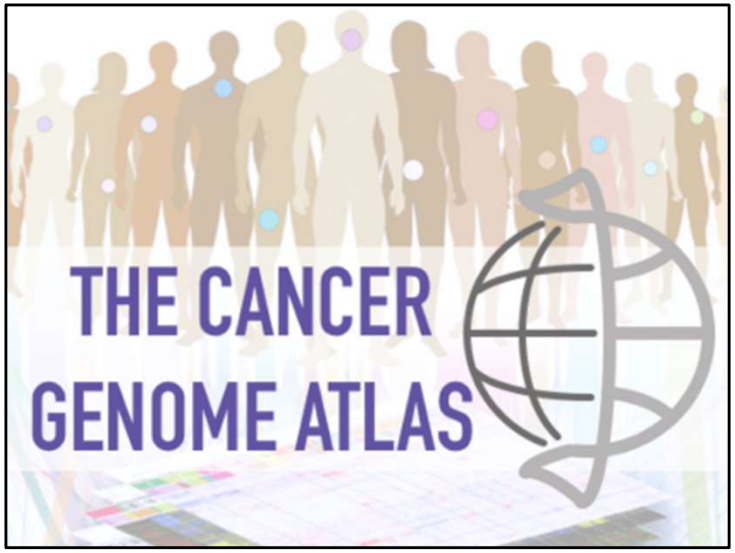
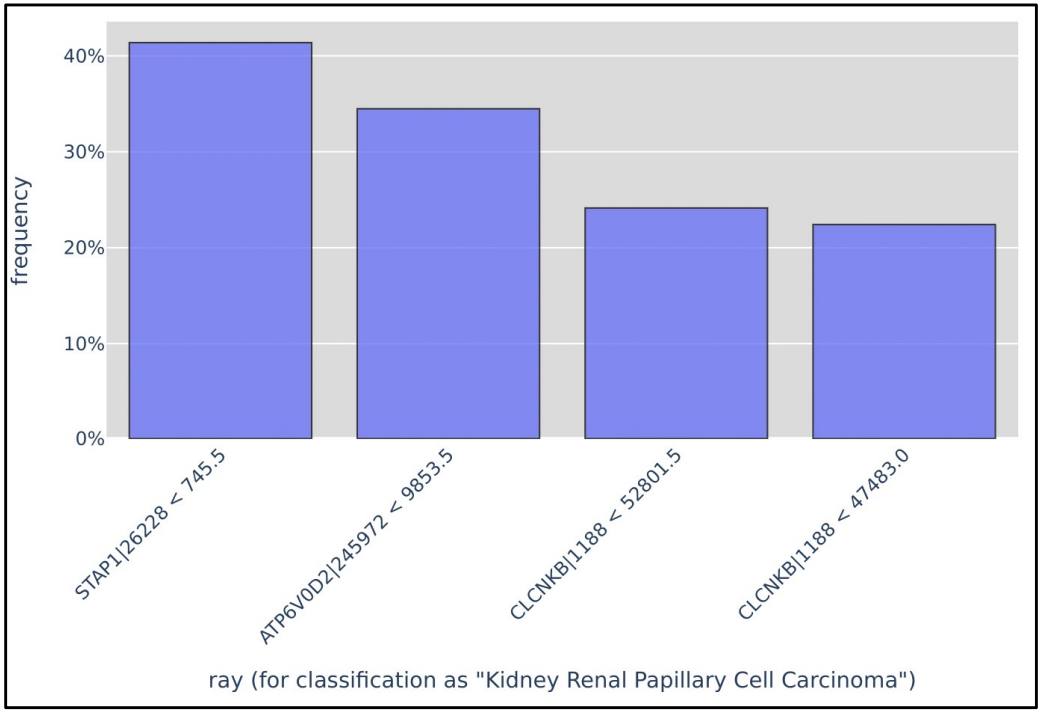
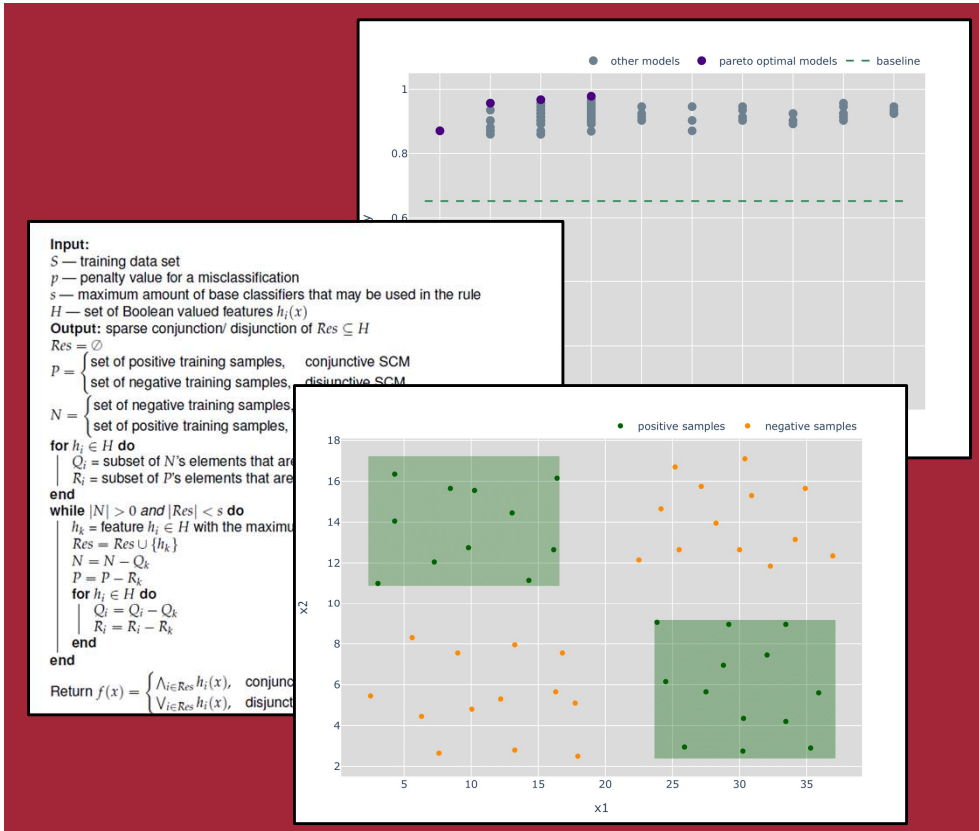


