# Dr. Eick

Third Draft

Assignment2 COSC 4335 Spring 2017

Clustering with K-Means and DBSCAN , Making Sense of Clustering Results

and Prediction

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



**Learning Objectives**:

1. Learn to use popular clustering algorithms, namely K-means 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 to use background knowledge to guide data mining algorithms to obtain better results.
6. Learn how clustering can be used to create useful background knowledge for prediction and classification problems.
7. You will learn how to create prediction models using R

**Deadlines**: March 19, 11p (students receive a 5% early submission bonus!); submissions will still accepted until March 24, 11p; the second deadline is a hard deadline!

**Last Updated**: March 4, 10a

**Datasets**: In this project we will use the *Complex9\_GN16 dataset[[2]](#footnote-2)* and the H*Abalone dataset* which is a modification of the Abalone Dataset (<http://archive.ics.uci.edu/ml/datasets/Abalone> ). The abalone shell and the meat is of value.[[3]](#footnote-3) The Complex9\_GN16 dataset is a 2D dataset with 9 classes[[4]](#footnote-4) and Abalone is an 8D dataset and one numerical output attribute—however, we use a transformed version of this dataset called *HAbalone* which has 10 attributes, including one numerical output attribute and one ordinal class attribute (5 classes are in the dataset: A, B, C, D, and E); the last attribute of each dataset serves as the class attibute which should be ignored when clustering the data sets; the 9th attibute of the HAbalone dataset should be ignored as well—however, the class attribute as well the numerical 9th attribute of the HAbalone dataset will be used in the post analysis of the clusters that have been generated by K-means and DBSCAN.

**Project2 Tasks**:

**0.** The orginal Abalone dataset has the following attibutes:

 Name Data Type Meas. Description

 ---- --------- ----- -----------

 Sex nominal M, F, and I (infant)

 Length continuous mm Longest shell measurement

 Diameter continuous mm perpendicular to length

 Height continuous mm with meat in shell

 Whole weight continuous grams whole abalone

 Shucked weight continuous grams weight of meat

 Viscera weight continuous grams gut weight (after bleeding)

 Shell weight continuous grams after being dried

 Rings integer +1.5 gives the age in years[[5]](#footnote-5)

Transform the Abalone dataset into a new 10D dataset called *HAbalone* as follows:

1. For the first attribute replace its values as follows: M🡪2, F🡪1, I🡪
2. Normalize the second through eighth attribute into z-scores
3. Keep the ninth attribute ‘Rings’ as it is
4. Introduce a new ordinal attribute called *Class* (attribute 10) based on the value of the 9th attribute ‘Rings’ as follows: 0-5🡪A, 6-8🡪B, 9-11🡪C, 12-17🡪D, 18-…🡪E.

Remark: When clustering the dataset only the first 8 attributes will be used; attributes 9 and 10 will be used to evaluate the quality of a clustering result. \*\*

**1.** Write an R-function purity(a,b) that computes the purity and the pecentage of outliers of a clustering result based on an apriori given set of class lables, 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; then

X={C1,…,Ck} is 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))

You can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters. The purity 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 ordinal-variation (B) that computes the original agreement of a bag B of ordinal classes associated with the instances of a cluster—the orignial classes are named A, B, C, D, and E in the H-ABALONE dataset. It is defines as follows:

Let φ be definied as follows: φ(A)=4, φ(B)=3, φ(C)=2, φ(D)=1, φ(E)=0,

If o is an object in the Abalone dataset, o.class denotes the value of the tenth attribute of o (which takes values A, B, C, D or E)

Let C be a cluster of Abalone objects, then the ordinal agreement in C[[6]](#footnote-6) is defined as follows:

Ordinal-variation(C) = (Σc,c’∈C and c≠c’ |φ(c)φ(c’)|)/|C|\*\*2-|C|)

