Lecturer: Jan Baumbach

Class: Introduction to Bioinformatics

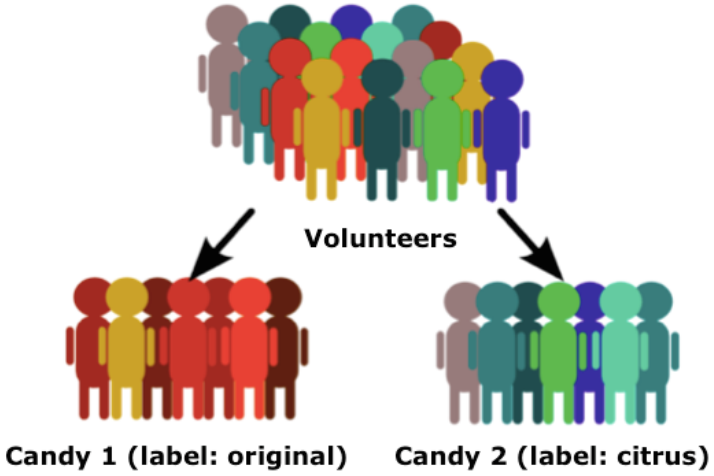
Student project – Breath Mining

In the class, we learned about data mining methods. These machine learners essentially train a classifier from a set of training data sets (where the class labels are known). Their robustness is usually evaluated with an x-fold cross-validation and they are afterwards applied to predict the unknown class of new samples.

Together, we went to the Odense University Hospital to collect breath metabolomics data to simulate a potential medical application. You were given the choice of two candies (Halls "original" and Halls "citrus") and asked to breath into a MCC-IMS BioScout machine to sample the expression of volatile metabolites in your exhaled air.

We provide you with raw data produced by the MCC-IMS BioScout. For each data set, we recorded which candy type you picked. This represents your training data.

The tree below illustrates their relationship.



A measurement is essentially represented by an intensity matrix I(d,t) for varying retention times (d) and drift times (t). An intensity peak at a certain position (d,t) corresponds to an exhaled molecule, i.e. a volatile organic compound (VOC). Due to slightly varying environmental conditions during sampling, the position of a peak that one substance will have can vary slightly from measurement to measurement. In other words: The same substance will occur as peak at approximately (!) the same position (d,t) in all measurements but its maximum will not be at exactly the same position.

TASKS:

Your first task is to implement a software that

Part I

(1) reads several raw data files,

(2) generates an density plot visualizing the intensity matrix (see Figure 1 for an example),

(3) calls the command line tool PEAX (URL2) to normalize the data and to extract a list of peaks, which represent volatile organic compounds (VOCs),

(4) implements a "peak alignment" (also known as "peak mapping") strategy (Figure 2). The underlying question is: Which of the peaks obtained in different measurements

correspond to the same VOC? Use this alignment method to assign unique IDs to peaks across different measurements. Keep in mind that the peaks positions may vary a bit. Feel free to contact the TAs to discuss your approach.

(5) reads the class labels file with the candy annotations (if you use R, these labels should be stored as a "factor", not as a "character vector"),

(6) creates a matrix M_{train}: *peaks x volunteers* that will serve as training data in part II (see Figure 3 for an illustration).

Note: For testing of Part I, we have prepared a toy data set consisting of six measurements with two classes (3x "original" and 3x "citrus" candy). You can download the raw data from URL1. You should then download & install the free software PEAX from URL2 and use it to create pick lists. These you can then compare to our peak picking results provided under URL3. The output of your peak alignment should look similar (but not necessarily identical!) to URL4 and your final indicator matrix should look like our example at URL5. Class labels for the toy data set can be found under URL6. To test your results, check if you find the same three top features as shown in Figure 4. As your peak IDs will be different from ours (Figure 4), refer to Table 1 to inquire the corresponding drift times and retention times.

It is also possible that you identify different peaks as discriminating features depending on the PEAX settings and peak alignment strategy you choose. Feel free to contact the TAs to discuss your results.