https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga





Robustifying the Set Covering Machine with Disjunctive Normal Forms and Nominal Features

Bachelor‘s Thesis in Software Engineering
by Rebekka Roßberg

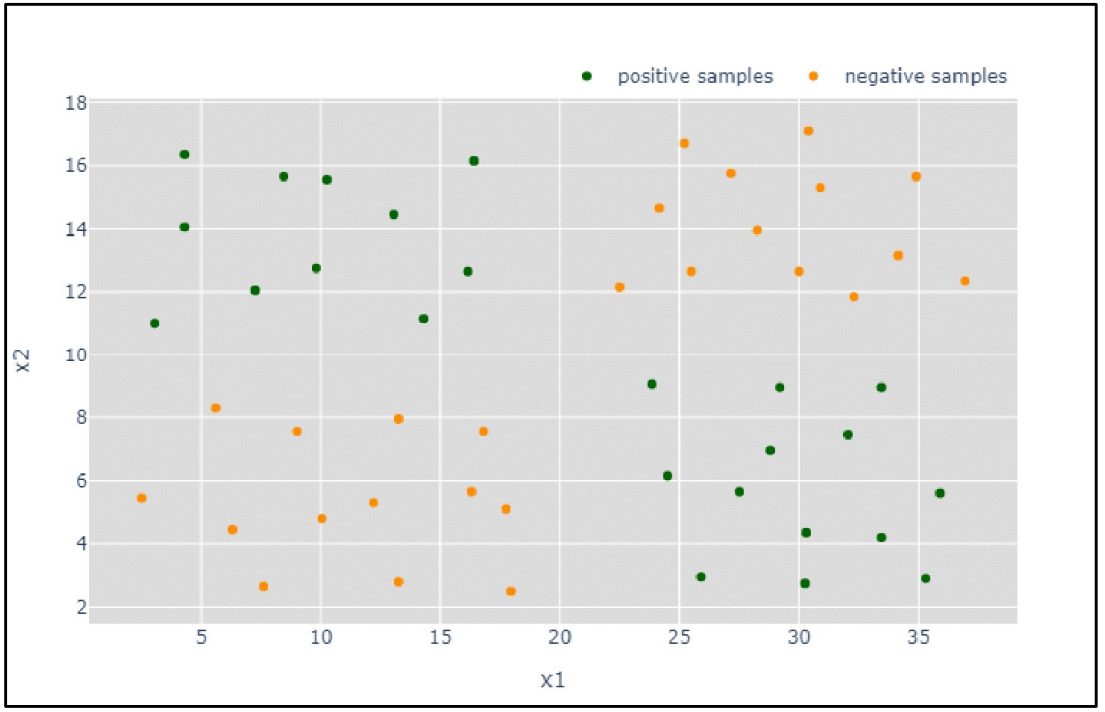
Institute of Medical Systems Biology

Marchand‘s Set Covering Machine

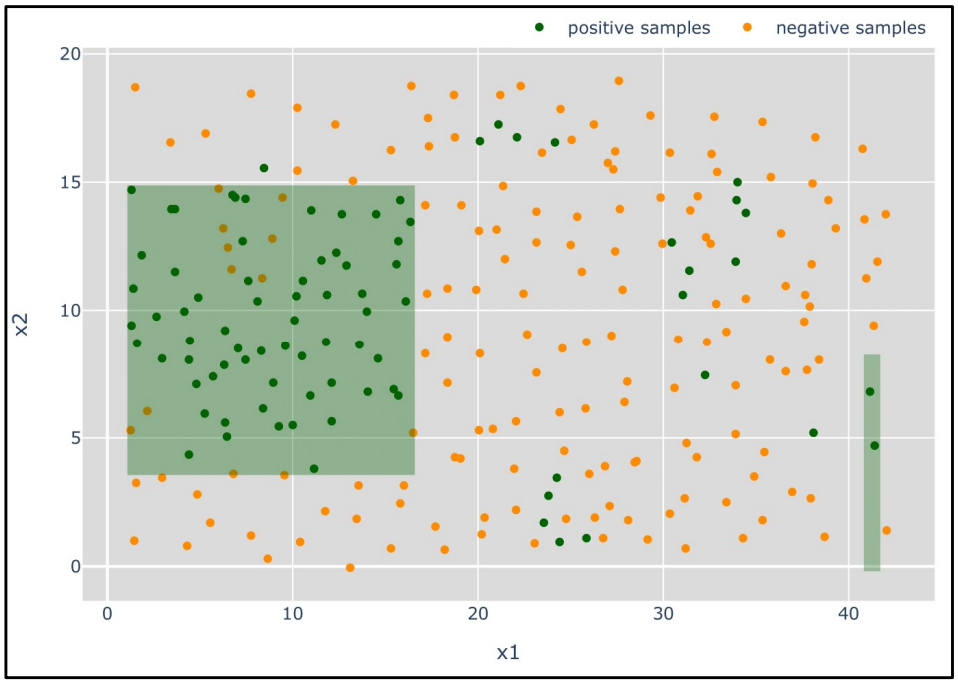
- Based on the algorithms of Valiant and Haussler
- Produces a conjunction *or* disjunction of base classifiers
