Data Mining and Machine Learning in Bioinformatics

Dr. Holger Fröhlich, Dr. Martin Vogt

Due: Apr 29, 10:30 am (by the end of the lecture)

Exercise Series 1

General: Exercises are to be solved and submitted in fixed groups by at most 3 students. Every member of a group should contribute solving each task and thus be able to answer questions to each task. No late submissions are accepted. Copying solutions will automatically lead to a point reduction of at least 50%. N – 1 homework assignments and N – 2 programming tasks have to be submitted in total.
A group can gain additional bonus points, if it presents its solution for a particular task during the tutorials. Accordingly, each task has a defined number of points as well as bonus points.

The famous iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. It can directly be loaded into the R workspace via

```
data(iris)
```

a) Compute a 150 x 150 Euclidean distance matrix between flowers. You can use R-function *dist* for that purpose, if you want. (4 points + 1 bonus)
b) Given that distance matrix calculate for each flower its so-called *nearest neighbor* (i.e. the one, which is most similar to it). Then construct a data.frame, with 4 columns:

Flower	Species	Nearest.neighbor	NN.species

Print the corresponding table on the screen and write it into a .csv file. (4 points + 1 bonus)
c) Calculate the percentage of flowers of species X that has a nearest neighbor of species Y. The output should be a 3 x 3 matrix X (because there are 3 species). Print that matrix on the screen and write it to a .csv file. (4 points + 1 bonus)
d) Generate three grouped barplots (R function *barplot*) visualizing the values of one row of X each. (4 points + 1 bonus)

e) Generate histograms (R function *hist*) visualizing the distribution of values for each of the variables sepal length, sepal width, petal length and petal height. The histograms should be plotted separately for each of the 3 species classes. (4 points + 1 bonus)

f) Explain in your own words what a histogram visualizes and how you have to interpret the figure. (4 points + 1 bonus)