Part II

In the next step, extend your software to

(7) create a peak alignment matrix M_{train} using the data you collected at the OUH (provided in URL7), which will now serve as basis/input for a classifier (machine learning tool),

(8) train a Random Forest classifier,

(9) implement a 5-fold cross-validation,

(10) report the mean accuracy, sensitivity and specificity,

(11) extract the five most discriminating features (peaks) by using the Gini index,

(12) learn and plot/report a decision tree by using only these five best features/peaks.

We suggest that you use the statistical learning suite R that you have learned about in the class. The TAs can provide support if you use R. Alternatively, you may choose to use Java together with the WEKA toolbox. For all steps, you may always use/include existing software libraries.

Part III

In addition, your teacher and TAs, together with the OUH, have kindly prepared an evaluation data set that we can use to evaluate the performance of your classifier. We also assembled a group of people, offered them candies, asked them to breath into the BioScout device and stored this data. We will provide you with this data (URL8) but without the class labels. Your task is to

(1) Use the code from part I to process and align this data set to create a second matrix M_{test}: *peaks x volunteers*, which will serve as test data.

(2) apply your classifier learned in part II to this new data sets.

(3) hand in the predicted test set class label annotations, i.e. your prediction of candies. For each measurement, give us the candy label. You can test your predictions by uploading a class label file, as exemplarily shown in URL6, to the web service at URL9.

RULES:

(1) You work in groups of 4-5 students on the software, results and documentation. You hand in as a group and learn how to work together.

(2) The short presentation during the oral exam, however, will be individual for each student, of course! So better contribute intensively to your group's project work.

(3) Prepare and hand in:

a. The software incl. source code and a short manual describing how to use and how to execute it.

b. A list of class labels for the samples from part III (see URL6).

c. Hand-in electronically by email to your TAs.

d. Deadline for final hand-in is January 5.

(4) You can always contact your TAs or your lecturer if you have questions or need clarification.

(5) You may utilize all kinds of existing software, libraries, internet, google, Wikipedia, etc. to solve your tasks.

(6) For the first part of your final (oral) exam, prepare a short presentation of 5 minutes about your project. Describe its motivation, the data sets you were given, your tasks, your general approach, and your results. Keep it short! This is not much time. You may use whatever (legal) tools you like (videos, Powerpoint, your Notebook, a book, …).

LINKS:

URL1:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws17-18/DM847/project/toy_data/candy_raw.zip

URL2:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws17-18/DM847/project/peax1.0-LinuxX64.zip

URL3:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws17-18/DM847/project/toy_data/candy_peax_processed.zip

URL4:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws17-18/DM847/project/toy_data/peak_alignment.txt

URL5:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws17-18/DM847/project/toy_data/indicator_matrix.txt

URL6:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws17-18/DM847/project/toy_data/class_labels.txt

URL7:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws17-18/DM847/project/data/labelled_candy_raw.zip

The ZIP file is password protected. The password will be sent to you via the Blackboard/e-learn.sdu framework and/or facebook.

URL8:

http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws17-18/DM847/project/data/unlabelled_candy_raw.zip

URL9:

http://shiny.compbio.sdu.dk/breath-checker/

ADDITIONAL LINKS:

Help with R: http://www.statmethods.net

WEKA: http://www.cs.waikato.ac.nz/ml/weka/

PAPERS:

Peak picking:

http://www.biomedcentral.com/1471-2105/15/25

http://pubs.acs.org/doi/abs/10.1021/ac60353a013

http://www.sciencedirect.com/science/article/pii/S002196730400490X#

Breath data mining:

http://www.mdpi.com/2218-1989/3/2/277/pdf

http://www.funpecrp.com.br/gmr/year2012/vol11-3/pdf/gmr2065.pdf

FIGURES:

