Portfolio 3

Kiri Koppelgaard

September 28, 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.

- recreate the models you chose last time (just write the model code again and apply it to your training data (from the first assignment))
- calculate performance of the model on the training data: root mean square error is a good measure. (Tip: google the function rmse())
- create the test dataset (apply the code from assignment 1 part 1 to clean up the 3 test datasets)
- test the performance of the models on the test data (Tips: google the functions "predict()")
- optional: predictions are never certain, can you identify the uncertainty of the predictions? (e.g. google predictinterval())

formatting tip: If you write code in this document and plan to hand it in, remember to put include=FALSE in the code chunks before handing in.

REPONSE: The quadratic model (mean length of utterance ~ visit + visit^2 + diagnosis + (VISIT+ VISIT^2|SUBJ)) produces a root mean square error of 0.289, when explaining the train data.The mean squared error increases when applying the model on the test set and goes from 0.289 to 0.773. Since the error increases when applied to the test set it appears the model does only has limited power to predict and generalize to the population. This could be due to overfitting of the data in the training set.

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)
- Test both of them on the test data.
- Report the results and comment on them.
- Now try to find the best possible predictive model of ChildMLU, that is, the one that produces the best cross-validated results.
- Which model is better at predicting new data: the one you selected last week or the one chosen via cross-validation this week?
- 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

RESPONSE: The quadratic model (mean length of utterance ~ visit + visitˆ2 + diagnosis + (VISIT+ VISITˆ2|SUBJ)) produces a mean squared error of 0.773, when applied to the test data compared to 0.658 produced by the best explaining model from last time (mean length of utterance ~ Diagnosis + visit+ visitˆ2 + ADOS + verbal IQ + nonverbal IQ + (VISIT+ I(VISITˆ2)|SUBJ)).Thus, the best model is still the last mentioned, which is able to predict the new data the best. Thus, it does not appear that the fancy model does not overfit as could have been expected.

Based on the mean of the mean squared errors, when applied to the cross-validated test data, the best predictive model is mean length of utterance ~ visit + diagnosis + MOT_MLU +ADOS + types_CHI + tokens_CHI + I(VISITˆ2)+(1+VISIT+ I(VISITˆ2)|SUBJ). This also counts when the model is applied to the ‘true’ test data, there it produces a root mean square error of 0.47 compared to the best explaining model from last week, which has a root mean square error of 0.66.

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.).
- 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)

Based on the best model from the cross-validation the typically developed child starts of with a mean length of utterance on 2.17, whereas Bernie starts 2.54. Thus, Bernie makes longer utterances when entering the experiment. At the 6th visit Bernie has a mean length of utterance of 3.17, whereas a typically developed child has a mean length of utterance of 2.40. Thus, Bernie is pretty genius. He both starts of with a higher MLU and develops faster, it would seem.

Based on the best predictive model found by cross-validation, Bernie is estimated to having a mean length of utterance of 3.28 at visit 6, which is relatively close to the true mean of 3.17. The model has slightly overestimated him and predicts he has a faster development than he in reality has, but it is close.

OPTIONAL: Exercise 4) Model Selection via Information Criteria

Another way to reduce the bad surprises when testing a model on new data is to pay close attention to the relative information criteria between the models you are comparing. Let's learn how to do that!

Re-create a selection of possible models explaining ChildMLU (the ones you tested for exercise 2, but now trained on the full dataset and not cross-validated).

Then try to find the best possible predictive model of ChildMLU, that is, the one that produces the lowest information criterion.

- Bonus question for the optional exercise: are information criteria correlated with cross-validated RMSE? That is, if you take AIC for Model 1, Model 2 and Model 3, do they co-vary with their cross-validated RMSE?

OPTIONAL: Exercise 5): Using Lasso for model selection

Welcome to the last secret exercise. If you have already solved the previous exercises, and still there's not enough for you, you can expand your expertise by learning about penalizations. Check out this tutorial: http://machinelearningmastery.com/penalized-regression-in-r/ and make sure to google what penalization is, with a focus on L1 and L2-norms. Then try them on your data!