Knowledge-Based Systems in Bioinformatics
Lab 3 – Discretization & Classification

Winter 2026

Teaching Assistants

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

1 Tasks

Consider two decision systems $A = (U, A \cup \{d\})$ and $B = (U, B \cup \{d\})$:
Discretize following datasets using **equal frequency** (with 3 levels) and **naive** methods. Answer the question:

- Which discretization algorithm will discretize these decision systems best and why?

The report should contain all calculations and tables with cuts instead of values. While defining the cuts use the following formula: $[c1, c2)$, where $c1$ is cut 1 and $c2$ is cut 2.

feature_1	decision
100	NO
50	NO
36	YES
40	YES
21	YES
104	NO
102	NO
73	YES
82	NO

Table 1: Decision system A

width [mm]	decision
0.82	B
0.44	A
0.66	A
0.68	A
0.12	A
0.22	B
0.24	B
0.96	A
0.34	B
0.04	B

Table 2: Decision system B

2 Introduction

R.ROSETTA is an R package built upon the ROSETTA rough set classifier. In addition to all the existing ROSETTA algorithms, we have added new functions especially useful in bioinformatics applications. These include: under sampling, rule p-value estimation, rule visualization methods and support sets detection. The package is publicly available on the GitHub repository: `https://github.com/komorowskilab/R.ROSETTA`

R.ROSETTA includes a sample dataset of gene expression measures. The objects are divided into two decision classes: male children with autism and healthy controls. The features are represented by genes. A following decision table contains features selected by a fast correlation-based filter (FCBF). Sample data is publicly available at GEO repository with the reference number GSE25507. More information about the data can be found in the NCBI webpage: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25507`.

3 Tasks

1. Install R.ROSETTA from GitHub repository and load the library.
   - The sample data is automatically loaded with the package and exists in the workspace as `autcon`.
2. Describe the dataset:
   (a) What is the number of features?
   (b) What is the number of objects in each class?
   (c) Do you think the distribution of objects is balanced?
3. Run `rosetta()` with the roc option set to `TRUE` and for autism class. Assign the results to `autconJohnson`. Use `autconJohnson$main` to retrieve the rule table information and store the

results in a separate variable. Use `autconJohnson$quality` and display the quality statistic of the model.

4. Answer the questions:
   (a) Define what is a cross-validation. How many cross-validations are performed in R.ROSETTA by default?
   (b) What is the default reduction method?
   (c) What is the default method of discretisation? Describe it shortly.
   (d) How many discretisation levels are produced by default?
   (e) What is the mean accuracy of the model?
   (f) What test shall be performed to check if model is statistically significant? Describe shortly the procedure.
   (g) How many rules are produced? Print out top 12 most significant rules (p-value < 0.05) for autism class. Use function `viewRules()` for printing rules.
   (h) Define the following measures of the rules: accuracy, support and cuts.
   (i) Display and describe top 12 rules for autism. What is their length? What is the quality of rules? Describe shortly what is the meaning of these rules.
   (j) Which class got more significant rules? What is the fraction of significant rules for each class? You can judge the significance of the rule if the (p-value < 0.05) .

5. Display ROC curve for the autism class. To plot the ROC curve use a function `plotMeanROC()`. Include the plot in the report and answer the questions:
   (a) What is the mean AUC of the model?
   (b) Define the values on the x and y axis of the ROC curve plot.
   (c) What is the interpretation of the area under the ROC curve?
   (d) Can you think of any problems occurring from comparing the performance of two classifiers using ROC?
   (e) Find the best threshold for the class? How do you define "best"?

6. Compare the significant features between the classes.
   (a) Use the function `getFeatures()`.
   (b) Display common genes for both classes.
   (c) Display unique genes for each class

7. Recalculate your model and visualize a rule information.
   (a) Use the function `recalculateRules()` and recalculate the model.
   (b) What information is stored in the column *supportSetRHS*?
   (c) Find the most significant rule for **control** and plot a heatmap. Use `plotRule()` function. Set labels according to this guideline:
   Include the plot in the report. What does the heatmap represent? Describe it shortly.

(d) For the same rule display a boxplot. Use `plotRule()` function. Set labels according to this guideline:

- 3 - high
- 2 - medium
- 1 - low

Include the plot in the report. Compare the gene expression of objects supporting the rule to the remaining objects.

8. Hieracrchical clustering of the rules

(a) Use the recalculated model

(b) Prune out the significant rules with p-value < 0.05

(c) Use the function `clusterRules()`.

(d) **Each cell in the heatmap displays either of the two values: 0 (indicating that the object does not support the rule) or 1 (indicating support for the rule).**

(e) Examine the clusters of autism and control in the rows, focusing on the rules that provide support.

(f) Analyze what the heatmap reveals about the rules and decisions.

(g) Briefly describe how effectively the rules classify the objects.

Hints

- Operations on vectors can be performed with: `union()`, `intersect()`, `setdiff()`
- Significance of the rules is defined as: ns $p > 0.05, *p \leq 0.05, **p \leq 0.01$ and $***p \leq 0.001$