- Different kinds of base classifiers possible, such as data-dependent rays or balls
- Results in a compact and easy to interpret classifier and therefore exposes relevant features
- Can handle high-dimensional data sets of low cardinality

Building a Disjunction of Conjunctive Rules

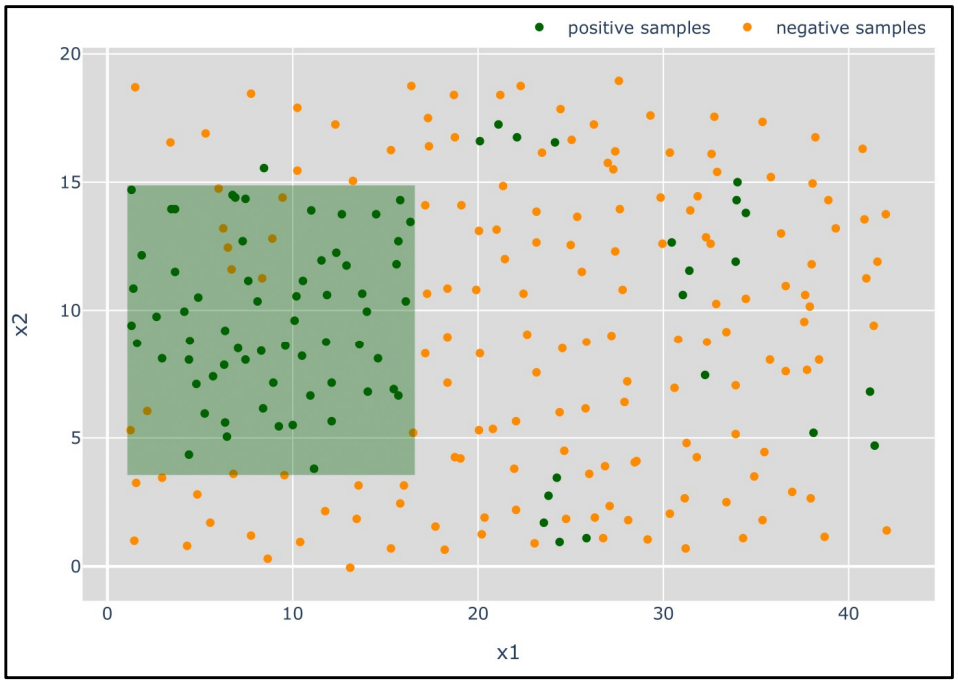
- Extending the expressiveness of the SCM
- Especially useful for data sets with multiple disjunct decision regions
- Increased possibilities to refine the classifier ↔ potential overfitting



