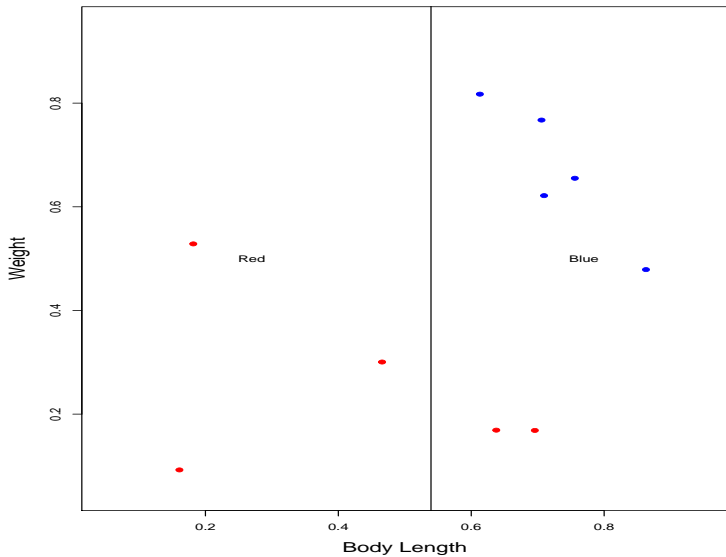


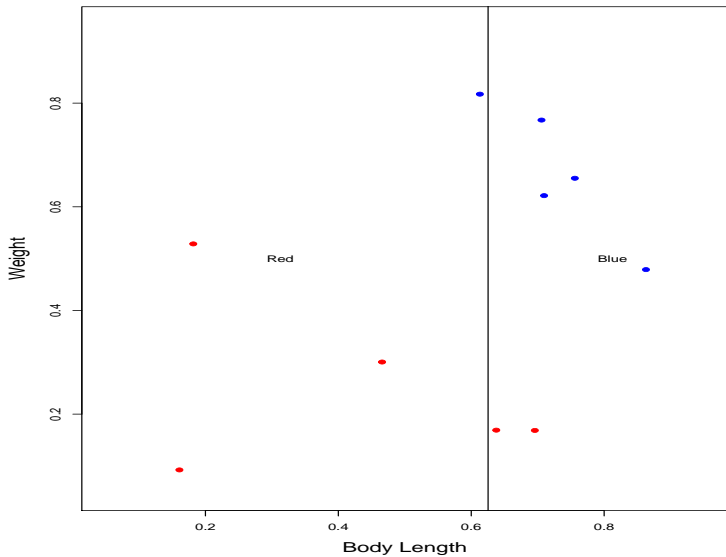


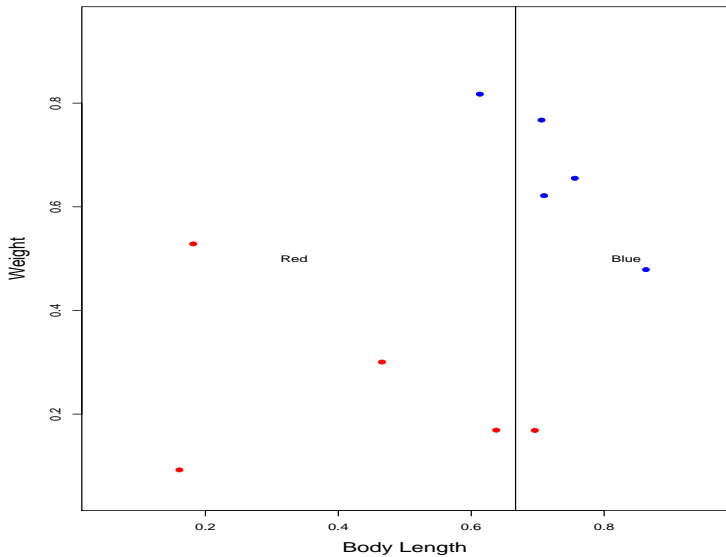


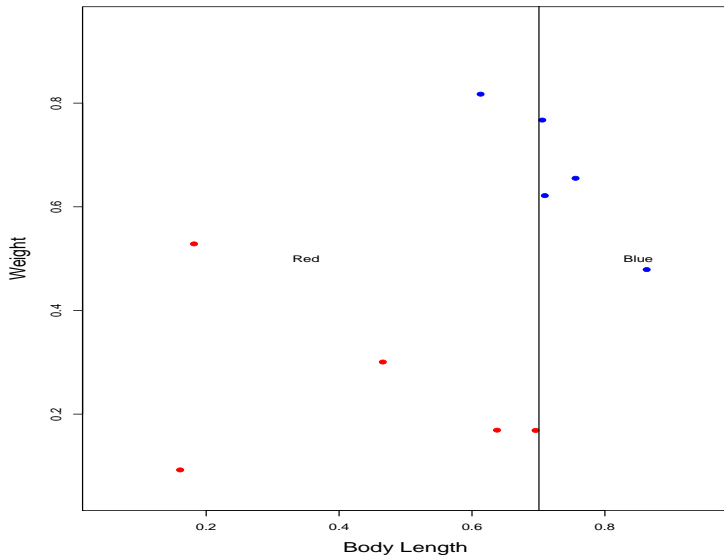


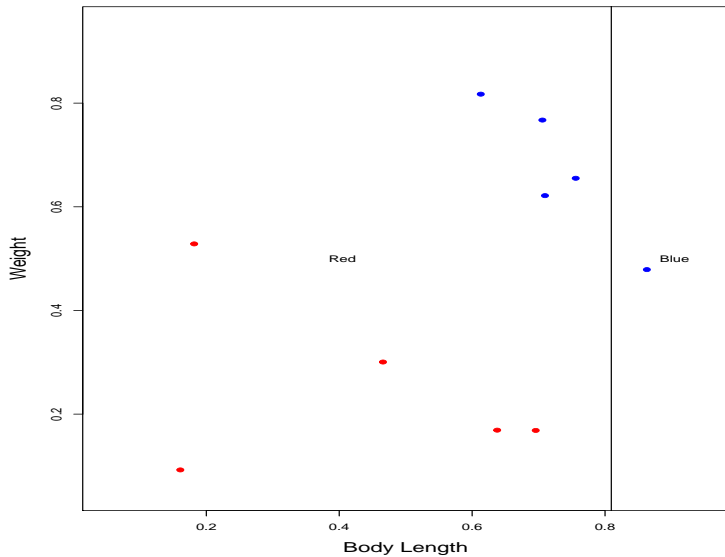


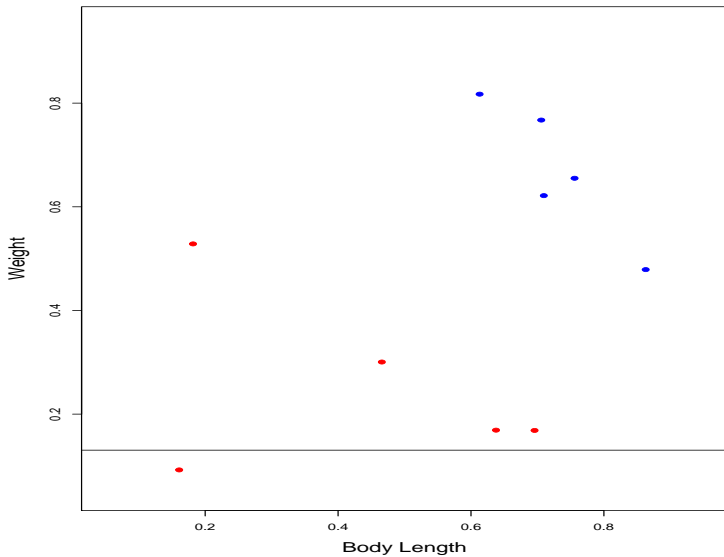


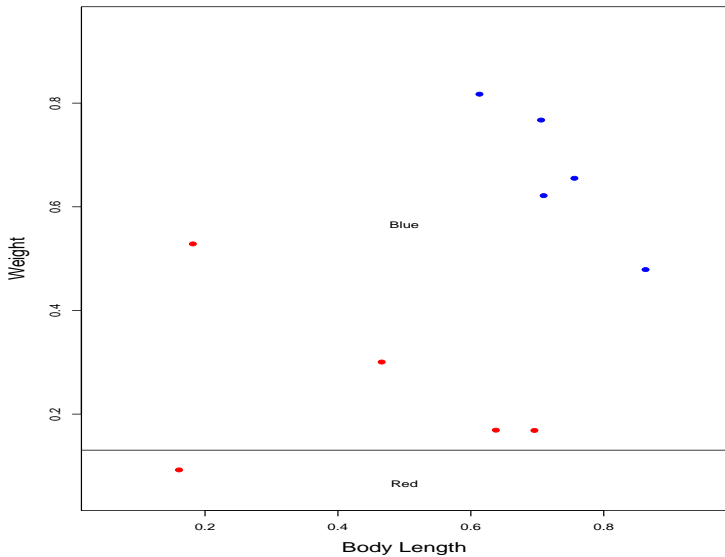


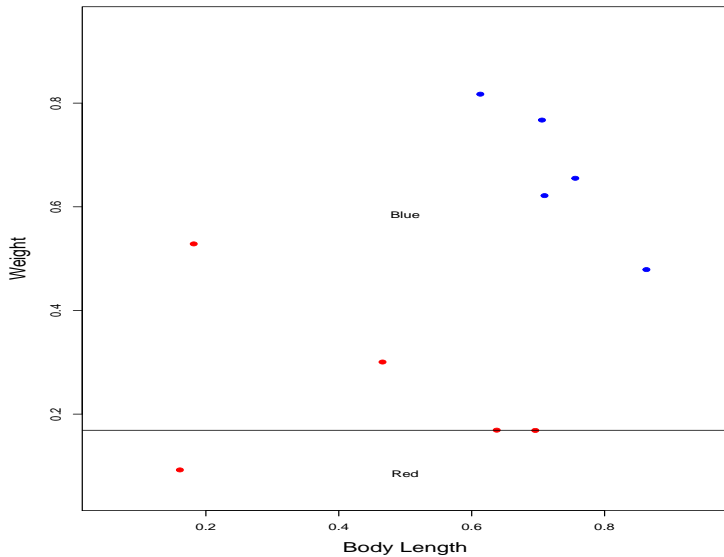


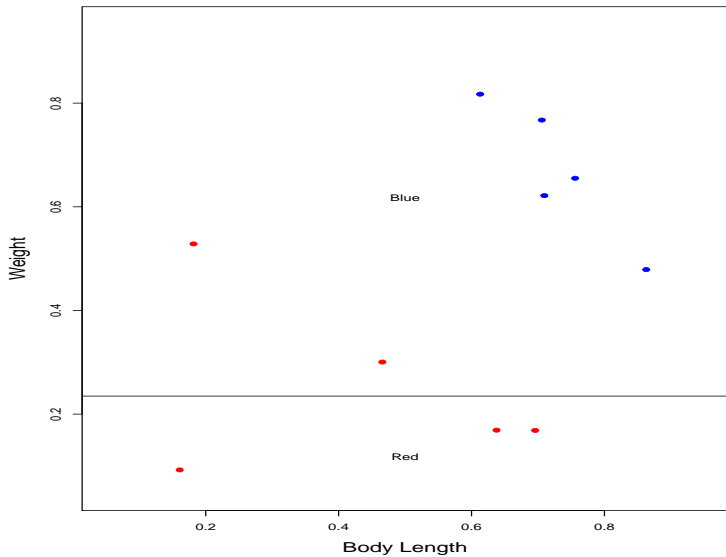


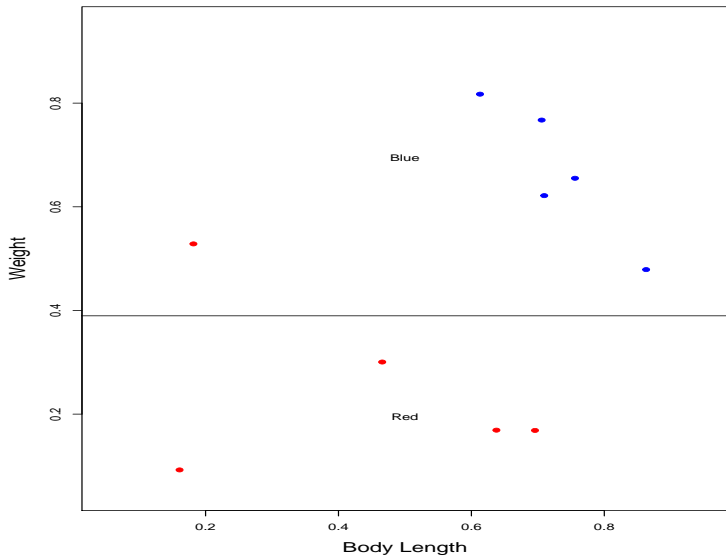


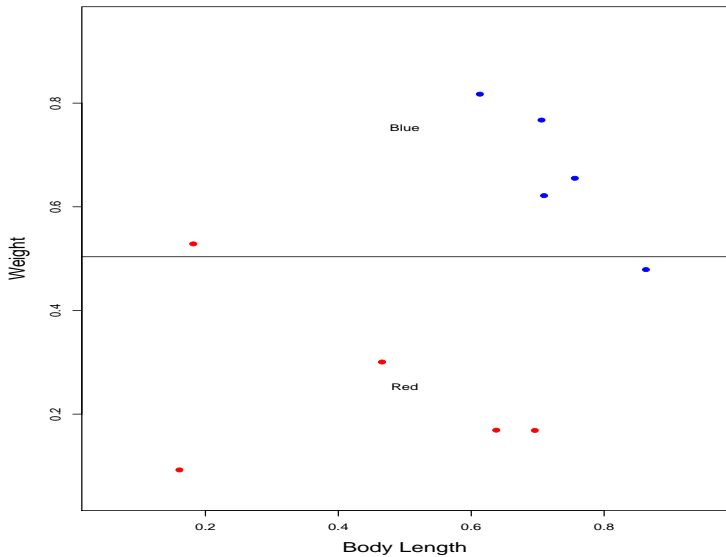


