CMSC 435 Assignment 2

Fall 2017
(individual work; 10 pts total)

This assignment asks you to implement and evaluate several algorithms for imputation of missing values using two provided datasets.

Datasets
There are three datasets. The *assignment2_dataset_complete.csv* file includes a complete dataset that does not have missing values. The *assignment2_dataset_missing004.csv* and *assignment2_dataset_missing20.csv* files include datasets with 0.4% and 20% of missing values, respectively. Both were generated from the *assignment2_dataset_complete.csv* file. You will impute the missing values in each of the latter two datasets and compare these imputed values to the true/correct values that are available in the *assignment2_dataset_complete.csv* file to evaluate and compare quality of imputation.
The three files are in the comma separated value (CSV) format. The first object defines the names of features and the remaining 3587 objects include the values of the corresponding 3587 objects. The first 84 features are numeric and continuous with values in the unit interval, while the last feature that is named class is nominal and binary with values F and C. The missing values are represented by ?. Note that the missing values are not present in the class feature.

Algorithms for missing data imputation
You will implement four methods for the imputation of missing values and apply each of them on the corresponding two datasets that have missing values: *assignment2_dataset_missing004.csv* and *assignment2_dataset_missing20.csv* files.

Mean imputation
Missing value for a specific feature and object is imputed with the mean value computed using the complete values of this feature.

Example

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	?	F
Object 2	0.41139	0.3014	?	C
Object 3	0.24752	0.32148	0.11169	F
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	0.0891	F

To impute the missing value of feature AAcomp_C for object 1, we compute the mean of all complete values of AAcomp_C: *mean* = (0.11169 + 0.13986 + 0.0891) / 3 = 0.11355.

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	0.11355	F
Object 2	0.41139	0.3014	?	C
Object 3	0.24752	0.32148	0.11169	F
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	0.0891	F

The imputed values **must not** be used to compute the means. In other words, all missing values for a given feature are imputed with the same mean value.

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	0.11355	F
Object 2	0.41139	0.3014	0.11355	C
Object 3	0.24752	0.32148	0.11169	F
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	0.0891	F

Conditional mean imputation

Missing value for a specific feature and object is imputed with the mean value computed using the complete values of this feature for objects that satisfy a condition defined by the class feature. For instance, a missing value for an object 1 for which class value = F is imputed based on the mean value computed using all objects for which class value = F.

Example

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	?	**F**
Object 2	0.41139	0.3014	?	C
Object 3	0.24752	0.32148	**0.11169**	**F**
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	**0.0891**	**F**

For object 1 for which class = F, the missing value for AAcomp_C feature is imputed as $\mathit{mean}_{\mathrm{F}}$ = (0.11169 + 0.0891) / 2 = 0.100395
For object 2 for which class = C, the missing value for AAcomp_C feature is imputed as $\mathit{mean}_{\mathrm{C}}$ = 0.13986 / 1 = 0.13986
The two imputed values are

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	**0.100395**	**F**
Object 2	0.41139	0.3014	0.13986	C
Object 3	0.24752	0.32148	**0.11169**	**F**
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	**0.0891**	**F**

Hot deck imputation

Missing values for features that have missing values in a given object are imputed with the values for the same features copied from another, most similar object. First, similarity of a given object that has missing values with every other object in the dataset is computed based on Euclidean distance. The object with the smallest distance is assumed to be most similar and its values are used for the imputation. If that object is missing some of the values that should be imputed then the second most similar object is used to impute these values, and so on. In other words, you should use the first complete value that you find by screening objects by their increasing values of distance. Given two objects $\mathbf{x} = \{x_1, x_2, \dots x_i, \dots, x_n\}$ and $\mathbf{y} = \{y_1, y_2, \dots, y_i, \dots, y_n\}$, the Euclidean distance is calculated as $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ where n is the total number of features (excluding the class future), x_i and y_i are values of feature *i* for objects **x** and **y**, respectively, and $x_i - y_i = 1$ if either x_i or y_i are missing values. The latter penalizes the use of objects that have missing values. Note that you **must not** use the class feature in the calculation of the distance.