Adjusting the Parameters: sC, sD, minConjSize

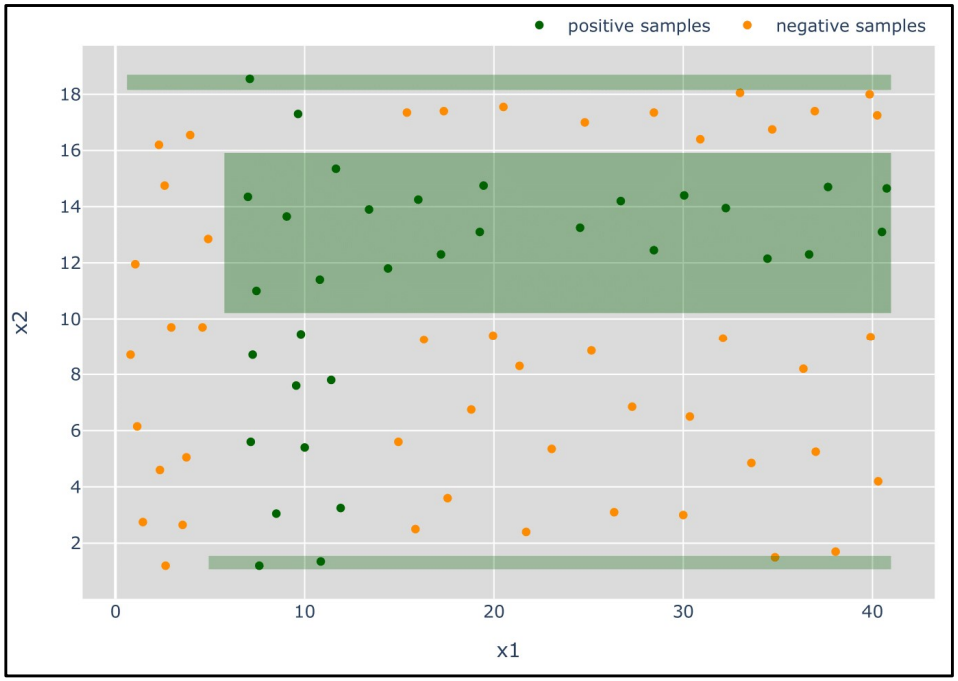


minConjSize = 1

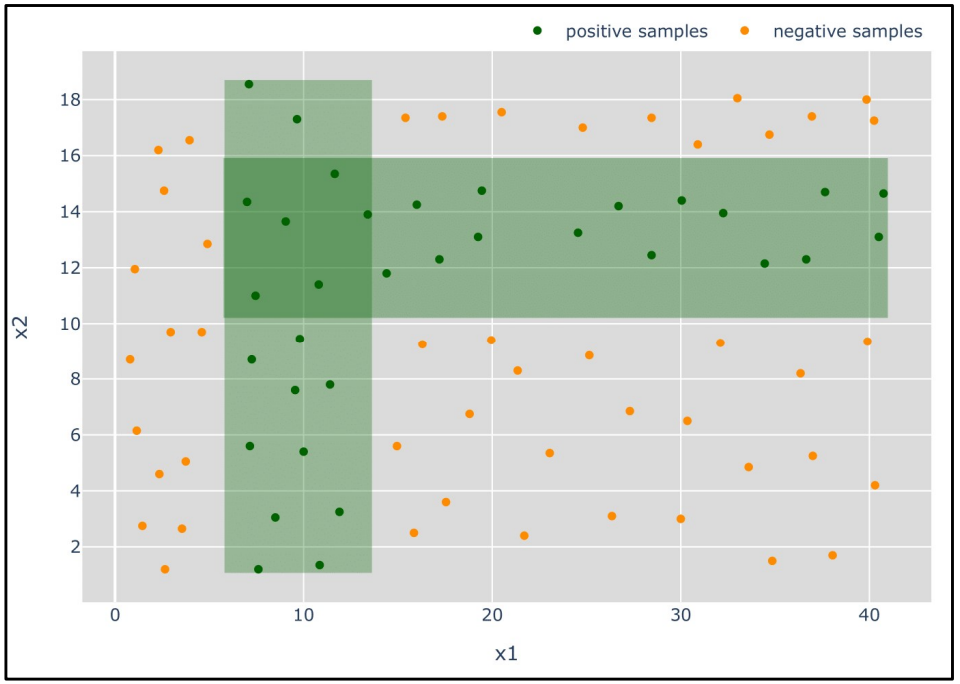


minConjSize = 5

Adjusting the Parameters: p



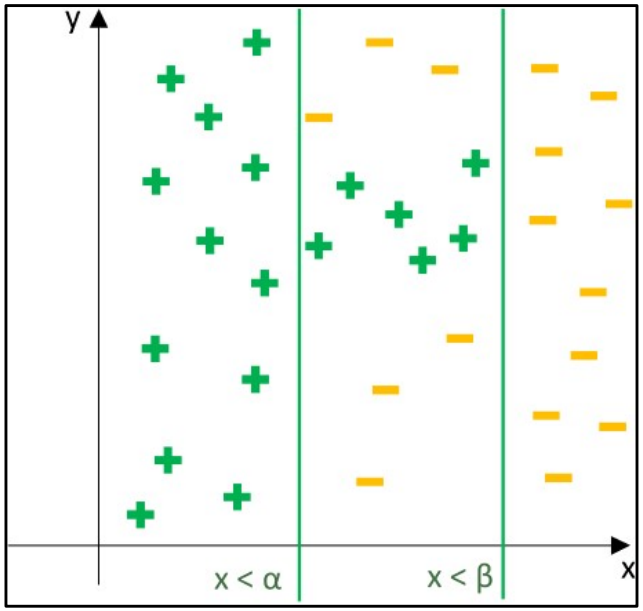
p = 1

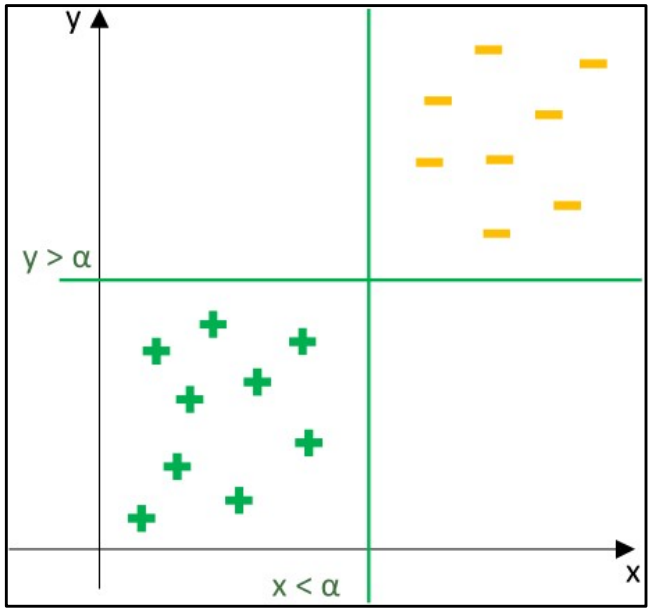


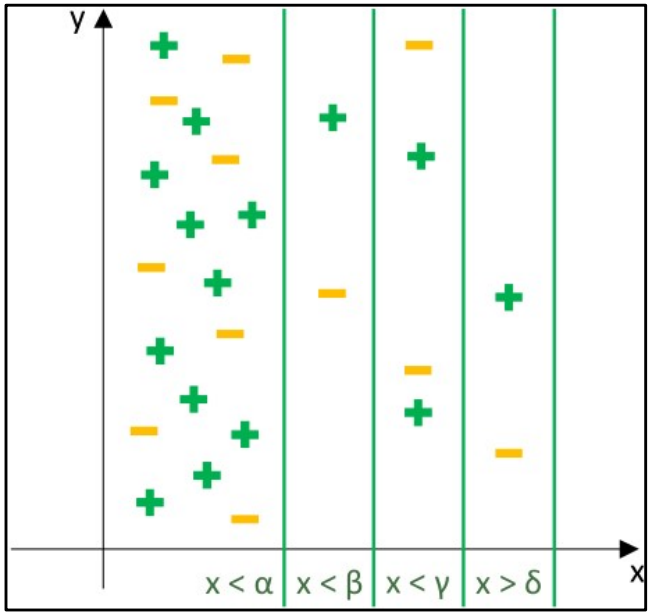