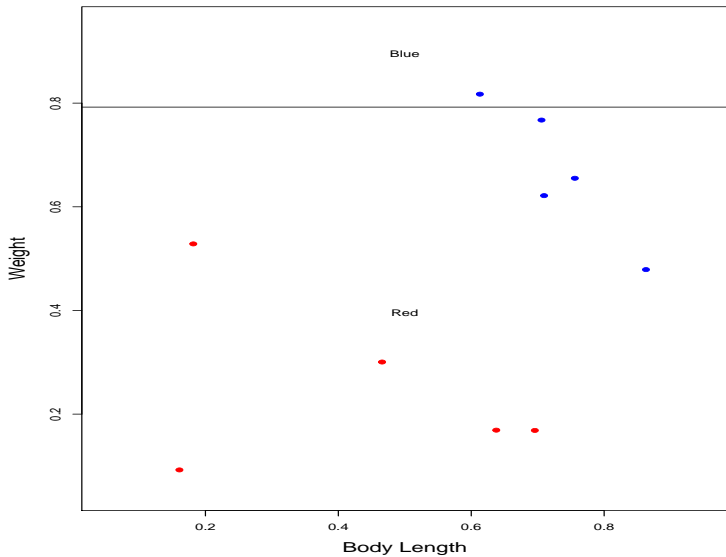


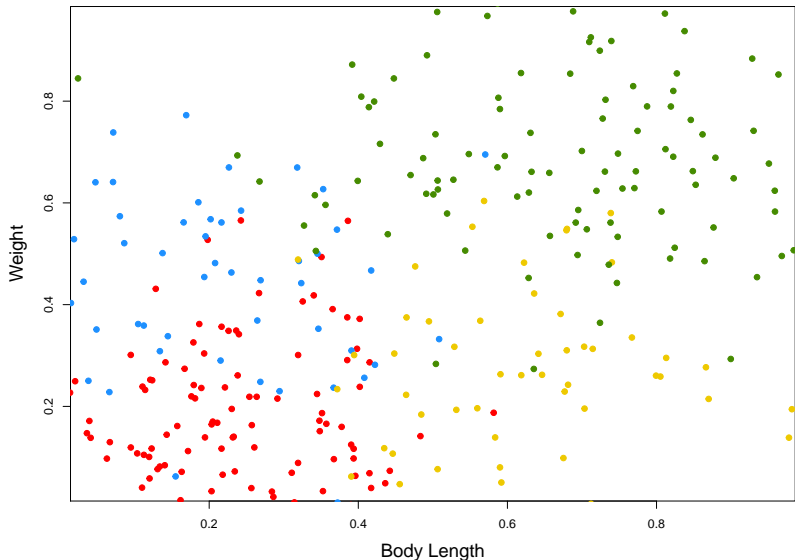


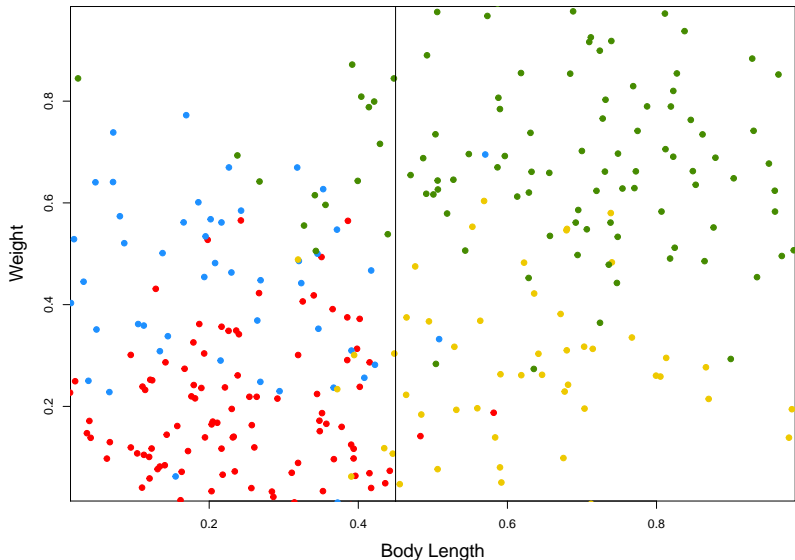


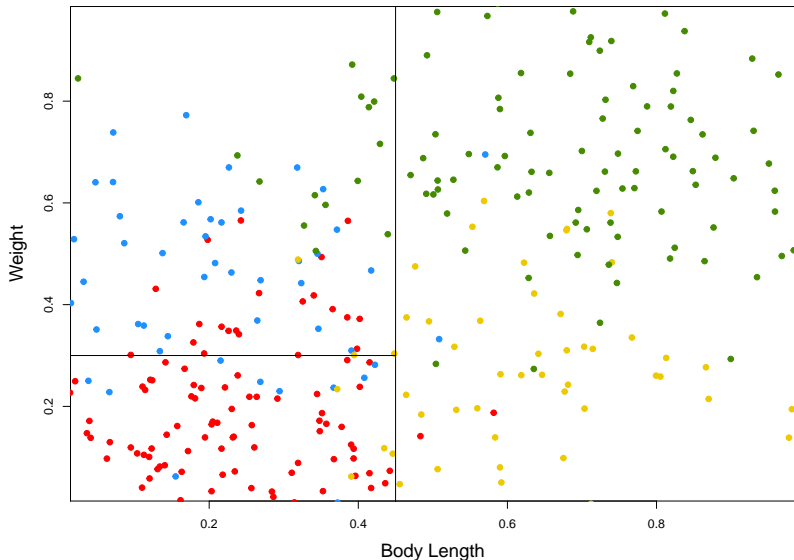


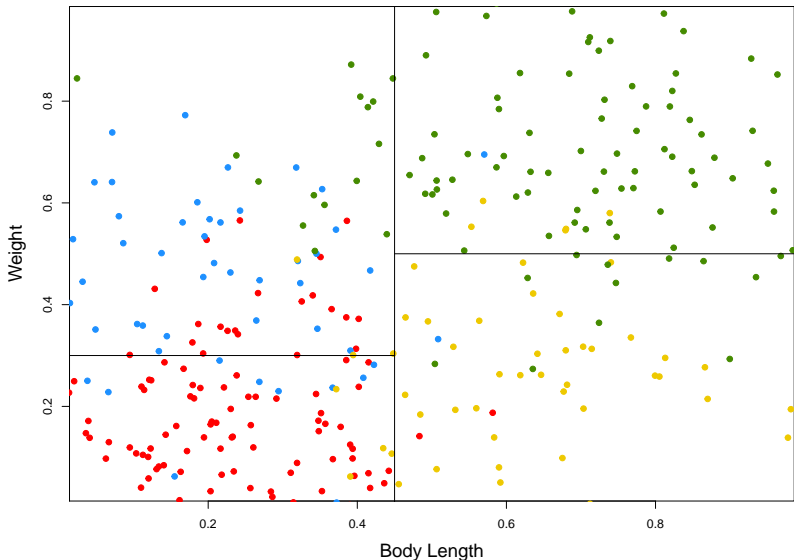


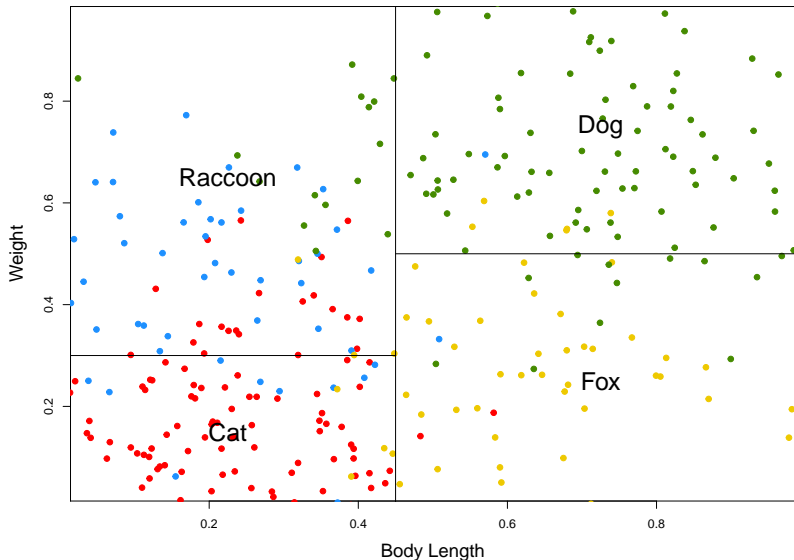












Valeurs manquantes

- ▶ On ne peut pas diviser les données en utilisant un prédicteur contenant des valeurs manquantes
- ▶ It's impossible to partition the data according to a variable containing missing values

