UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

Master in Innovation and Research in Informatics

Master's Thesis

Graph Neural Network Applications

Author:
Pau Rodríguez Esmerats

Professor:
Marta Arias Vicente

August 4, 2019

Contents

Abstract

abstract-text

Contents

1 Introduction

Introduction...

2 State of the art

SOTA...

3 Methodology

Methodology...

4 Implementation

Implementation..

5 Experiments

This section exposes the selected approach to perform feature and model selection through different experiments. An experiment consists on training a model over a selected dataset, using cross validation to select the hyper parameters that produce a model with the minimum validation error. Our strategy consists on explore the space of combinations of models and datatsets in order the find the best performing model over the best dataset. The steps we follow are summaryzed here:

- First we start from a group of different datasets, each containing features produced on different analyses.
- Then we train all the models that we have considered for the project over each of those datasets, performing cross validation each time.
- After that, we select the top 3 best pairs model-dataset, the ones with minimum validation error, and perform feature selection over the dataset using forward selection.
- Finally, for each pair, the model is trained again over the resulting dataset.

The next table shows fragments of all the experiments and their error measures. For each experiment we printed the normalized root mean squared error obtained during the cross validation (shown as Validation.NRMSE), obtained as a mean of all the nrmse of each model and fold. This is the value that we use to select the best model, by selecting the minimum normalized root mean squared error. We also show the normalized mean squared error that the selected model, which is then trained again over all the training data set, obtains when predicting over the test set. This is the error that helps us see how well the selected model will generalize over new data.

