**Homework 2**

**R programming**

**Done by**

**Agnes Priyadharshini**

**Briefly describe the mechanisms of operation of four different data imputation methods.**

The four different data imputation method used are as follows,

* **MICE (Multivariate Imputation via Chained Equations):**

**Mechanism:**

The working mechanism of MICE is similar to regression. It takes the entire instance or observation into account. If one feature is missing in an observation, that missing feature is considered to be dependent variable and the other features are considered to be independent variables. Using independent variables the dependent variable is predicted.

**Example:**

If the observation has 5 features namely, X1,X2,…,X5 and if feature X3 is missing. MICE considers X1,X2,X4,X5 as independent variables and it predicts the dependent variable X3.

**Methods:**

Linear regression – Used to predict continuous missing values

Logistic regression – Used to impute categorical missing values

Bayesian polytomous regression – Used to impute factor variables

**Syntax:**

library(mice)

For continuous missing values, uses predictive mean matching

mice(dataset,m5,maxit = 50,method = ‘pmm’,seed = 500)

For categorical missing values, uses logistic regression

mice(dataset,m5,maxit = 50,method = ‘logreg’,seed = 500)

For missing factor variables, uses Bayesian regression

mice(dataset,m5,maxit = 50,method = ‘polyreg’,seed = 500)

where,

m - refers to number of imputed datasets. Values range from 1 to n

maxit – refers to number of iterations taken to impute the missing values

method – refers to the method used to impute missing values

dataset – refers to the dataset with missing values.

**Cons:**

* + - It works on the assumption that values are missing at random.
    - Worst case scenario, if all the features in an observation are missing then it cannot impute any feature.
* **Amelia:**

**Mechanism:**

Amelia does multiple imputations to impute the missing values. The main reason behind multiple imputations is that it reduces bias and increases efficiency. Also, it uses EMB algorithm. It assumes that the dataset has Multivariate Normal Distribution and that the missing values are at random.

**EMB Algorithm:**

It is an iterative method for approximating the maximum likelihood of a function

**Example:**

If the value of m (explained in the sections below) is 5. It takes 5 samples and applies EMB algorithm on all the five samples separately and computes its mean and variance. The first set of estimates are used to impute the first set of missing values and so on.

**Syntax:**

library(amelia)

amelia(dataset,m = 5, parallel = ‘multicore’)

where,

m - refers to number of imputed datasets. Values range from 1 to n

dataset – refers to the dataset with missing values.

idvars – all other ID variables that do not need to be imputed

noms – keep nominal variables

parallel – type of parallel operation to be used (if any)

**Cons:**

* If the dataset is not normally distributed, it has to be transformed into approximate normality
* Cannot manage imputation of variables defined on a subset of data
* **missForest:**

**Mechanism:**

It is an implementation of random forest algorithm. It builds a random forest model for each variable and uses that model to predict missing values in that variable with the help of observed values. It also computes the error for each variable. The error rate can be reduced by altering number of variables that are randomly sampled at each split and by altering number of trees grown in the forest.

**Example:**

If a dataset has 5 features, it has trees for each feature in the dataset. If feature 1 has missing values, using observed values in the tree for feature 1 it imputes the missing values.

**Syntax:**

missForest(dataset)

where,

dataset – refers to the dataset with missing values.

**Cons:**

* + - Since it uses random forest, it is easy to train but takes time to impute missing values
* **mi (Multiple Imputation With Diagnostics):**

**Mechanism:**

It builds multiple imputation models to estimate impute missing values. Also, it uses predictive mean matching method. It detects irregularities in the data. It also adds noise to the imputation process so as to solve the problem of additive constraints.

**Example:**

If the observation has 5 features namely, X1,X2,…,X5 and if the feature X3 is missing. MI follows predictive mean matching technique as used by MICE to impute the missing value.

**Syntax:**

mi(dataset)

where,

dataset – refers to the dataset with missing values.

The default parameters are,

n.imp – refers to number of imputations which is 3 by default

n.iter – refers to number of iterations which is 30 by default

**Cons:**

* Worst case scenario, if all the features in an observation are missing then it cannot impute any feature.

**Using the iris dataset, randomly create missing values to occupy x% of the data. Use R to apply 3 different methods of imputation and compare their performance when x=2, 5, 10 , 15, 20, 25 (All code snippets must be included in the report appendices and on github). Use the following methods to evaluate the performance of the methods: Root mean square error (RMSE) between imputed and true values Supervised classification error (Use k-NN classifier) Discuss your observations**

**Dataset with missing values:**

**Overall summary:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SepalLengthCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** |
| **Min** | 4.30 | 2.0 | 1 | 0.1 |
| **1st Qua** | 5.10 | 2.8 | 1.6 | 0.3 |
| **Median** | 5.80 | 3.0 | 4.4 | 1.3 |
| **Mean** | 5.84 | 3.05 | 3.79 | 1.19 |
| **3rd Qua** | 6.40 | 3.3 | 5.1 | 1.8 |
| **Max** | 7.90 | 4.4 | 6.9 | 2.5 |

**Overall Observation:**

**NA’s :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SepalLengthCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** |
| **2%** | 2 | 3 | 3 | 3 |
| **5%** | 10 | 10 | 12 | 9 |
| **10%** | 12 | 13 | 13 | 10 |
| **15%** | 23 | 25 | 21 | 23 |
| **20%** | 36 | 28 | 35 | 27 |
| **25%** | 44 | 42 | 28 | 35 |