p = 1.5

Resolving Tie Situations

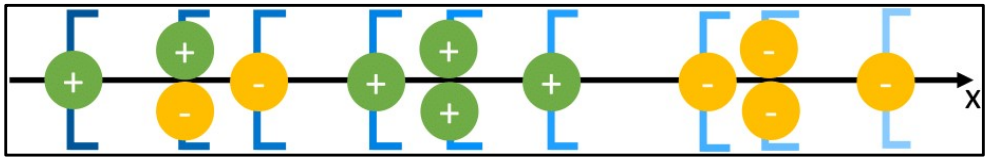
1. Save history and option 2
2. Finish conjunction with option 1
3. Use option 2 as the starting point for the next conjunction



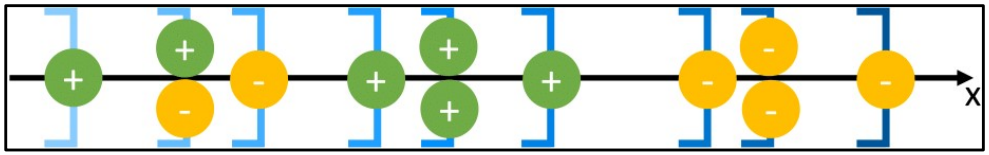




Optimizing Rays: Reducing the Complexity by O(|samples|)