Branch-Exclusive Splits Trees (BEST)

Motivation

- ▶ Données concernant les résultats d'étudiants universitaires
- ▶ Data set about university student's record

- ▶ Le nombre de crédit et la note moyenne dans chaque département sont observés
- ▶ Contains number of credits and averaged grades in every department

- ▶ Nous voulons faire de la prédiction en utilisant la première année de cours
- ▶ We were aiming at predicting various things using the first year of courses

Motivation

- ▶ La note est manquante si l'étudiant ne s'est inscrit à aucun cours
- ▶ Grades are missing if the number of credit attempted is 0
- ▶ Nous pouvons utiliser le *nombre de crédit* pour déterminer quand la *note moyenne* est une variable manquante
- ▶ We can use the *number of credits* variable to define where the *averaged grade* variable is missing

Le modèle

- ▶ Modèle très général où seulement une partie des prédicteurs est considérée pour le partitionnement
- ▶ General framework where only a determined subset of variable is available for partitioning process
- ▶ La liste des prédicteurs considérés est mise-à-jour après chaque partitionnement
- ▶ The list of predictors available is updated after a split is performed

Le modèle

- ▶ Des variables peuvent être ajoutées ou retirées de l'ensemble de prédicteurs analysé lors du partitionnement
- ▶ Variables can be added or removed as desired depending on previous split variable and split point
- ▶ Nous construisons alors un arbre dans lequel certaines variables sont exclusives à certaines branches
- ▶ We effectively build a tree with branch-exclusive variables
- ▶ Peut être utile pour imposer une structure particulière, construire des forêts ou gérer les valeurs manquantes
- ▶ Can be used for multiple purposes : Force a particular ordering, build special kinds of forests or manage missing values.

