**Project 1 – Data Preprocessing**

**CS539 Machine Learning – Fall 2014**

**Prof. Carolina Ruiz**

**Student’s Name:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |
| --- | --- | --- |
|  | Weka | R |
| **Data Exploration [15 points]** | **/ 15** | |
| **Discretization [20 points]**  Code (Weka) / function (R) description  Results & analysis | unsuperv. supervised  /03 /05  /02 /02 | /05  /03 |
| **Missing Values [10 points]**  Code (Weka) / function (R) description  Results & analysis | /03 | /04  /03 |
| **Feature Selection [20 points]**  Correlation matrix  Correlation based feature selection (CFS) in Weka  Feature selection in R (functionality & experiments) | /05 | /05  /06 /04 |
| **Feature Extraction [15 points]**  Comparison of PCA results in Weka and R | /04 | /04 /04  /03 |
| **Model Construction [30 points]**   * Code (Weka) / function (R) description * Experimentation with different preprocessing * Experimentation with different testing modes * Presentation of resulting model and evaluation * Observations and analysis of results | ZeroR OneR  /01 /01  /01 /01  /01 /01  /01 /01 | ZeroR OneR  /01 /05  /02 /02  /02 /02  /02 /02  /02 /02 |
| TOTALS | **/32** | **/63** |

Total Project: \_\_\_\_\_\_\_\_\_\_\_\_/110 = \_\_\_\_\_\_\_\_\_\_\_/90

Class Presentation / Slides: \_\_\_\_\_\_\_\_\_\_\_/05

Class Participation during project presentation: \_\_\_\_\_\_\_\_\_\_\_/05

Total Project 1: \_\_\_\_\_\_\_\_\_\_\_/100

**I. Data Exploration AT MOST 1 PAGE**

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

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

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

**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**