If |C|=1 Ordinal\_variation(C)=0

In the above formulas ”|…|” represents the set cadinality functions

Moreover, assuming X={C1,…,Ck} is a clustering consisting of classes C1,…,Ck,

Ordinal-variation(X) is the number of instances weighted sum of Ordinal-variation(C1),…,Ordinal-variation(Ck). However, we give X in the form of (a,b) where *a* gives the assignment of objects in O to clusters, and *b* is class variable assoicated with each with each object in O. That is, implement a function ordinal-variation(a,b) instead; ignore all instances of cluster 0 from ordinal agreement computations, as those example represent outliers. \*\*\*

**3**. Write an R-function variance(a,b) which computes the variance of the clustering result X based on an apriori given set of numerical observations—one numerical observation is associated with with each object, where *a* gives the assignment of objects in O to clusters, and *b* is the numerical observation associated with each object in O. The *variance of a clustering* is the weighted sum of the variance[[7]](#footnote-7) observed in each cluster with respect to the numerical variable. The observed cluster variance is weighted by number\_of\_example\_in\_the cluster/total number of examples in all clusters; the same way how variance is assessed by regression tree learning algorithms.

In general, the function variance returns a vector: (<variance>,<percentage\_of\_outliers). If the used clustering algorithm supports outliers, outliers should be ignored in variance computations; you can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters. For example if the function variance returns (2.8, 0.3) this would indicate that the variance of the evaluated clustering is 2.8 and that 30% of the objects in the clustered dataset are outliers. If cluster 0 does not exist, assume that there are no outliers! \*\*

**4**. Run K-means for k=9 and k=13 twice for the Complex9-GN16 dataset[[8]](#footnote-8). Visualize and interpret the obtained four clusterings! Also compute the purity of the clustering results using the function you developed earlier. Interpret the clustering result! \*\*

**5.** Run K-means for k=5 and k=10 for the HAbalone dataset 20 times (set seed 4335, before running k-means), reporting the result with the lowest SSE. Report the best clustering found, its purity, ordinal ageement and variance (using the ninth and tenth attribute). \*\*

**6.** Run DBSCAN for the Complex9-GN16 data set trying to find a clustering with the highest purity (try to find good parameters by manual trial and error) with 20% or less outliers. Report the best clustering you obtained including its purity and how you found it! Do the same for the HAbalone dataset minimizing the orginal variation of the obtained clustering result! Report the obtained 2 clustering results including the purity of the first results and ordinal-variation of the second result!\*\*\*\*

**7.** Write a search procedure in R that looks for the “best” K-means clustering for the HAbalone dataset—trying to minimize the variance of the 9th attribute—assuming k=6 by exploring different distance metrics for the HAbalone dataset. Distance metrics are modified by multiplying the HAbalone with weight vectors (a1,…,a8) with each weight being a number in [0,∞) and then running K-means[[9]](#footnote-9) for the transformed dataset. The developed search procedure returns the “best” K-Means clustering found—the one for which the variance is the lowest[[10]](#footnote-10)—, the weight vector used to obtain this result and the accomplished variance as well each cluster’s size and variance, and the seed used when running k-means; please limit the number of tested weight vectors to 5000 in your implementation! Report the best clustering you found using this procedure; if you run a probabilistic search procedure report 3 clustering results for 3 runs of your search procedure. Also report the purity of the best clustering(s) you found! What does this result/these results tell you about the importance of the 8 attributes for predicting the number of rings of an abalone? Explain how the search procedure you deleloped works! \*\*\*\*\*\* (and up to \*\*\*\* extra credit for more sophisticated search procedures and other sophisticated approaches to solve the problem at hand).

There will be a ***COSC 4335 Abalone Data Mining Cup*** associated with task 6; the student who finds the clustering with the lowest variance will win a prize (*TBDL what it will be; may be an Abalone Dinner in Japan, if we find a budget for that…)* and there will also be a second place prize. To be eligible for the competition submit the following to Romita, in a separate e-mail, before the submission deadline:

1. Weight vectors for attributes you used
2. Seed you used when running k-means to obtain the clustering result[[11]](#footnote-11)
3. Variance achieved

Also save the modified Abalone dataset with the clustering result attached as an additional attribute called ‘Cluster’, just in case; you do not need to submit this file.

