

Knowledge-Based Systems in Bioinformatics

Lab 1 – Clustering

Teachers:

Doğa Yalova	doga.yalova.3905@student.uu.se
Matthew Redmayne	matthew.redmayne.0786@student.uu.se
Lukas Bleichner	lukas.bleichner.5753@student.uu.se
Girish Pulinkala	girish.pulinkala@icm.uu.se

Introduction

Iris setosa, *Iris versicolor* and *Iris virginica* are three closely related flowers. The dataset consists of the following measurements (features): sepal length, sepal width, petal length and petal width. You may read more about the dataset on the UCI Machine Learning Repository <http://archive.ics.uci.edu/ml/datasets/Iris>. The task is to use clustering methods to group the flowers. Ideally, we would get one group for each species. Read the sample script `clustering.r`, run the code and answer the questions.

Tasks

1. In this exercise you shall use the sample R code for clustering from `clustering.R` file.

The iris dataset exists in R workspace under `iris`.

- a. Start with installing the `cluster` package.
- b. How many flowers per species are included in the data?
- c. What is the output of `summary` function? What is average petal and sepal length for all flowers?
- d. What is the structure of object `iris$Species`?
- e. Choose two features for the analysis e.g., sepal length and petal length.

2. Cluster the iris dataset using k-means approach. Plot the results (mark plot as Fig.1).

- a. Describe briefly the algorithm for k-means clustering
- b. Plot results using function `clusplot`. Mark plot as Fig.1.
- c. Will the clusters be the same each time you run it?
- d. Are clusters separate or unite?

3. Cluster the iris dataset using hierarchical approach. Plot the results (mark plot as Fig.2a and Fig.2b).

- a. Describe briefly the algorithm for hierarchical clustering
- b. Plot results using function `clusplot`. Mark plot as Fig.2a.
- c. Will the clusters be the same each time you run it?
- d. Are clusters separate or unite?
- e. Plot dendrogram. Mark plot as Fig.2b.
- f. What is the main advantage of using dendrogram?
- g. Describe shortly main features of dendrogram.

4. Display both clustering methods on single plot with the species names as points (mark plot as Fig.3).

- a. Can you see wrongly clustered species? If yes, which species and for which method?

b. What do the yellow squares in Fig.3 correspond to?

5. Answer the general questions.

- a. What is the idea behind unsupervised learning?
- b. Give at least two pros and cons for *k-means* and hierarchical clustering.
- c. Give two examples of biological applications where clustering can be used.