Example

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	?	F
Object 2	0.41139	0.3014	?	C
Object 3	0.24752	0.32148	0.11169	F
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	0.0891	F

To impute the missing value of feature AAcomp_C for object 1, you first compute the four distances to every other object

$$d(obj1, obj2) = \sqrt{(0.40256 - 0.41139)^2 + (0.1497 - 0.3014)^2 + 1} = 1.0115$$

$$d(obj1, obj3) = \sqrt{(0.40256 - 0.24752)^2 + (0.1497 - 0.32148)^2 + 1} = 1.0264$$

$$d(obj1, obj4) = \sqrt{(0.40256 - 0.24609)^2 + 1 + 1} = 1.4228$$

$$d(obj1, obj5) = \sqrt{1 + (0.1497 - 0.58306)^2 + 1} = 1.4791$$

Since object 2 that is the most similar to object 1 has a missing value for feature AAcomp_C then the second nearest object 3 is used and the missing value is imputed as follows

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	0.11169	F
Object 2	0.41139	0.3014	?	C
Object 3	0.24752	0.32148	0.11169	F
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	0.0891	F

The imputed values **must not** be used to compute the distances. In other words, all missing values for each feature are imputed based on the distances that use the dataset before the imputation. This ensures that the errors inherent in the imputed values are not propagated to compute the imputation.

Conditional hot deck imputation

Missing values for features that have missing values in a given object are imputed with the values for the same features copied from another, most similar object that satisfies a condition defined by the class feature. For instance, a missing value for an object 1 for which class value = F is imputed based on similarity only to the objects for which class value = F. The calculation of the similarities follows the unconditional hot deck imputation.

Example

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	**?**	**F**
Object 2	0.41139	0.3014	?	C
Object 3	0.24752	0.32148	**0.11169**	**F**
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	**0.0891**	**F**

To impute the missing value of feature AAcomp_C for object 1, you first compute the two distances to the object that share the same value of the class feature

$d(obj1, obj3) = \sqrt{(0.40256 - 0.24752)^2 + (0.1497 - 0.32148)^2 + 1} = 1.0264$

$d(obj1, obj5) = \sqrt{1 + (0.1497 - 0.58306)^2 + 1} = 1.4791$

The missing value is imputed with the value for the same feature AAcomp_C from the closest object 3 as follows.

	AAcomp_A	AAcomp_R	AAcomp_C	class
Object 1	0.40256	0.1497	**0.11169**	**F**
Object 2	0.41139	0.3014	?	C
Object 3	0.24752	0.32148	**0.11169**	**F**
Object 4	0.24609	?	0.13986	C
Object 5	?	0.58306	**0.0891**	**F**

Like for the unconditional hit deck imputation, the imputed values **must not** be used to compute the distances.

Calculation of the imputation error

You will use the two datasets that were imputed with the four methods to calculate the corresponding eight imputation errors. You will compute the errors based on the Mean Absolute Error (MAE) between the imputed values and the corresponding complete values that are available in the *assignment2_dataset_complete.csv* file. This dataset should be used only to calculate MAE values, not to perform the imputations. The MAE values should be used to judge and compare the quality of each imputation.

Given the imputed values $\mathbf{x} = \{x_1, x_2, \dots x_i, \dots, x_N\}$ computed from a dataset that has missing values and the corresponding complete values $\mathbf{t} = \{t_1, t_2, \dots, t_i, \dots, t_N\}$ in the complete dataset, MAE is defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |x_i - t_i|$$

where N is the total number of missing values, x_i is a the imputed value in the dataset that has missing values, x_i and t_i are values for the same object and same feature in the two datasets, and $|\cdot|$ denotes the absolute value.

