Report

Vincent Guidoux & Arnold Von Bauer

0. Preparatory stage (WBCD)

Question 1 - Comment the plot below

question_1

Considering there's two classes, there are evenly distributed.

1. Launch the Trefle experiments (or modeling runs)

We choose `vec_weight = [0.0, 0.33, 0.66, 1]` to have a good view on the distribution but not too long for the model to be tested

Question 2 - Explain what is the meaning/role of the alpha value? Why we "play" with it? How is it related with the weights given to sebnsitivity and specificity? What would imply a high weight for sensitivity, respectively specificity?

remeinder :

- **sensitivity** answer the question : How likely is the test to detect the presence of a characteristic in someone with the characteristic ?
- **specificity** answer the question : Hoe likely is the test to detect the absence of a characteristic in someone without the characteristic ?

	Low weight	**middle weight**	**high weight**
sensitivity	low	high	high
specificity	high	high	low

Question 3 - Decide on an alpha value to be used to define the (fitness) weights for sensitivity and specificity. Explain your choice

We choosed Sensitivity, because it's not because someone is diagnosed that he's going to go through the whole treatment, doctors expert will look at that result and take measures.

It's better to have more people diagnoied than less, at forgot people on the road.

Question 4 - Analyze the graphic above and decide on a weight for the RMSE contribution to the fitness funnction. Justify your choice. What are your final values for the three weights? How do you interpret them?

RMSE fitness score will be low like 0.2,

- sensitivity : 0.7
- specificity : 0.3
- RMSE : 0.1

Question 5 - Explain what are the implications of these two parameters (i.e., number of rules and number of variables per rule) on the models, in terms of both performance and interpretability.

	number of rule & number of variable per rule	
performance	The more there are, the performance the better	
interpretability	The less there are, the interpretability the better	

Question 6 - If you have setted your algorithm up to use 6 rules and 5 variables per rule on a dataset composeed of 100 features, how many features could be used at most by a model?

6 * 5 = 30 features could be used at most by a model that is settled to use 6 rules and 5 variables per rule

Question 7 - In your opinion, why did we decide to first explore the number of rules instead of the number of variables per rule?

The impact is bigger of the number of rule, if you made an error on setting the number of rule it's more adjustable, than making an error with number of variable per rule, and after, correcting the number of rule

Question 8 - Which values have you decided to test at this stage? Why this range?

In the course, we saw that a number of rules > 7 +/- 2 was too big. So we decide too look at 3, 5, and 7 rules, thinking that 9 was too much and 3 was a compromised minima

Question 9 - On the base of the graphic above, select a narrower range for the number of rules to be explored in the next step. Justify your choice.

We see that the higher value for sensitivity is around 5 and it deacays less to the left, so we choose 4 and 5 rules.

Question 10 - Then, define a range of values for the number of variables per rule. How did you decide on them? Why?

For interpretability, we choose 2,3,5. 5 is the value use during all the lab, it seemed to big for us, but we wanted to see, and 2, and 3 is for the check.

Question 11 - In your opinion, which values/ranges of both parameters: number of rules and vars per rule, should you choose to obtain the best models? (comment briefly on the plot and include it into to report)

2 Model selection

Question 12 - Explain your choice of the threshold values for the sensitivity and specificity. (Save both plots into your reports)

Question 13 - Explain your choice of the threshold. (Save both plots into your report)

3. Analysis of the selected models

Question 14: Among the final models, select three of them as follows: the smallest one (in terms of rules and variables), the best one (in terms of performance), and one in the "middle" that you consider as being a good trade-off between size and performance. With them: