University of California, Los Angeles
Department of Statistics

Imputation of Missing Values in CART

An Application of the EM Algorithm

Author:
Thomas Maierhofer
SID 905078249

Stats 201C:
Prof. Quing Zhou

Final Project Report

Spring Quarter 2019,
June 14, 2019

Contents

Abstract

Classification and Regression Trees (CARTs) are one of the most widely used algorithms in machine learning. There are multiple ways of treating missing values in covariates, many of which work well in different scenarios. This report proposes a novel strategy for imputing missing values based on an EM algorithm. In the expectation step (E step), missing values are imputed based on the leaf node in which the corresponding observation falls given the current CART. In the maximization step (M step), the CART is fit on the current training data. These steps are repeated until convergence, i.e. the imputed values do not change anymore. Using a simulation study where missing values are introduced missing at random into a real data set, this EM algorithm is shown to achieve superior predictive performance in comparison with a number of standard approaches to handling missing values in CART.

1 Introduction

This report introduces a novel approach to handling missing values in CARTs. It is structured as follows: Section 2 introduces CARTs and the EM algorithm, the two main concepts required for the proposed method for handling missing values in CARTs in Section 3. Section 4 showcases the superior performanc in comparison with standard treatments of missing values in a real data set with artificially introduced missing values. The report ends with an outlook on potential future research in Section 5.

2 Background

This section introduces the two cornerstones of this report: Classification and Regression Trees (CARTs) in Section 2.1 and the EM algorithm in Section 2.2.

2.1 Classification And Regression Trees

CARTs were originally introduced by Breiman et al. (1984) and have quickly gained popularity due to their competitive predictive power, ease of implementation, and compellingly simple concept. CARTs create a model based on recursively splitting the covariate space into more homogeneous (w.r.t. the target variable) subspaces, a.k.a. nodes, using binary splits on individual covariates. The final nodes are called leaves and are used for prediction. The basic algorithm for CART works as follows:

Creating CART

1. Initialize empty tree with the root node containing all observations
2. For each node not satisfying the stopping criterion:
   - 2.1 Find the split that minimizes the node impurity in its resulting child nodes
   - 2.2 Add child nodes to the tree.
3. While stopping criterion of the tree is not reached repeat Step 2
4. Return tree with splits and nodes

Note that all splits are binary splits on one of the covariates. Common stopping criteria for nodes are its minimal nodesize or minimal impurity. Nodes that are not split any further are called leaf nodes. A common stopping criteria for the overall tree is its depth, i.e. maximal number of recursive splits. Common measures of node impurity are Gini Index for categorical target variables and MSE for numerical target variables.

2.2 Expectation Maximization Algorithm

The EM algorithm is an iterative algorithm for obtaining the maximum likelihood estimator for parameters in a statistical model under the presence of unobserved variables. The basic concept of the EM algorithm has been used for a long time but was explicitly formalized by Dempster et al. (1977). The EM algorithm iterates until convergence between the expectation step (E step), where the parameters of the model are used to obtain the (expectation of) the unobserved variables, and the maximization step (M step), where the model parameters are chosen to maximize the likelihood given the estimated (originally unobserved, now observed) variables.

3 Treatment of Missing Values in CART

This section introduces an EM algorithm for the treatment of missing values in CARTs in Section 3.2 after a brief review of commonly used methods in Section 3.1.

3.1 Traditional Methods

The most obvious way of treating missing values in a CART (and any other statistical model) is to simply delete observations with missing values. This is problematic if the data is not missing completely at random (MCAR) or if there are not enough complete pbservatiosn left to fit a reasonable model. Alternatively, missing values can be imputed using arithmetic means for scalar variables and modes for categorical variables. This ad-hoc method generally provides good results for MCAR data. Both approaches are very problematic if the data is jsut missing at random (MAR) conditional on the observed data. This is a cmmon assumption, mainly because without it asymptotically true estimators cannot be obtained.