**Rmse:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MICE** | | | | |
|  | **SepalLenghtCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** |
| **2%** | 0.04396969 | 0.04690416 | 0.05830952 | 0.03366502 |
| **5%** | 0.14628739 | 0.14899664 | 0.16512621 | 0.03366502 |
| **10%** | 0.1860108 | 0.1331666 | 0.1645195 | 0.0697615 |
| **15%** | 0.22876480 | 0.17776389 | 0.14628739 | 0.07071068 |
| **20%** | 0.2916619 | 0.2018250 | 0.4849742 | 0.1812917 |
| **25%** | 0.2861235 | 0.2860070 | 0.4061199 | 0.1796292 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Amelia** | | | | |
|  | **SepalLenghtCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** |
| **2%** | 0.006421378 | 0.083291920 | 0.039864706 | 0.032985595 |
| **5%** | 0.08735461 | 0.07927714 | 0.09392485 | 0.05121226 |
| **10%** | 0.14585519 | 0.07643965 | 0.14736170 | 0.04763895 |
| **15%** | 0.2168240 | 0.1884153 | 0.2174612 | 0.1101037 |
| **20%** | 0.2207145 | 0.2061662 | 0.2968290 | 0.1338118 |
| **25%** | 0.2554428 | 0.2722320 | 0.3065405 | 0.1789488 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **missForest** | | | | |
|  | **SepalLenghtCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** |
| **2%** | 0.03300298 | 0.04319034 | 0.01366136 | 0.02047167 |
| **5%** | 0.05322328 | 0.07888986 | 0.10216081 | 0.05204880 |
| **10%** | 0.09173400 | 0.08137266 | 0.10521685 | 0.04167438 |
| **15%** | 0.13979954 | 0.12361676 | 0.09207207 | 0.08613107 |
| **20%** | 0.1701083 | 0.1357687 | 0.1779475 | 0.1099485 |
| **25%** | 0.2099667 | 0.1692750 | 0.2242650 | 0.1222717 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Mi** | | | | |
|  | **SepalLenghtCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** |
| **2%** | 0.04173904 | 0.11639861 | 0.05044305 | 0.03017918 |
| **5%** | 0.06998713 | 0.10711724 | 0.15688378 | 0.05075442 |
| **10%** | 0.10317619 | 0.14771295 | 0.17978449 | 0.06426186 |
| **15%** | 0.2054814 | 0.2453946 | 0.1780511 | 0.1268568 |
| **20%** | 0.2987464 | 0.2474126 | 0.5458942 | 0.2659223 |
| **25%** | 0.4467000 | 0.3035081 | 0.4646785 | 0.2142657 |

**Recall:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **MICE** | | | **Amelia** | | | **missForest** | | | **Mi** | | |
|  | **Class 1** | **Class 2** | **Class 3** | **Class 1** | **Class 2** | **Class 3** | **Class 1** | **Class 2** | **Class 3** | **Class 1** | **Class 2** | **Class 3** |
| **2%** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| **5%** | 1.0 | 0.98 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.96 | 1.0 |
| **10%** | 1.0 | 0.96 | 0.97 | 1.0 | 0.98 | 1.0 | 1.0 | 0.97 | 0.96 | 1.0 | 1.0 | 0.98 |
| **15%** | 1.0 | 0.96 | 1.0 | 1.0 | 0.98 | 1.0 | 1.0 | 0.98 | 1.0 | 1.0 | 0.98 | 1.0 |
| **20%** | 1.0 | 0.94 | 0.94 | 1.0 | 0.89 | 0.97 | 1.0 | 0.92 | 0.95 | 0.97 | 0.85 | 0.95 |
| **25%** | 1.0 | 0.92 | 0.93 | 1.0 | 0.93 | 0.90 | 1.0 | 0.96 | 0.96 | 1.0 | 0.85 | 0.93 |

**Accuracy:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **MICE** | **Amelia** | **missForest** | **Mi** |
| **2%** | **1.0** | **1.0** | **1.0** | **1.0** |
| **5%** | **0.99** | **1.0** | **1.0** | **0.98** |
| **10%** | **0.98** | **0.99** | **0.98** | **0.99** |
| **15%** | **0.98** | **0.99** | **0.99** | **0.99** |
| **20%** | **0.96** | **0.95** | **0.96** | **0.92** |
| **25%** | **0.95** | **0.94** | **0.97** | **0.92** |

Hence based on the observations above, when the percentage of missing values increases missForest has lower root mean square error and high accuracy when compared to MICE, Amelia and MI.

**Repeat the above process, albeit with the missing values not randomly created (Come up with your own scheme where some classes, feature ids or features having a certain size have higher probability of having missing values. Be sure to explain your scheme and clearly show its implementation).**

**Method used:** Not missing at random

**Scenario:**

The equipment used to measure the sepal length, sepal width, petal length and petal width has an error. So whenever it measures sepal length greater than 6.0. It has a high probability that the next value of sepal width will be missing.

**Overall observation:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **SepalLenghtCm** | **SepalWidthCm** | **PetalLengthCm** | **PetalWidthCm** |
| **NA** | 0 | 61 | 0 | 0 |

Methods to impute SepalWidth:

|  |  |  |
| --- | --- | --- |
|  | **RMSE** | **Overall Accuracy** |
| **MICE** | 0.3244482 | 0.9933 |
| **Amelia** | 0.2808357 | 0.9867 |
| **missForest** | 0.2365418 | 0.9867 |
| **MI** | 0.325297 | 0.9933 |

Hence, based on the above graph the missForest method has the lowest error rate but the accuracy is high in Mice and Mi methods.