**Project 1 – Data Preprocessing**

**CS539 Machine Learning – Fall 2014**

**Prof. Carolina Ruiz**

**Student’s Name:** Daniel Fitzgerald

**I. Data Exploration**

**Three observations of good things about the dataset (include visualizations) [6 Points]:**

* **The data has many attributes.** There are 24 dimensions plus 24 features for medication. In general, the larger the dataset, the more certain we can be that any trends we observe in it are accurate predictive models.
* **The dataset is large.** There are 82599 entries, collected over 10 years. This was a very large study. In addition, there is a large variation of data. Almost every combination of attributes appears. This is good because if some combination doesn’t appear, we can know its probability is very low, but we can’t know how low beyond some point. If it appears at-least once, then we can estimate its probability as we do all other combinations.
* **The data contains many nominal attributes.** Nominal attributes are good for making predictive models because they lend themselves to binary classification algorithms such as decision trees. Numerical values must be discretized for such categorical methods to apply.

**Three observations of bad things about the dataset (include visualizations) [6 Points]:**

* **The data is sparse.** Many of the values are missing and it was collected from disparate unstandardized sources. Some of the numeric and nominal values are based on somewhat arbitrary thresholds determined by the scientists. For example, the 30 day readmission limit was “was chosen based on criteria often used by funding agencies.” 30 days has no special significance in the context of diabetes diagnosis, testing, or treatment.
* **The data is noisy.** As with all real-world data, there are not only missing values, but there are likely mistakes or random variations in some of the data. For example, “race” is not a clear-cut category. For example, some patients could be Asian-Hispanic and they would have had to choose one of those races or the “other” option. Some doctors may have multiple medical specialties, or experience in areas outside their formal specialty.
* **The data contains irrelevant features.** Payer code, for example, has no bearing on any medical procedures, diagnosis, etc. This is not a problem in practice because irrelevant data can easily be identified and ignored by a human interpreter (data scientist)

**List of all attributes that you would remove from the dataset right away and why [3 Points]**

* **Weight:** “considered to be too sparse and it was not included in further analysis” (Strack *et al.*)
* **Payer Code**: “had a high percentage of missing values and it was not considered relevant to the outcome” (Strack *et al.*)
* **Patient Number**: Like payer code, these are arbitrary numbers that have no relationship to medical conditions, diagnoses, treatment, etc.
* **Encounter ID**: Like payer code, these are arbitrary numbers that have no relationship to medical conditions, diagnoses, treatment, etc.

**II.1. Data Preprocessing: Discretization AT MOST 1 PAGE**

1. **Using Weka:**
2. **Description of the supervised discretization results [2 Points]**

**Description of the Java code implementing the supervised discretization filter [5 Points]**

1. **Description of the unsupervised discretization results [2 Points]**

**Description of the Java code implementing the unsupervised discretization filter [3 Points]**

1. **Using R:**
2. **Description of the R functions used for discretization [5 Points]**
3. **Description of the results obtained with these discretization functions [3 Points]**

**II. 2. Data Preprocessing: Missing Values. AT MOST 1 PAGE**

1. **Using Weka: Description of the results of replacing missing values [3 Points]**

1. **Using R:**
2. **Description of the R functions used for replacing missing values [4 Points]**
3. **Description of the results obtained with these functions [3 Points]**

**II.3 Data Preprocessing: Attribute/Feature Selection AT MOST 2 PAGES FOR II.3 AND II.4 COMBINED**

1. **Using R.**
2. **Include either the Correlation Matrix, a visualization of it, or both. [4 Points]**
3. **Which 3 attributes would you remove based on the correlation matrix and why? [1 Points]**
4. **Using Weka. Correlation Based Feature Selection.**
   1. **Result of applying CfsSubsetEval. [1 Points]**

**Explain in your own words what property this subset of attributes satisfies. [3 Points]**

* 1. **Comparison of results with your answers to part (a) above. [1 Points]**

1. **Using R. Description of Feature Selection functions in R [6 Points]**

**and results of using them on the dataset [4 Points] (may continue on next page)**

**II.4. Data Preprocessing: Attribute/Feature Extraction. AT MOST 2 PAGES FOR II.3 AND II.4 COMBINED**

1. **Using Weka. Principal Components Analysis (PCA) Results and Discussion [4 Points]**
2. **Using R. Principal Components Analysis (PCA) Results and Discussion [4 Points]**

**Specify what functions you used in R. [4 Points]**

1. **Comparison of the Weka and the R results [3 Points]**

**III. Model Construction. AT MOST 1 PAGE**

Summarize the experiments you ran in the table below. Add more table rows as needed. [24 Points]

* **What code/functions did you used to run ZeroR experiments in R? [1 Points]**
* **What code/functions did you used to run OneR experiments in R? [5 Points]**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tool** | **ML**  **technique** | **Pre-**  **processing** | **Testing**  **method** | **Resulting**  **model** | **Evaluation:**  **accuracy,**  **conf. matrix** | **Observations and Analysis of your results** |
| Weka?  or  R? | ZeroR?  or  OneR? | none? |  |  |  |  |
| … |  |  |  |  |  |  |
| … |  |  |  |  |  |  |
| … |  |  |  |  |  |  |
| … |  |  |  |  |  |  |
| … |  |  |  |  |  |  |

**Optional additional page to include any interesting work that you did in the project that you want to show, but didn’t have enough space to include on the previous pages. AT MOST 1 PAGE**