Locating the optimal lower border



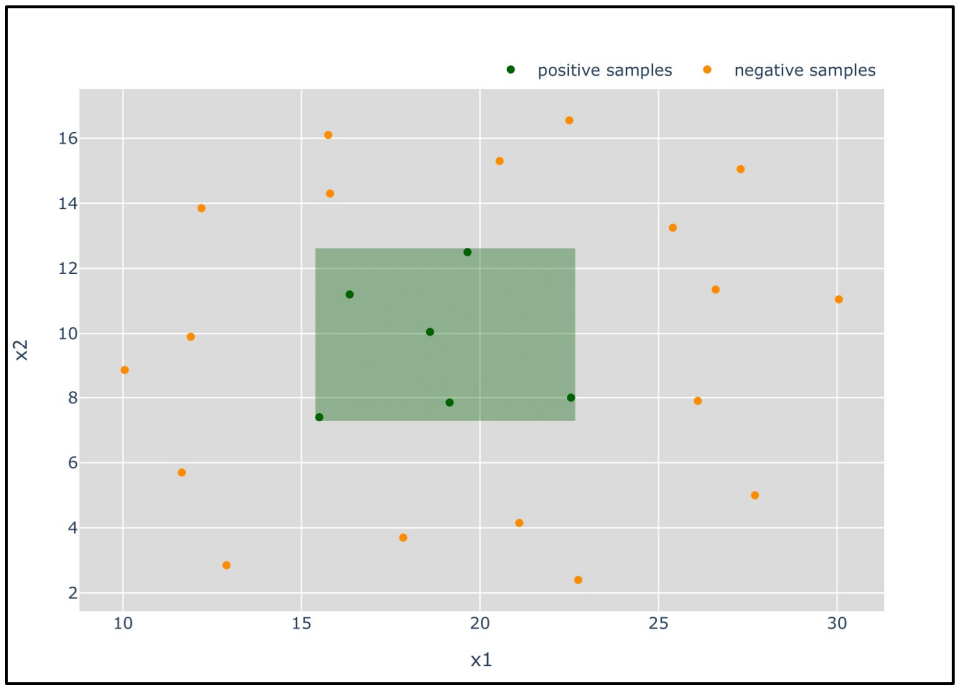
Locating the optimal upper border

Optimizing Rays: Allowing Re-correction

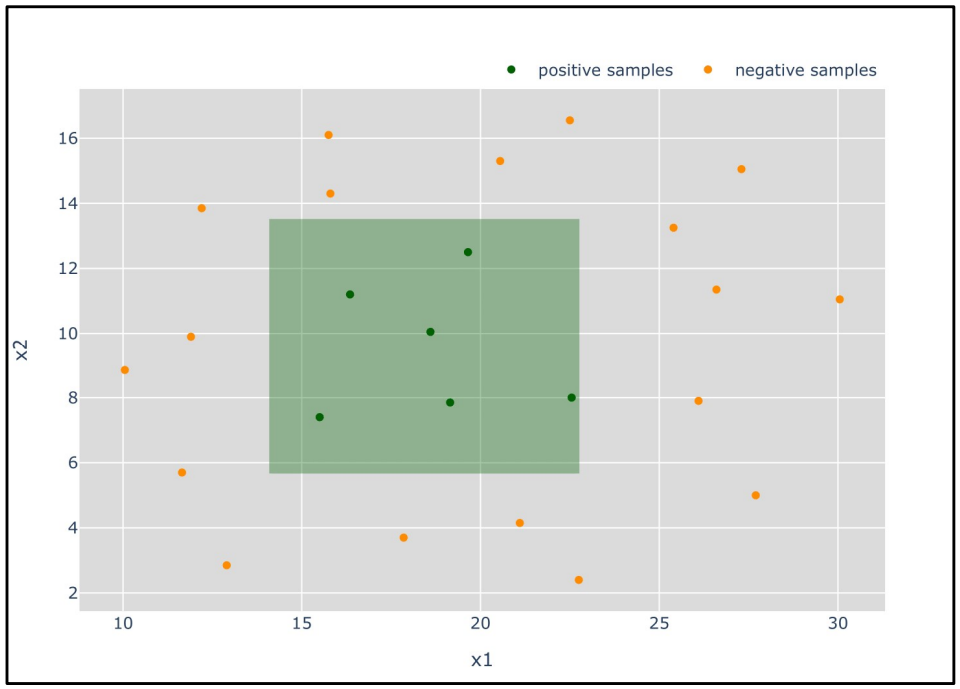
Without re-correction

With re-correction

Optimizing Rays: Clever Placement of Thresholds



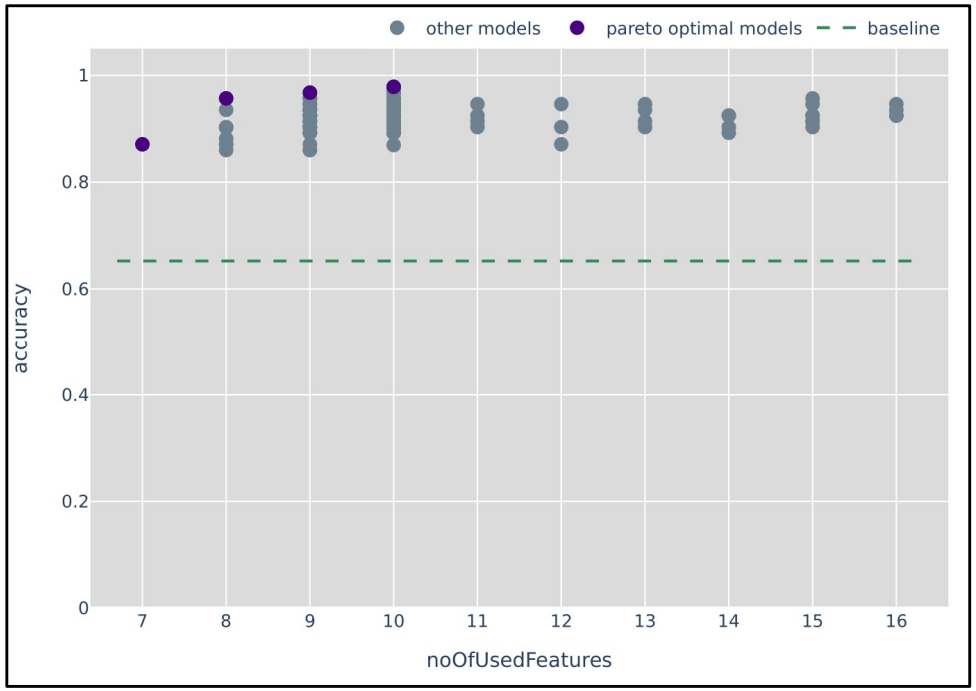
Thresholds on the data points



