**Project 4 – Clustering**

**CS548 Knowledge Discovery and Data Mining - Spring 2016**

**Prof. Carolina Ruiz**

**Students:** <replace this with your names in alphabetical order by last name>

|  |  |  |
| --- | --- | --- |
| **Dataset :**   * Dataset Description * Data Exploration * Initial Data Preprocessing (if any) | **Dataset** | |
| /05  /10  /05 | |
| **Code Description:** At leasttwo Clustering algorithms | **Weka**  /20 | **Python**  /10 |
| **Experiments:**   * Guiding Questions | /10 | |
| K-means - Sufficient & coherent set of experiments | /05 | /05 |
| * Objectives, Parameters, Additional Pre/Post-processing | /05 | /05 |
| * Presentation of results | /05 | /05 |
| * Analysis of individual experiments’ results | /05 | /05 |
| Hierarchical - Sufficient & coherent set of experiments | /05 | /05 |
| * Objectives, Parameters, Additional Pre/Post-processing | /05 | /05 |
| * Presentation of results | /05 | /05 |
| * Analysis of individual experiments’ results | /05 | /05 |
| DBSCAN - Sufficient & coherent set of experiments | N/A | /05 |
| * Objectives, Parameters, Additional Pre/Post-processing | N/A | /05 |
| * Presentation of results | N/A | /05 |
| * Analysis of individual experiments’ results | N/A | /05 |
| Quantitative Analysis of Results and Discussion | /10 | |
| Qualitative Analysis of Results, Discussion, and Visualizations | /20 | |
| Advanced Topic | /30 | |
| Total Written Report Project 4 | /220 = /100 | |

**Dataset Description, Exploration, and Initial Preprocessing: (at most 1 page)**

**[05 points] Dataset Description: (e.g., dataset domain, number of instances, number of attributes, distribution of target attribute, % missing values, …)**

**[10 points] Data Exploration: (e.g., comments on interesting or salient aspects of the dataset, visualizations, correlation, issues with the data, …)**

**[05 points] Initial data preprocessing, if any, based on data exploration findings: (e.g., removing IDs, strings, necessary dimensionality reduction, …)**

**Weka Code Description: Inputs, output, and process followed by Weka’s code for clustering (at most 2/3 page)**

**[10 points] Code Description for the first algorithm of your choice:**

**[10 points] Code Description for the second algorithm of your choice:**

**[10 points] Python Packages and Functions used for Clustering. Describe inputs & outputs (at most 1/3 page)**

**[10 points] Three Guiding Questions about the dataset domain (at most 1/3 page):**

1. **…**
2. **…**
3. **…**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Experiments with Partitional Clustering (k-means).** *At most 1 page.* | | | | | | | | | |
|  | **Tool** | **Pre-process** | **# clusters** | **Distance**  **function** | **#**  **iterations** | **SSE** | **% of instances**  **per cluster** | **Observations about experiment**  **Observations about visualization**  **Interpretation of centroids**  **Classes to cluster evaluation?** | **You can add**  **other columns** |
| P1 | Weka?  Python? |  |  |  |  |  |  |  |  |
| P2 | … |  |  |  |  |  |  |  |  |
| P3 | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Experiments with Hierarchical Clustering (single link, complete link, average, centroid, Ward).** *At most 1 page.* | | | | | | | | | |
|  | **Tool** | **Pre-process** | **# clusters** | **Link**  **type** | **#**  **iterations** | **Time**  **taken** | **% of instances**  **per cluster** | **Observations about experiment**  **Observations about visualization**  **Classes to cluster evaluation?** | **You can add**  **other columns** |
| H1 | Weka?  Python? |  |  |  |  |  |  |  |  |
| H2 | … |  |  |  |  |  |  |  |  |
| H3 | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[40 points] Summary of Experiments with DBSCAN in Python.** *At most 2/3 page.* | | | | | | | | | |
|  | **Pre-process** | **Epsilon** | **minPts** | **#**  **clusters** | **Time taken** | **% of instances**  **per cluster** | **Observations about experiment**  **Observations about visualization**  **Interpretation of means & std dev**  **Classes to cluster evaluation?** | **You can add**  **columns** |
| D1 |  |  |  |  |  |  |  |  |
| D2 |  |  |  |  |  |  |  |  |
| D3 |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |

**[10 points] Quantitative Analysis of Weka and Python Results and Discussion (at most 1/3 page)**

**[20 points] Qualitative Analysis of Weka and Python Results on and Visualizations (at most 1 page)**

(Remember also to analyze the results from the point of view of the dataset domain, and discuss the answers that the experiments provided to your guiding questions.)

**Advanced Topic: <include name of the topic here>**

**[7 points] List of sources/books/papers used for this topic (include URLs if available):**

* …
* …
* …

...

**[20 points] In your own words, provide an in-depth, yet concise, description of your chosen topic. Make sure to cover all relevant data mining aspects of your topic.**

**[3 points] How does this topic relate to clustering?**

**Authorship:** Although each student on the team is expected to be involved in every aspect of the project, describe in detail here the main contributions that each of the team members made to this project. This authorship description must accurately reflect the work done by each team member, and must be approved by all of the members of the team (at most 1/3 page)