**CS548 Knowledge Discovery and Data Mining. Spring 2014. Project 4: Clustering. Student’s Name: *write your name here***

**[2 points] Dataset Description, and initial data preprocessing if any (at most ½ page):**

**[3 points] Three Guiding Questions about the dataset domain (at most ½ page):**

1. **…**
2. **…**
3. **…**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[5 points] Summary of Experiments with Partitional Clustering (k-means).** *At most 1/2 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?  RapidM?  Matlab? |  |  |  |  |  |  |  |  |
| P2 | … |  |  |  |  |  |  |  |  |
| P3 | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[5 points] Summary of Experiments with EM.** *At most 1/2 page.* | | | | | | | | | |
|  | **Tool** | **Pre-process** | **# clusters** | **Distance**  **function** | **#**  **iterations** | **Log likelihood** | **% of instances**  **per cluster** | **Observations about experiment**  **Observations about visualization**  **Interpretation of means & std dev**  **Classes to cluster evaluation?** | **You can add**  **columns** |
| E1 | Weka?  RapidM?  Matlab? |  |  |  |  |  |  |  |  |
| E2 | … |  |  |  |  |  |  |  |  |
| E3 | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **[6 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?  RapidM?  Matlab? |  |  |  |  |  |  |  |  |
| H2 | … |  |  |  |  |  |  |  |  |
| H3 | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |
| … | … |  |  |  |  |  |  |  |  |

**[7 points] Analysis of Results:** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the Domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include and explain (some of) the best / most interesting results you obtained in your experiments. 4. Include a visualization of the best k-means clustering, the best hierarchical clustering, and the best EM clustering you obtained.

**[7 points] Advanced Topic: <include name of the topic here>**

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

* …
* …
* …

...

**Description of the topic and summary of what you learned:**

**How does this topic relate to clustering and the material covered in this course?**