Problem Set 3 - Dealing with missing values

NB

1) Which programming languages to use?

You can use Python, R or both of them (but with one limitation: one strategy - one language).

2) What libraries/packages to use?

You are free to choose any appropriate libraries (good choice would be pandas, numpy, scikit-learn for Python and dplyr or tidyr for R).

3) How to summarize my homework?

The best way is to create an Jupyter/R notebook with code and explanations for each strategy. In case you are not familiar with these tools, you can create a Python/R scripts and write explanations as comments. However, we strongly recommend you to use Jupyter/R notebooks, as those are #1 tools in applied data analysis nowadays.

4) Useful links

1. Missing Data Conundrum: Exploration and Imputation Techniques.
2. Dealing with Missing Data Using R
3. Scikit-Learn Imputer Documentation
4. The use of KNN for missing values

5) Grading

The complete solution of this problem set amounts to 5 points of your final grade. There is a bonus problem in this homework for which you can get 2 points.

Tasks

This homework is centered around the process of dealing with missing values. You will have to apply few important strategies in order to make your dataset suitable for further data analysis. Besides it, you will have to compare different methods of dealing with missing values by training a separate logistic regression model for each strategy and measuring their performance.

This time you will work with the dataset from your previous homework: **(Census Income dataset)**.

1) Logistic/Linear regression.

First, learn how to train logistic and linear regression model in Python or R. You don't need to implement them by your own, use library utilities (**scikit-learn** in Python and **glm/lm** in R).

Python examples:

- Classifying Flowers Using Logistic Regression in Sci-Kit Learn
- Building A Logistic Regression in Python, Step by Step
- Linear Regression in Python from Scratch

R examples:

- How to Perform a Logistic Regression in R
- Customer Churn Prediction Model Using Logistic Regression
- Linear Regression in R

The problem we have to solve with our Census dataset is the problem of binary classification.

In case you have never worked with logistic regresssion before, ask course TA (Andrii) or your classmates (who took ML course last year) to help you. This concept is pretty intuitive if you follow the pattern and don't dive deep into theory behind it.

Please note, that logistic regression is not the main topic of this homework. So don't bother yourself tuning your models. Method of dealing with missing values should be the only property that differentiates the models you create in this homework. This is the best way to observe the effects that different strategies of dealing with missing values have on the accuracy of classification.

By the way, we'll share a template of linear/logistic regression implemented in Python for this task.

2) Preparing your dataset for modeling.

To prepare the dataset for further modeling, you need to convert categorical values to numerical ones.

For example, you can replace values in column **'sex'** with 0's for Males and 1's for Females. For columns that have more than 2 categories, generate a set of dummy variables. For example. you can replace column **'education'**, which has 16 distinct values with 16 columns: isBachelors, isHS-grad, is11th... and so on. There are library functions in Python and R that can do it for you automatically.

In general, there are 5 attributes in this dataset that contain missing values: workclass, occupation, capital_gain, capital_loss and native_country. **Transform them only after you have dealt with missing values.**

The dataset is already divided into test and training set. Please remember that each separate strategy of dealing with missing values should be applied to both training and test set.

3) [1pt] Strategy 1: Do not deal with missing values.

3.1. Drop all the rows that contain missing value in any of its **categorical** columns. Don't do anything with missing values (0's) in **numeric** columns.

3.2. Transform categorical columns to numeric values as described in item **2)**.

3.3. Train your model.

3.4. Calculate train and test classification accuracy. Save ther values for further comparison.

3.5. What are the drawbacks of this method? Describe in 3-4 sentences.

4) [1pt] Strategy 2: Global most common substitution.

4.1. Replace the missing values in **categorical** columns with their most common value (mode). Replace missing values (0's) in **numeric** columns with attribute average (mean).

4.2. - 4.4. Same as 3.2. - 3.4.

4.5. What are the limitations of this strategy? Are there any cases when it could be useful? Describe in 4-5 sentences.

5) [2pt] Strategy 3: Regression imputation.

Important!: For **native_country** column, please aggregate values into parts of the world (Europe, Asia, Oceania, North America, South America, Africa), so it would be way easier to create a multivariate classification model to predict missing values in this column.

5.1. For each **categorical** column with missing values, create a dataset in which this column is a target variable. In the dataset obtained, fill the other missing values using methods described in **Strategy 2**.

5.2. Train logistic regression model for each of those datasets (5.1.). Use this model to predict missing values.

5.3. For each **numerical** column with missing values, create a dataset in which this column is a target variable. In the dataset obtained, fill the other missing values using mathods described in **Strategy 2**.

5.4. Train linear regression model for each for those datasets (5.3.). Use this model to predict missing values.

5.5. After filling all the missing values using regression imputation, repeat steps 3.3. - 3.4. with your dataset.

5.6. What are the main drawbacks of this approach? Describe in 3-4 sentences.

6) [1pt] Conclusions

6.1. What strategy of dealing with missing values gave you the best classification accuracy? (3-4 sentences)

6.2. Describe what are the use-cases for different methods of missing value imputation? (4-5 sentences)

7) (Bonus problem) [2pt] Strategy 4: KNN imputation.

7.1. Impute missing values in Census dataset using KNN algorithm. You can use library utilities for this task as well.

7.2. - 7.4. Same as 3.2. - 3.4.

7.5. Compare the results with those obtained using Strategies 1-3.