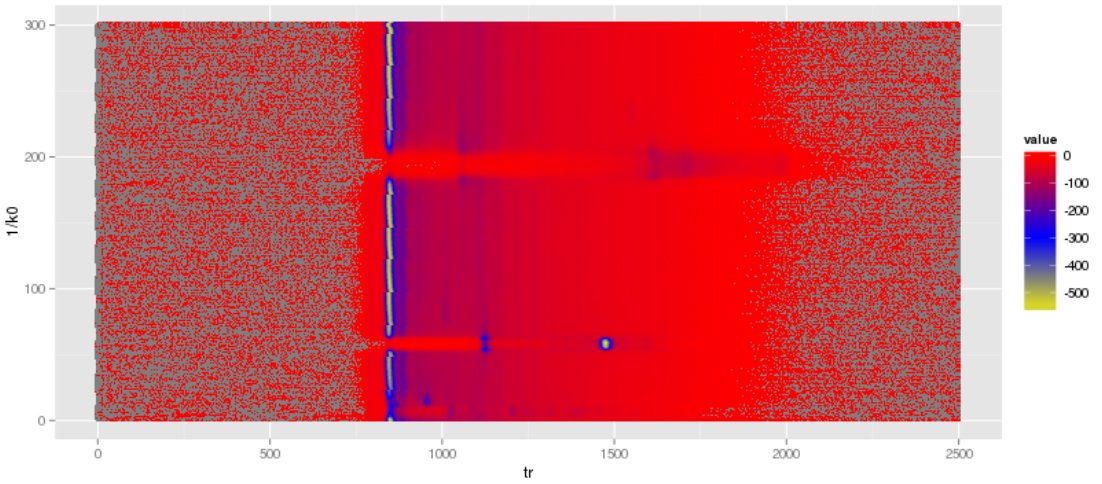


Figure 1: A crude example of a density plot for IMS raw data (the intensity matrix). This plot was created with the ggplot2 package in R.

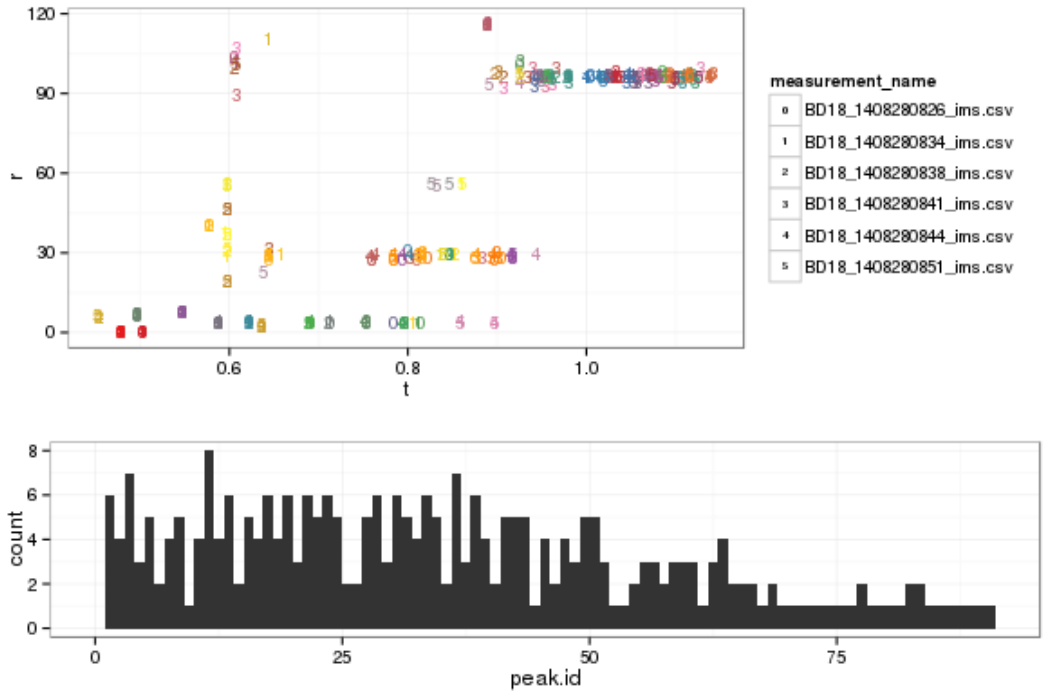


Figure 2: Example for a peak alignment result. Peaks from different measurements are aligned by their Euclidean distance and a given distance threshold. Different colors correspond to different peaks. The histogram on the bottom shows how often each peak is represented across different measurements.

