Towards Data Mining – Exercise 6: noisy data, outliers, and signal saturation

To pass this exercise, you need to follow these guidelines and answers questions that are underlined in Moodle.

Most of the needed code is given to you (ex6_code.R).

You can get help on any of the commands in R by typing *?commandname*. You can refer to a variable in a data.frame with the dollar sign: *my.data.frame$my.variable*. You might find the first weeks exercise also helpful.

Part 1. Noisy data

1. Load the data *iris_noisy.csv*.
2. Install and load the package *“e1071”*.
3. Fit a naive Bayes model to the data and check the training accuracy (note: normally a separate data for training and testing would be used). In the formula “Species ~.” means that we want to model Species as a function of every other variable.
   a. *fit.noisy <- naiveBayes(as.factor(Species) ~., data = noisy)*
   b. *pred.noisy <- predict(fit.noisy, noisy)*
   c. *table(pred.noisy, noisy$Species)*
4. What is training accuracy (classification rate for the training data) (Q1)?
   How many versicolor flowers were correctly classified (Q2)?
   What is the total number of correctly classified instances in the noisy dataset (Q3)?
5. Repeat with the complete data set. It is in file *iris_complete.csv.* What is the training accuracy for the complete data (Q4)?
6. Check the summary of the noisy data and compare to the summary of the complete data. Are there any differences (Q5)?
7. Can you figure out what is causing the difference in classification results? Try analyzing the data by stratifying based on the class label (Species) and comparing to the stratified complete data set. You can use e.g. the *subset* function for stratification and *summary* and *boxplot* for analysis. What is this type of noise called (Q6)? What is the correction procedure of this type of noise called?

Part 2. Outlier detection

8. Take a subset of the noisy data from the previous part where Species is “setosa”.
9. Use *densityMclust* to cluster the data. Plot the data in the same way as in task 11 d) i in the last week’s exercise. What is the average Euclidean distance of the normal samples to the cluster center (Q7)? What is the Euclidean distance of the outlier to the cluster (Q8)?

10. What type of outlier is it (Q9)? What is the primary purpose of using clustering in the analysis of the Setosa subset (Q10)?

Part 3. Signal saturation

11. Load the data *saturated.csv*.

12. Plot the signal. Can you see signs of signal saturation?

13. What are the minimum and maximum values of the signal (variable Acc) (Q11-Q12)?

14. How many percent of the signal consists of the minimum values? How about maximum values (Q13)?

15. Use the *rle* command to find the runs of equal values in the signal.
    a. How long is the longest run? What value is repeated in that run?
    b. How long is the longest run of a value other than the saturation point? What value(s) are repeated in that/those run(s)?
    c. How many runs of at least four consecutive values are there in the signal (Q14)?