Thresholds between the data points

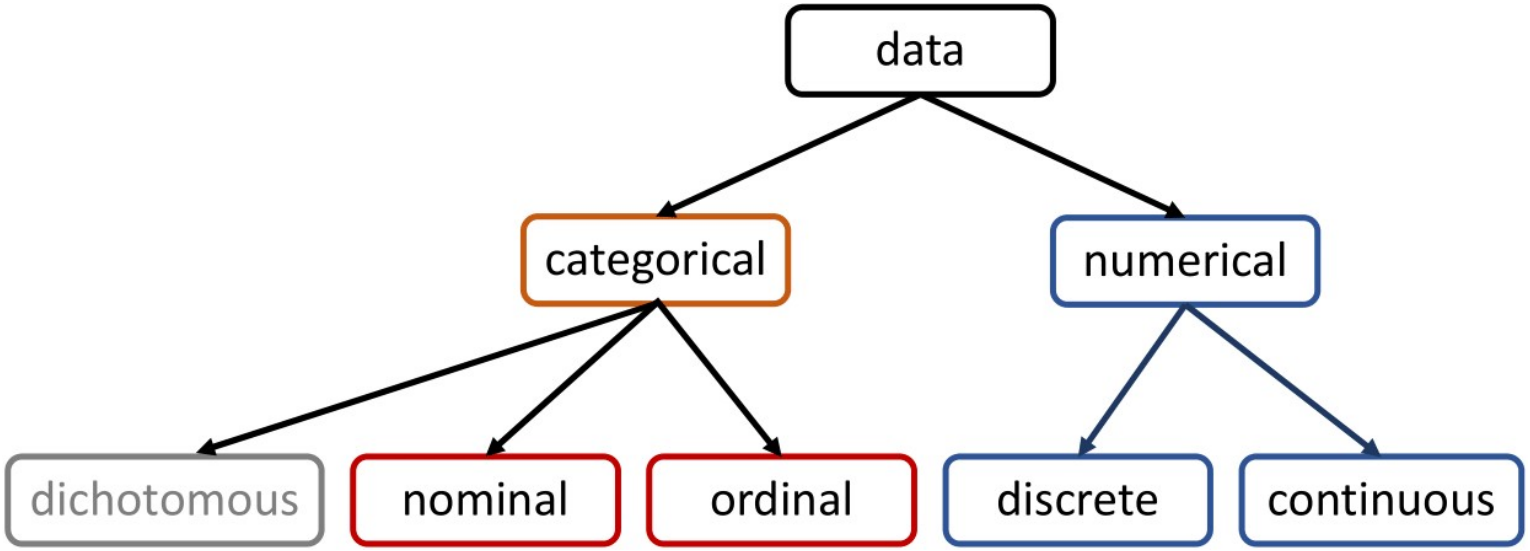
Application on Gene Expression Data

- Kidney Chromophobe
- Kidney Renal Papillary Cell Carcinoma
- Kidney Renal Clear Cell Carcinoma
- Cholangiocarcinoma
- Pancreatic Adenocarcinoma
- Liver Hepatocellular Carcinoma
- Colon Adenocarcinoma
- Rectum Adenocarcinoma