**8**. Learn a linear model that predicts the 9th attribute using the first 8 attributes for the HAbalone dataset. Interpret the obtained coefficents of the obtained linear model and access its quality of the obtained regression function and the importance of the 8 attributes. Compare this task’s finding with the findings of the previous task! Next, learn a different prediction model of your own liking[[12]](#footnote-12) for the same task Report the mean-square error and its R2 for the two models you obtained!\*\*\*\*

**9**. Summarize to which extend the K-Means and DBSCAN where able to rediscover the classes in the COMPLEX9-GN16 and HAbalone dataset! \*\*

**Training Cases for R-Functions that need to be developed for tasks 1-3:**

*1st case*

Let a, b, c be the following vectors:

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

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

c=(3,2,4,19,20,13,15,10)

purity(a,b)=0.714
ordinal-variation(a,b)=1.523
variance(a,c)=52.902

2nd case

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

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

c=(3,2,4,19,20,13,15,10)

purity(a,b)=0.8333
ordinal-variation(a,b)=0.3333
variance(a,c)=3.665

**Deliverables for Assignment2:**

1. A Report[[13]](#footnote-13) which contains all deliverables for all tasks of Project2.
2. An Appendix which describes how to run the procedure that you developed for Task 7.
3. An Appendix which contains the R-functions you wrote for tasks 0,1, 2, 3, 7, 8 should be included.
4. All R codes to be submitted in a compressed folder along with a readme file if necessary.
5. Delivery of Project2 Reports: use blackboard to submit your assignment and call the attached files *<last name>*\_P2.docx (or *<last name>*\_P2\_.pdf ) and <*lastname*>\_P2.zip/rar/7z
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/Complex9\_GN16.txt](http://www2.cs.uh.edu/~ml_kdd/restored/Complex%26Diamond/Complex9_GN16.txt) 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. <http://www.opentable.com/north-beach-restaurant> [↑](#footnote-ref-3)
4. It has been obtained by modifying the original Complex9 dataset, by exposing 16% of its examples to Gaussian noise. [↑](#footnote-ref-4)
5. For details see: <http://superimportexportwallpaper.blogspot.com/2013/01/abalone-age.html> ‘ [↑](#footnote-ref-5)
6. Assume C is given a vector of class memberships, and the ordinal agreement adds up the values of a distance matrix and then divides it by the number of elements in the distance matrix, excluding the diagonal. [↑](#footnote-ref-6)
7. If a cluster contains only 1 object, its variance is defined to be 0. [↑](#footnote-ref-7)
8. It can be found at: [http://www2.cs.uh.edu/~ml\_kdd/Complex&Diamond/Complex9.data](http://www2.cs.uh.edu/~ml_kdd/Complex%26Diamond/Complex9.data) ; it has been visualized at: [http://www2.cs.uh.edu/~ml\_kdd/Complex&Diamond/2DData.htm](http://www2.cs.uh.edu/~ml_kdd/Complex%26Diamond/2DData.htm) [↑](#footnote-ref-8)
9. Run k-means as follows: kmeans(<dataset>,<number of clusters); do not use other parameters! [↑](#footnote-ref-9)
10. The variance of the 9th attribute is low in a clustering this would indicate that the clusters contain examples of abalones that have a similar age / number of rings. [↑](#footnote-ref-10)
11. We will need the seed to reproduce your clustering result. [↑](#footnote-ref-11)
12. We recommend to use SVM regression or regression trees to obtain the second prediction model! [↑](#footnote-ref-12)
13. Single-spaced; please use an 11-point or 12-point font! [↑](#footnote-ref-13)