Example

Incomplete dataset

	AAcomp_A	AAcomp_R	AAcomp_D	AAcomp_C	class
Object 1	0.40256	0.1497	0.1687	?	F
Object 2	0.41139	0.3014	0.47033	?	C
Object 3	0.24752	0.32148	0.41167	0.11169	F
Object 4	0.24609	?	?	0.13986	C
Object 5	?	0.58306	0.52568	0.0891	F

Dataset where values were imputed using the unconditional mean imputation

	AAcomp_A	AAcomp_R	AAcomp_D	AAcomp_C	class
Object 1	0.40256	0.1497	0.1687	0.11355	F
Object 2	0.41139	0.3014	0.47033	0.11355	C
Object 3	0.24752	0.32148	0.41167	0.11169	F
Object 4	0.24609	0.33891	0.394095	0.13986	C
Object 5	0.32689	0.58306	0.52568	0.0891	F

Complete dataset

	AAcomp_A	AAcomp_R	AAcomp_D	AAcomp_C	class
Object 1	0.40256	0.1497	0.1687	0	F
Object 2	0.41139	0.3014	0.47033	0.14175	C
Object 3	0.24752	0.32148	0.41167	0.11169	F
Object 4	0.24609	0.21359	0.24071	0.13986	C
Object 5	0.70541	0.58306	0.52568	0.0891	F

Given the above imputation, the MAE is calculated as follows.

$$MAE = \frac{1}{5}(\,|0.11355 - 0| + |0.11355 - 0.14175| + |0.33891 - 0.21359| + |0.394095 - 0.24071| + |0.32689 - 0.70541|\,) = 0.1598$$

The MAE values should be computed with precision of **four digits** after the decimal point.

Implementation

Your code must perform imputation, display the eight values of MAE on the screen and save the eight imputed datasets in the csv format. The imputed datasets should be named as follows:
Vnumber_a2_missing004_imputed_mean.csv
Vnumber_a2_missing004_imputed_mean_conditional.csv
Vnumber_a2_missing004_imputed_hd.csv
Vnumber_a2_missing004_imputed_hd_conditional.csv
Vnumber_a2_ missing20_imputed_mean.csv
Vnumber_a2_ missing20_imputed_mean_conditional.csv
Vnumber_a2_ missing20_imputed_hd.csv
Vnumber_a2_ missing20_imputed_hd_conditional.csv
where *Vnumber* is your V number, e.g., *V12345678_a2_missing004_imputed_mean.csv*

The MAE values should be displayed on the screen in the following format

MAE_004_mean = 0.1234
MAE_004_mean_conditional = 0.5678
MAE_004_hd = 0.1234
MAE_004_hd_conditional = 0.5678
MAE_20_mean = 0.1234
MAE_20_mean_conditional = 0.5678
MAE_20_hd = 0.1234
MAE_20_hd_conditional = 0.5678

You must use Java to implement all computations including loading the datasets from the csv files, coding the four imputation methods, calculation of the MAE values, printing the MAE values on the screen, and saving of the eight imputed datasets. You may use multiple classes and functions, but they must be included in **a single source code (.java) file**. This java file must successfully compile and produce the above mentioned outputs. Use appropriate data types to ensure the required precision of results.

Deliverables

1. Java source code in a single .java file
2. A pdf document with answers to the following five questions

2a. What are the MAE values for the eight results? You should simply copy the output from the screen.

2b. Which of the considered four imputation methods is the most accurate for the *assignment2_dataset_missing004.csv* dataset? Explain **why**.

2c. Which of the considered four imputation methods is the most accurate for the *assignment2_dataset_missing20.csv* dataset? Explain **why**.

2d. Are the results for the two datasets similar? Explain **why**.

2e. Which of the two unconditional methods (mean vs. hot deck) is faster, i.e., requires fewer computations?

Notes

- Include your **name, class number and title, and assignment number** at the top of the pdf file. Use separate and **clearly marked paragraphs** for each of the five questions.
- Copy the submission email to yourself to have a proof for the time of your submission.
- Do not procrastinate and start early. Late submissions will be subject to deductions: 15% in first 12 hours and 30% for between 12 and 48 hours. We will not accept submissions that are over 48 hours late.
- TA will grade the assignments by checking if the source code runs correctly, validating the results on the screen and in the files, and marking the answers to the five questions.
- We will **deduct** points if the files names and/or the outputs on the screen do not follow the above defined format.
- We will check against **plagiarism**. Make sure that you write your own code and provide your own answers.

Due Date

Your assignment must be received by 12:30pm Eastern Time (beginning of the lecture) on September 26 (Tuesday), 2017. Send the two files in a single email to the TA for this class, Mr. Chen Wang, at wangc27@vcu.edu