Gérer les valeurs manquantes avec BEST

- ▶ L'idée est d'utiliser un ensemble de prédicteurs pour définir la région où un prédicteur avec des valeurs manquantes est disponible pour le partitionnement
- ▶ The idea is to use a set of predictors to define regions where predictors with missing values are available for the data partitioning process
- ▶ Pour notre exemple précédent : Quand le nombre de crédits est plus grand que zéro, alors la note moyenne devient disponible
- ▶ For the Student's Data Set : When the *number of credits* > 0 , then the grade variable becomes available for that branch

Gérer les valeurs manquantes avec BEST

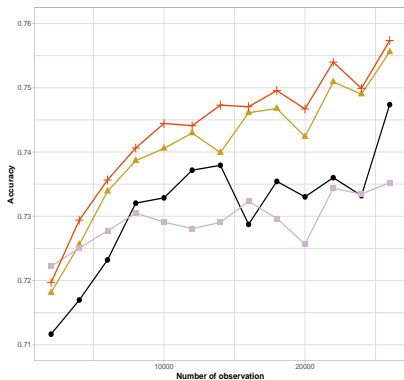
- ▶ S'il n'y existe pas de variable de la sorte, nous pouvons ajouter une variable binaire
- ▶ If the data set contains no such variable, we can simply add a dummy variable
- ▶ Les résultats sont faciles à interpréter et nous n'avons fait aucune hypothèse par rapport aux valeurs manquantes.
- ▶ The results are highly interpretable, works for both fitting and predicting and we did not assume any missingness structure

Résultats

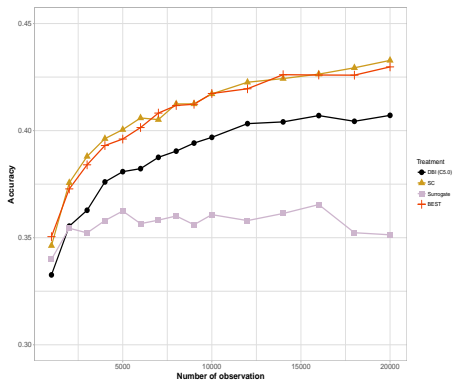
Données Réelles

- ▶ BEST est plus précis que les autres techniques mises à l'épreuve
- ▶ BEST produces a more accurate classifier than other missing value techniques
- ▶ Le classificateur obtenu est facile à interpreter
- ▶ The classifier is way more interpretable

Grades data sets



(a) Program Completion



(b) Choice of Major

Figure: Accuracy of various algorithms on a real data set.

Merci!

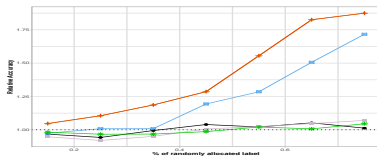
Fitting the parameters

- ▶ We then proceed at pruning the tree
- ▶ Control for the number of leaves by adding a cost for additional leaves
- ▶ The tuning parameters are usually selected using cross-validation or validation set

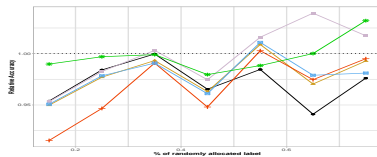
Simulated data sets

- ▶ Number of observation is not an important factor.
- ▶ The strength of the relation between the predictors and the response and the missingness pattern has an effect on performance.

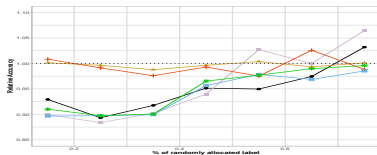
Simulated data sets



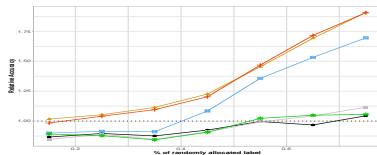
(a) MAR : Y



(b) MAR : X



(c) MNAR



(d) Mixed

Treatment

- ◆ DBI (C5.0)
- Surrogate
- VI (Mode)
- ▲ SC
- + BEST
- * EMPOL