→ **Classifiers like:** IF (EPHA3|2042 > 666.5 AND ADORA2B|136 > 305.5 AND ACO1|48 > 2297.5) OR (ELAVL2|1993 > 209 AND DDC|1644 < 5461.5 AND ABCA4|24 < 8.5) THEN class ‘Rectum Adenocarcinoma’

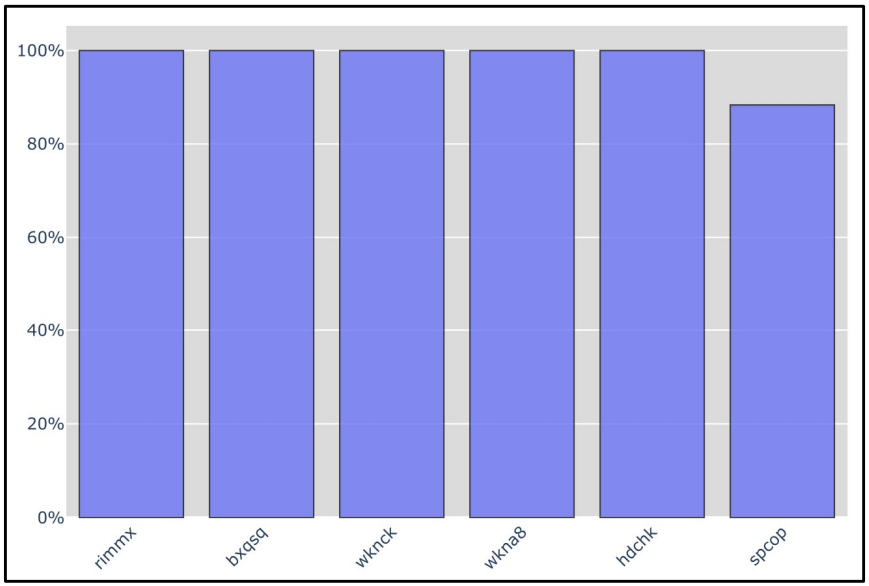
Extension to Nominal Features

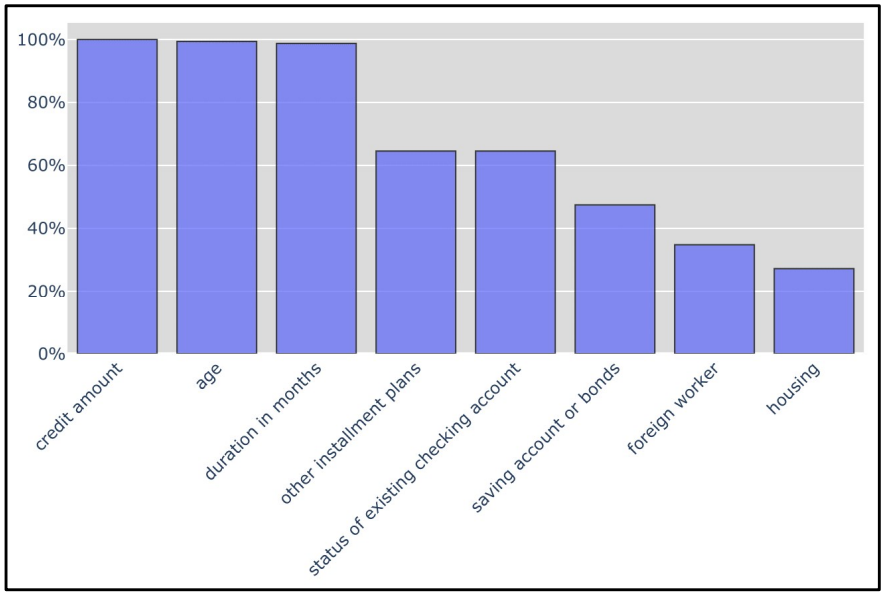


Application on UCI Data Sets

Chess:

IF (bxqsq=f AND wknck=f AND wkna8=f AND hdchk=f AND spcop=f) OR (rimmx=t) THEN won’





German:

IF (status of existing checking account = no checking account AND other installment plans = none AND age > 22.5 AND credit amount < 9504.0 AND age < 66.5) OR (duration in months < 8.5 AND age > 25.0 AND credit amount < 3015.5) OR (credit amount < 421.0) THEN good

Conclusion

- Feasible run times of the optimized algorithm in Julia
- The uniform handling of both, numerical and nominal, features widens the field of possible use cases by a lot

- Using a DNF *can* improve a SCM‘s ability to create an accurate classifier
- Mostly helpful for disjunct and low dimensional data
- However chance is often unused, for example in 5/7 gene expression data sets
- Good parameter adjustmenet is essential, especially of `minConjSize`