GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
INSTITUTE OF COMPUTER SCIENCE
Software Engineering for Distributed Systems
http://www.swe.informatik.uni-goettingen.de

Data Science and Big Data Analytics
WS 2016/2017
Dr. Steffen Herbold

Exercise 2 · Due at 2016-12-01

Statistical Tests and Clustering

Execute the following tasks with R[1]:

1. Create three data samples with normally distributed data:
   - S1: 100 values, $\mu = 0.0, \sigma = 1.0$
   - S2: 100 values, $\mu = 1.5, \sigma = 1.0$
   - S3: 10 values, $\mu = 1.5, \sigma = 1.0$
2. Plot the densities of S1, S2, and S3 separately.
3. Plot the densities of S1 and S2, as well as S1 and S3 together.
4. Interpret the above density plots. What do they indicate?
5. Perform a t-test between S1 and S2, as well as between S1 and S3. How significant is the difference between the samples?
6. Apply the kmeans algorithm to the columns `Petal.Width` and `Petal.Length` of the `iris` data set
   - Three times for $k = 2$
   - Three times for $k = 3$
   - Three times for $k = 4$
7. Visualize the results of each clustering (Hint: look at the R documentation to see how to do that). Do the clusters remain the same? Are the results as you would expect them to be?

[1] You can start RStudio by typing rstudio into the bash in the CIP pool.