# waDr. Eick

Assignment2 COSC 4335 Spring 2018

Similarity Assessment, Clustering and Using Clustering to Create Background Knowledge for Classification Tasks

Individual Project[[1]](#footnote-1)

Second Draft



Due: Saturday October 20, 11p (3% bonus), submissions will still be accepted without penalty until Monday, October 22, 11p (electronic Submission)

Last Updated: October 6, 7:01pm

Tentative weight of Assignment2: about 32-37% of the points allocated to the four assignments

**Learning Objectives**:

1. Learn to use popular clustering algorithms, namely K-means, K-medoids/PAM and DBSCAN
2. Learn how to summarize and interpret clustering results
3. Learn to write R functions which operate on the top of clustering algorithms and clustering results
4. Learning how to make sense of unsupervised data mining results
5. Learn how clustering can be used to create useful background knowledge and classification problems.
6. Learn how to create distance function and distance matrices in R and learn about their importance for clustering.

**Datasets**: In this project we will use the *Diamond9 dataset[[2]](#footnote-2) which is a 2D dataset with ; it can be downloaded at:* [*http://www2.cs.uh.edu/~ml\_kdd/restored*](http://www2.cs.uh.edu/~ml_kdd/restored) *just click on the Dataset link, and download the file called Diamond9.data)* and the cleaned Pima Indian dataset (the original can be found at: [https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes](https://archive.ics.uci.edu/ml/datasets/pima%2Bindians%2Bdiabetes))which is a “null-value cleaned” version of the original dataset, called Pima dataset, in the following which can be found at: <http://www2.cs.uh.edu/~ceick/UDM/Pima.csv>

The Pima dataset has 8 numerical attributes and a binary class variable (1 indicates that the person is assumed to have diabetes), indicating the following information:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

5 Examples in the modified Pima Indians Diabetes Dataset:

6,148,72,35,156,33.6,0.627,50,1

1,85,66,29,156,26.6,0.351,31,0

8,183,64,29,156,23.3,0.672,32,1

1,89,66,23,94,28.1,0.167,21,0

0,137,40,35,168,43.1,2.288,33,1

**Assignment2 Tasks**:

**0.** Transform the Pima dataset into a dataset ZPima by z-scoring the first 8 attributes of the dataset, and copying the 9th attribute of the dataset \*

**1.** Write an R-function purity(a,b,outliers=FALSE) that computes the purity of a clustering result based on an apriori given set of class labels, where a gives the assignment of objects in O to clusters, and b is the “ground truth”. Purity is defined as follows: Let

O be a dataset

X={C1,…,Ck} be a clustering of O with Ci ⊆O (for i=1,…,k), C1∪…∪Ck ⊆O and Ci∩Cj=∅ (for i≠ j)

PUR(X)= (number\_of\_majority\_class\_examples(X)/(total\_number\_examples\_in\_clusters(X))

If the used clustering algorithm supports outliers, outliers should be ignored in purity computations; you can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters. If the parameter outliers is set to FALSE, the function just returns a floting point number of the observed purity, if parameter outliers is set to T the function returns a vector: (<purity>,<percentage\_of\_outliers); e.g. if the function returns (0.98, 0.2) this would indicate that the purity is 98%, but 20% of the objects in dataset O have been classified as outliers. \*\*\*\*

**2.** Write an R-function wabs-dist(u,v,w) that takes two vectors u, v as an input and computes the distance between u and v as the sum of the absolute weighted differences—using the weights in w[[3]](#footnote-3).\*

**3** Write an R-function create-dm(x,w) that returns a distance matrix for the objects in dataframe x by calling the wabs-dist(a,b,w) for all pairs of objects a and b belonging to x. Next, call create-dm for the first 8 attributes of the dataset Zpima and save the obtained distance matrix, we call Zpima-dist. \*\*\*

**4**. Learn a linear model that predicts the 9th attribute using the first 8 attributes for the ZPima dataset. Interpret the obtained coefficents of the obtained linear model and access its quality and the importance of the 8 attributes. \*\*\*

**5**. Run K-means on the ZPima dataset for k=5 and k=8 and nstart=20; next run PAM for k=5 with distance matrices that have been created using the following weight vectors for the ZPima dataset─obtaining 3 different PAM clustering results:

a. (1,1,1,1,1,1,1,1)

b. (0.2,1,0,0,0,1,0.2.0.2)

c. (0,1,0,1,0,1,0,0)

