Assignment 1 - Language Development in ASD - part 3

Riccardo Fusaroli

August 10, 2018

## Welcome to the third exciting part of the Language Development in ASD exercise

In this exercise we will delve more in depth with different practices of model comparison and model selection, by first evaluating your models from last time, then learning how to cross-validate models and finally how to systematically compare models.

N.B. There are several datasets for this exercise, so pay attention to which one you are using!

1. The (training) dataset from last time (the awesome one you produced :-) ).
2. The (test) datasets on which you can test the models from last time:

* Demographic and clinical data: <https://www.dropbox.com/s/ra99bdvm6fzay3g/demo_test.csv?dl=1>
* Utterance Length data: <https://www.dropbox.com/s/uxtqqzl18nwxowq/LU_test.csv?dl=1>
* Word data: <https://www.dropbox.com/s/1ces4hv8kh0stov/token_test.csv?dl=1>

### Exercise 1) Testing model performance

How did your models from last time perform? In this exercise you have to compare the results on the training data () and on the test data. Report both of them. Compare them. Discuss why they are different.

My model from the last time performed with an rmse of 0.2894077 on the training data, and 1.00916 on the test data.

### Exercise 2) Model Selection via Cross-validation (N.B: ChildMLU!)

One way to reduce bad surprises when testing a model on new data is to train the model via cross-validation.

In this exercise you have to use cross-validation to calculate the predictive error of your models and use this predictive error to select the best possible model.

* Use cross-validation to compare your model from last week with the basic model (Child MLU as a function of Time and Diagnosis, and don’t forget the random effects!)
* (Tips): google the function “createFolds”; loop through each fold, train both models on the other folds and test them on the fold)

Which model is better at predicting new data: the one you selected last week or the one chosen via cross-validation this week?

* Test both of them on the test data.
* Report the results and comment on them.

“mean rmse for quadratic model on test data: 0.721611653317222” “mean rmse for linear model on test data: 1.08373418211466”

Quadratic model = (CHI\_MLU ~ Diagnosis \* VISIT + I(VISIT^2) + (1+VISIT+ I(VISIT^2)|SUBJ)) Linear model = (CHI\_MLU ¨Diagnosis \* VISIT + I(VISIT^2) + (1+VISIT + I(VISIT^2)|SUBJ))

The above quadratic model performed better on the training data than the basic linear model from last week. The development of CHI\_MLU seems visually to be fit better by a quadratic model, so this is as expected so far. Unsuprisingly, both models performed better on the training data which indicates an overfit. Increasing the amount of data or experimenting with amount of folds could influence these results.

* Now try to find the best possible predictive model of ChildMLU, that is, the one that produces the best cross-validated results.

CHI\_MLU ~ ADOS \* Age (1+Age|SUBJ) performed best after cross validation, rmse = (0.62). As ADOS is a more nuanced description of severity of ASD than Diagnosis is, this makes intuitive sense.

* Bonus Question 1: What is the effect of changing the number of folds? Can you plot RMSE as a function of number of folds?
* Bonus Question 2: compare the cross-validated predictive error against the actual predictive error on the test data

[HERE GOES YOUR ANSWER]

### Exercise 3) Assessing the single child

Let’s get to business. This new kiddo - Bernie - has entered your clinic. This child has to be assessed according to his group’s average and his expected development.

Bernie is one of the six kids in the test dataset, so make sure to extract that child alone for the following analysis.

You want to evaluate:

* how does the child fare in ChildMLU compared to the average TD child at each visit? Define the distance in terms of absolute difference between this Child and the average TD. (Tip: recreate the equation of the model: Y=Intercept+BetaX1+BetaX2, etc; input the average of the TD group for each parameter in the model as X1, X2, etc.).

Using the quadratic model I have estimated that Bernie differs from subjects with a TD diagnosis with the values following values pr. visit:

Visit 1 = 0.7444560 Visit 2 = 0.7644444 Visit 3 = 1.1131915 Visit 4 = 0.5630986 Visit 5 = 0.2532523 Visit 6 = 0.3084127.

* how does the child fare compared to the model predictions at Visit 6? Is the child below or above expectations? (tip: use the predict() function on Bernie’s data only and compare the prediction with the actual performance of the child)

Bernie’s MLU is 3.45 , much better than the quadratic model’s prediction of MLU (1.88). Upon inspecting Bernies ADOS value it should be noted that it is only 7. Arguably this indication of the low severity of Autism in Bernie explains his high MLU.