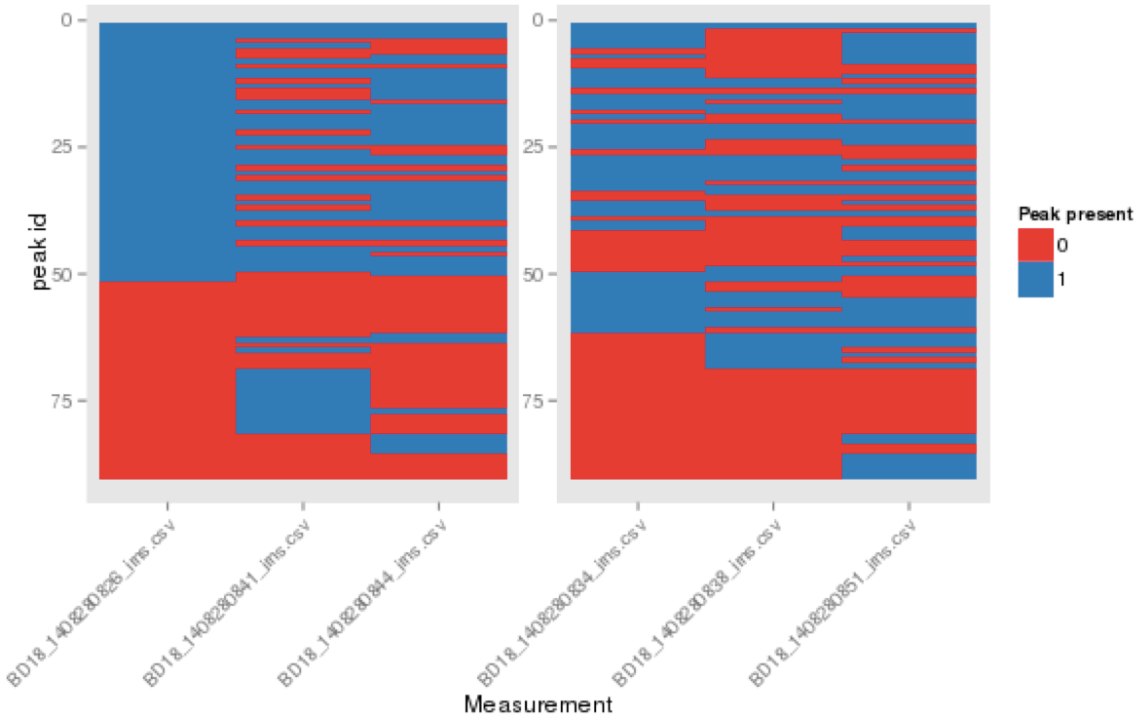


Figure 3: Illustration of an indicator (feature occurrence) matrix created from different measurements (volunteers) for the two class toy data set. Left side: "original", right side: "citrus".

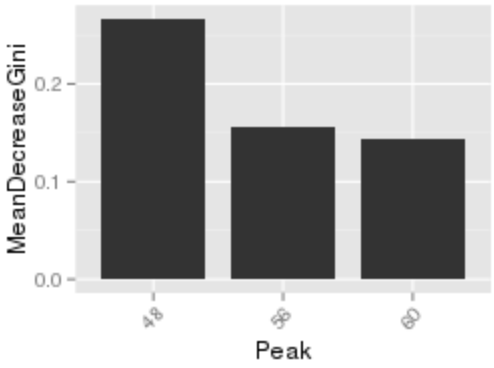


Figure 4: Top three discriminating peaks in the toy data set, sorted by their mean decrease in Gini index.

TABLES:

Peak No.	D – drift time	t – retention time
48 (id 48)	0.87471	56.754
56 (id 181)	0.59674	75.41
60 (id 331)	0.92535	195.874

Table 1: Top three discriminating peaks in the toy data set as illustrated in Figure 4. Here, we also give the drift time and retention times of the peaks, as your peak IDs would be different from ours.