For the 5 obtained clustering results report the overall purity, the purity of each cluster, the majority class and the centroid/medoid of each cluster. Next, visualize the clustering result of the K-means run for k=5 and for the 3 PAM results in the Plasma glucose/Body mass index (Attribute 2&6 Space) for the original dataset[[4]](#footnote-4). Interpret the obtained results! Does changing the distance metric affect the PAM results? Do the results tell you anything which attributes are important for diagnosing diabetes, and about the difficulty diagnosing diabetes? \*\*\*\*\*\*\*\*\*

**6.** Run DBSCAN on the ZPima dataset; try to find values for MinPoints and epsilon parameters, such that the number of outliers is less than 20% and the number of clusters is between 3 and 13; visualize the obtained clustering result in the Plasma glucose/Body mass index attribute Space on the original dataset and report its purity. Comment on the quality of the obtained clustering result. \*\*\*

**7.** Run K-means with nstart=20 for k=9 and k=12 for the Diamond9 dataset; visualize the results, compute their purity, and discuss the obtained clustering results and its quality.\*\*

**8.** Next, run DBSCAN for the Diamond9 dataset; try to find “good” parameter settings[[5]](#footnote-5) for epsilon and Minpoints such that purity of the obtained clustering result is maximized and the number of outliers is less than 10%[[6]](#footnote-6). Visualize the “best” clustering result you found, report the epsilon and Minpoint parameters you used to obtain this clustering result and report its purity. Results with higher purity will receive higher scores; students that find the “best” result will get extra credit! \*\*\*\*(and up to \*\* extra credit for the “best”
 solution)**.**  Also compare the DBSCAN results with those you obtained for K-means in Task 7!

**9**. Summarize to which extend the K-Means and DBSCAN where able to rediscover the classes in the Diamond9 and Pima/ZPima datasets! Moreover, did your expermental results reveal anythign interesting about diabetes? \*\*\*

**Training Cases for R-Functions that need to be developed for tasks 1:**

*1st case*

Let a, b be the following vectors:

a=(0,1,1,1,1,2,2,3)

b=(A,A,A,E,E,D,D,C)

purity(a,b)=0.714

2nd case

a=(1,1,1,0,0,2,2,2)

b=(A,A,A,E,E,D,D,C))

purity(a,b)=0.8333

**Training Cases for R-Functions that need to be developed for tasks 2 and 3:**

Let x be a data frame containing vectors a, b and c where:

a: (1, 2)

b: (4, 5)

c: (9, 12)

w is weight vector: (0.2, 0.3)

wabs-dist(a,b,w) = 1.5

wabs-dist(b,c,w) = 3.1

create-dm(x,w)=

|  |  |  |
| --- | --- | --- |
| 0 | 1.5 | 4.6 |
| 1.5 | 0 | 3.1 |
| 4.6 | 3.1 | 0 |

**Note:** The \* at the end of each problem denotes its weight. Therefore, higher the number of \*, more points that problem will carry.

**Submission Guidelines:**

Create a folder and name it as LastName\_StudentId\_HW2. HW2 folder should include:

* R code for the tasks.
* The data files needed to run the R codes.
* The assignment report containing all the plots and results along with the interpretations for them.

Submit the LastName\_StudentId\_HW2 folder in a zipped file through Blackboard.

**Points will be deducted for incomplete submission**

1. No collaboration with your class mates is allowed! [↑](#footnote-ref-1)
2. It can be found at: [http://www2.cs.uh.edu/~ml\_kdd/restored/Complex&Diamond/diamond.data](http://www2.cs.uh.edu/~ml_kdd/restored/Complex%26Diamond/diamond.data)

it has been visualized at: [http://www2.cs.uh.edu/~ml\_kdd/restored/Complex&Diamond/2DData.htm](http://www2.cs.uh.edu/~ml_kdd/restored/Complex%26Diamond/2DData.htm) [↑](#footnote-ref-2)
3. It basically, computes sum((abs(a-b))\*w) [↑](#footnote-ref-3)
4. Add the $cluster column to the Pima dataset (and not the ZPima dataset) and visualize the clusters. [↑](#footnote-ref-4)
5. You might write a program that runs DBSCAN automatically for different parameter settings; you might also change the distance function, using the same procedure we used for PAM, to obtain “better” results. [↑](#footnote-ref-5)
6. That is, if you report a clustering result with more than 10% outliers, you will get a low score for this task! [↑](#footnote-ref-6)