3.2 EM Algorithm

The EM algorithm for the treatment of missing values in CARTs finds the maximum likelihood estimator for the model given its unobserved observations. It has the advantage, that asymptotically it will recover the true model parameters even if the missing values are MAR, i.e. they depend on the observed data. In its E step, missing values are imputed based on the leaf node in which the observations falls given the current CART. Missing values in scalar covariates are imputed as the arithmetic mean of all observations within the leaf node, categorical covariates as the majority class (mode). In its M step, the CART is fit on the current data. These steps are repeated until convergence, i.e. the imputed values do not change anymore. In pseudocode, this algorithm is:

EM for Missing Values in CART

1. Initialize missing values as random draws from their (observed) marginal distribution
2. Fit initial CART on completed data
3. **E step**: Impute missing values as the covariates mean/mode in corresponding leaf node in current model
4. **M step**: Fit CART to current data
5. Repeat Steps 3 and 4 until convergence

4 Simulation Study

The methodology proposed in Section 3 is compared using a simulations study. The well known iris data set (Anderson; 1935), see Figure 3, is used to have a realistic data generating process. A logistic regression model is used to introduce missing values (Section 4.1). The results of the performance comparison are summarized in Section 4.2.

4.1 Introduction of Missing Values

For the purpose of this simulation missing values are introduced into the iris data set at random (MAR). The probability of missing observation X_{ij} for observation $i = 1, \ldots, n$, and covariate $j = 1, \ldots, p,$, i.e. the propensity of X_{ij}, is sampled from a logistic regression model where the target variable is the indicator whether or not X_{ij} is missing and the predictor variables are all other covariates $(X_{ik}), k \neq j$. The general model formula can be written as

$$\begin{aligned} P(X_{ij} \text{ missing}) &= \text{logistic}(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_{j-1} X_{i(j-1)} + \\ &\qquad \beta_{j+1} X_{i(j+1)} + \ldots + \beta_p X_{ip}), \end{aligned} \tag{1}$$

where the model coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ are fixed and known simulation parameters. Here, the parameter values $\beta = (0, 0.2, 0.4, -0.2, -0.2)$ corresponding to the intercept and the covariates "Sepal.Length", "Sepal.Width", "Petal.Length", and "Petal.Width", were chosen to achieve a high probability of missingness in the highly predictive covariates "Petal.Length" and "Petal.Width" and lower probabilities of missingness in "Sepal.Length", "Sepal.Width", in order to create a data set that was notably more difficult to train on than the fully observed data. The marginal shares of missing values sampled by this strategy are visualized in Figure 1.

4.2 Performance Comparison

The predictive performance of the EM algorithm is compared to the strategies proposed in Section 3.1. Additionally, a decision tree based on the fully observed data is included as a reference for the best possible performance given the dataset and general modeling strategy. The results of a 10 fold cross validation are plotted in Figure 2. The accuracy was computed on predictions of fully observed variables. The EM algorithm performs similarly to the full data set and clearly outperforms the model only using observations without any missing values or imputing marginal means. The average performance across the 10 folds of the cross validation is summarized in Table 1. This shows that the proposed EM algorithm is capable of recovering high-dimensional dependencies in a data set with missing values, which results in an improved predictive power in comparison to simpler methods ignoring the MAR structure in the missing values.

Table 1: Average accuracy across the 10 fold cross validation of the classification tree, see Figure 2.

	Full Data	EM	Remove NA	Impute NA
Accuracy	0.93	0.93	0.79	0.69

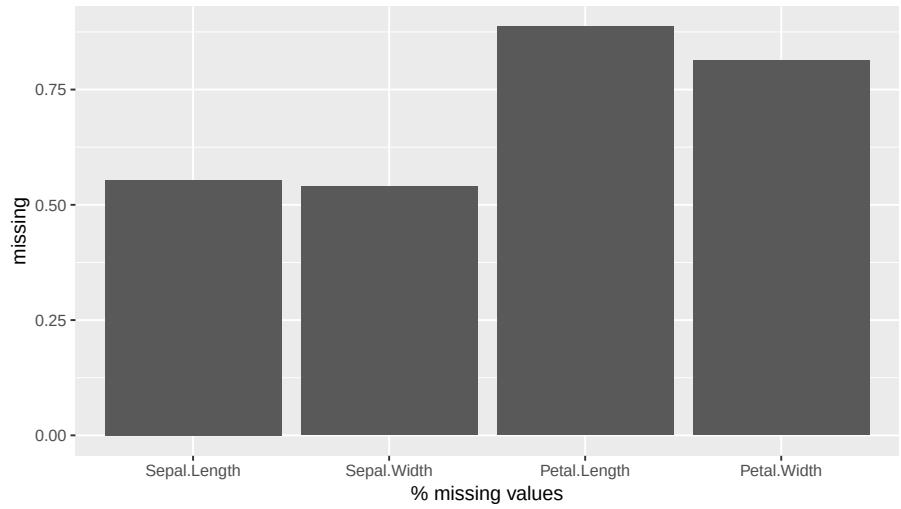

Figure 1: Marginal share of missing values per covariate using the MAR strategy described in Equation (1) and $\beta = (0, 0.2, 0.4, -0.2, -0.2)$.

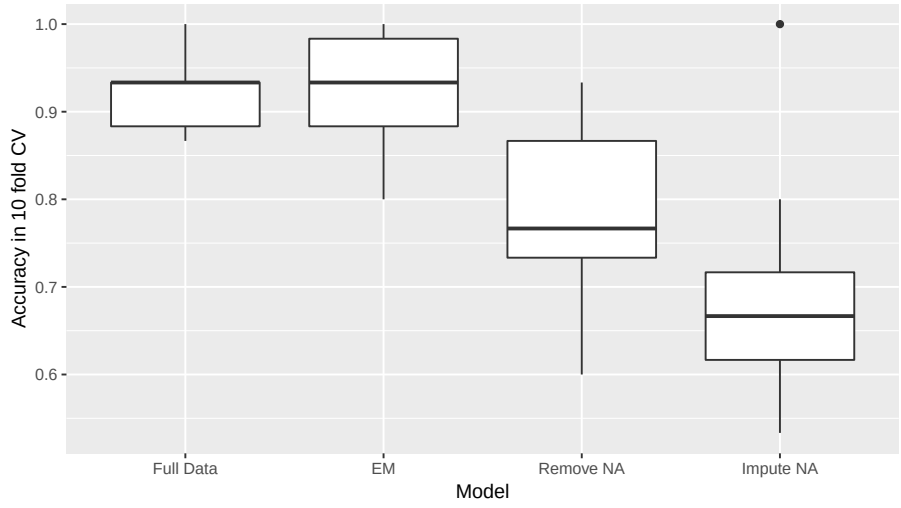

Figure 2: Boxplots of the performance per fold of the classification tree based on the original data without missing values (Full Data), imputation using the EM algorithm proposed in Section 3 (EM), using only observations without missing values (Remove NA), and imputing marginal means (Impute NA).

Figure 3: Iris data: Density of the covariates per species on the diagonal, scatterplots between all covariates in the sub-diagonal plots, and overall and within species correlation coefficients of the covariates in the upper-diagonal plots.

5 Future Research

A high-yield expansion of this project is to generalize the proposed EM algorithm to Random Forests. Random forests are an ensemble of decorrelated CARTs (Breiman; 2001) that generally has increased predictive power in comparison with CART. The EM algorithm for the entire ensemble would simplify to applying the EM algorithm proposed in Section 3.2 separately to each of the CARTs in the ensemble. This should result in similar performance improvements.

Generalizing the proposed EM algorithm to data with missing values in the target variable is straight forward. In the E step of the algorithm proposed in Section 3, missing values in the target variable would be assigned to the mean/mode of the observations in the leaf node they fall into. This should work reasonably well for observations missing their target variable for which relevant covariates are fully observed. This would place the originally mislabeled observation into a leaf node with a majority of its actual class, which would lead to a reclassification in the E step. A similar strategy could also be used to predict labels for observations with missing values in covariates.

A variation on the proposed algorithm would be to create a Gibbs sampler that samples missing values from the marginal distribution of the observed values in the corresponding leaf node. This would lead to slower convergence, but should reduce the probability of getting stuck in a local maximum of the distribution due to the random initial assignment.

References

Anderson, E. (1935). The irises of the Gaspe peninsula, *Bulletin of the American Iris Society* **59**: 2–5.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984). *Classification and Regression Trees*, CRC Press.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1): 1–22.

Digital Appendix

The R code implementing the EM algorithm proposed in Section 3 and the simulation study described in Section 4 can be found in the publicly accessible Github repository `https://github.com/maierhofert/EMforCART.git`. It also contains all figures and the Latex code used to generate this report.