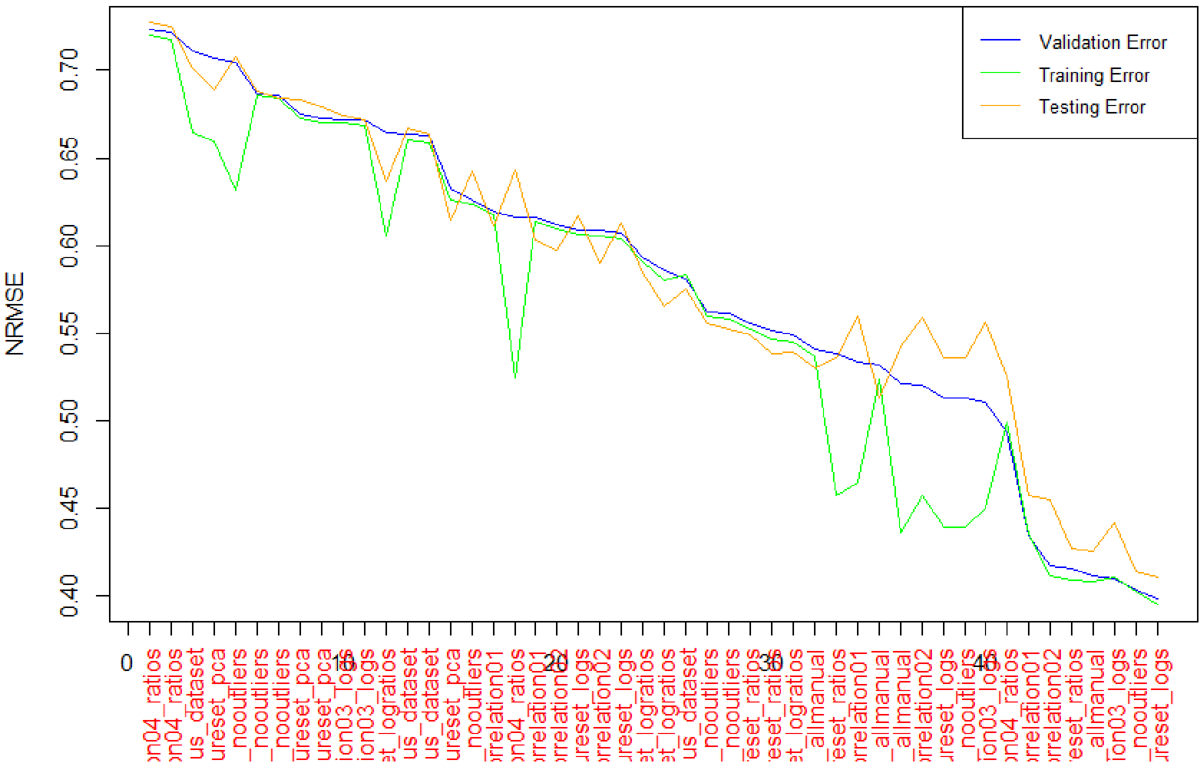

Figure 1: Evolution of the NRMSE for the validation, training and testing set

	subset[, 2]	Comment	Validation.NRMSE	Testing.NRMSE
51	regression_randomforest	featureset_logs	0.40	0.41
56	regression_randomforest	featureset_original_nooutliers	0.40	0.41
54	regression_randomforest	featureset_nocorrelation03_logs	0.41	0.44
49	regression_randomforest	featureset_allmanual	0.41	0.43
59	regression_randomforest	featureset_ratios	0.42	0.43
53	regression_randomforest	featureset_nocorrelation02	0.42	0.45
52	regression_randomforest	featureset_nocorrelation01	0.43	0.46
55	regression_randomforest	featureset_nocorrelation04_ratios	0.49	0.53
42	regression_tree_rpartlib	featureset_nocorrelation03_logs	0.51	0.56
39	regression_tree_rpartlib	featureset_logs	0.51	0.54
44	regression_tree_rpartlib	featureset_original_nooutliers	0.51	0.54
41	regression_tree_rpartlib	featureset_nocorrelation02	0.52	0.56
37	regression_tree_rpartlib	featureset_allmanual	0.52	0.54
25	lasso regression GLMNET	featureset_allmanual	0.53	0.51
40	regression_tree_rpartlib	featureset_nocorrelation01	0.53	0.56
47	regression_tree_rpartlib	featureset_ratios	0.54	0.54
13	ridge regression GLMNET	featureset_allmanual	0.54	0.53
50	regression_randomforest	featureset_logratios	0.55	0.54
35	lasso regression GLMNET	featureset_ratios	0.55	0.54
23	ridge regression GLMNET	featureset_ratios	0.56	0.55
32	lasso regression GLMNET	featureset_original_nooutliers	0.56	0.55
20	ridge regression GLMNET	featureset_original_nooutliers	0.56	0.56
60	regression_randomforest	raw_continuous_dataset	0.58	0.58
26	lasso regression GLMNET	featureset_logratios	0.59	0.57
14	ridge regression GLMNET	featureset_logratios	0.59	0.58
27	lasso regression GLMNET	featureset_logs	0.61	0.61
29	lasso regression GLMNET	featureset_nocorrelation02	0.61	0.59
15	ridge regression GLMNET	featureset_logs	0.61	0.62
17	ridge regression GLMNET	featureset_nocorrelation02	0.61	0.60
28	lasso regression GLMNET	featureset_nocorrelation01	0.62	0.60
43	regression_tree_rpartlib	featureset_nocorrelation04_ratios	0.62	0.64
16	ridge regression GLMNET	featureset_nocorrelation01	0.62	0.61
57	regression_randomforest	featureset_pca_nooutliers	0.63	0.64
58	regression_randomforest	featureset_pca	0.63	0.61
36	lasso regression GLMNET	raw_continuous_dataset	0.66	0.66
24	ridge regression GLMNET	raw_continuous_dataset	0.66	0.67
38	regression_tree_rpartlib	featureset_logratios	0.66	0.64
30	lasso regression GLMNET	featureset_nocorrelation03_logs	0.67	0.67
18	ridge regression GLMNET	featureset_nocorrelation03_logs	0.67	0.67
34	lasso regression GLMNET	featureset_pca	0.67	0.68
22	ridge regression GLMNET	featureset_pca	0.68	0.68
33	lasso regression GLMNET	featureset_pca_nooutliers	0.69	0.68
21	ridge regression GLMNET	featureset_pca_nooutliers	0.69	0.69
45	regression_tree_rpartlib	featureset_pca_nooutliers	0.70	0.71
46	regression_tree_rpartlib	featureset_pca	0.71	0.69
48	regression_tree_rpartlib	raw_continuous_dataset	0.71	0.70
31	lasso regression GLMNET	featureset_nocorrelation04_ratios	0.72	0.72
19	ridge regression GLMNET	featureset_nocorrelation04_ratios	0.72	0.73

Table 1: Experiments performed and their NRMSE for validation and testing

We also show in the fig.1 how the mean of the normalized root mean squared error (NRMSE) when performing cross validation evolves in the different datasets, and compare it to how the training and test NRMSE performs.

Our top 3 best performers consists of Random forest, Decision tree and Lasso regression models trained over feature datasets that contains logarithms of continuous variables, manually selected continuous vars that are not correlated between them or all the original dataset.

	Model	Dataset	Validation NRMSE	Testing NRMSE
51	regression_randomforest	featureset_logs	0.40	0.41
42	regression_tree_rpartlib	featureset_nocorrelation03_logs	0.51	0.56
25	lasso regression GLMNET	featureset_allmanual	0.53	0.51

Table 2: Experiments performed and their NRMSE for validation and testing

Once we have found our best performing model and datasets, we run a forward selection algorithm to perform feature selection over each dataset with its corresponding model.

	Model	Feature set	Validation.NRMSE	Testing.NRMSE
1	regression_randomforest	featureset_logs	1.00	1.00
2	regression_randomforest	featureset_logs 1_7_8_13	0.98	0.98
3	regression_randomforest	featureset_logs 1_7_8_14	0.96	0.96
4	regression_randomforest	featureset_logs 1_7_8_15	0.99	1.00
5	regression_randomforest	featureset_logs 1_7_8_16	0.74	0.75
6	regression_randomforest	featureset_logs 1_7_8_17	0.83	0.81
7	regression_randomforest	featureset_logs 1_7_8_18	0.95	0.94
8	regression_randomforest	featureset_logs 1_7_8_19	0.82	0.83
9	regression_randomforest	featureset_logs 1_7_8_20	0.97	0.97
10	regression_randomforest	featureset_logs 1_7_8_16_13	0.73	0.76
11	regression_randomforest	featureset_logs 1_7_8_16_14	0.74	0.76
12	regression_randomforest	featureset_logs 1_7_8_16_15	0.73	0.74
13	regression_randomforest	featureset_logs 1_7_8_16_17	0.72	0.71
14	regression_randomforest	featureset_logs 1_7_8_16_18	0.70	0.71
15	regression_randomforest	featureset_logs 1_7_8_16_19	0.71	0.72
16	regression_randomforest	featureset_logs 1_7_8_16_20	0.72	0.75
17	regression_randomforest	featureset_logs 1_7_8_16_18_13	0.70	0.71
18	regression_randomforest	featureset_logs 1_7_8_16_18_14	0.71	0.71
19	regression_randomforest	featureset_logs 1_7_8_16_18_15	0.71	0.73
20	regression_randomforest	featureset_logs 1_7_8_16_18_17	0.67	0.67
21	regression_randomforest	featureset_logs 1_7_8_16_18_19	0.69	0.69
22	regression_randomforest	featureset_logs 1_7_8_16_18_20	0.69	0.71
23	regression_randomforest	featureset_logs 1_7_8_16_18_17_13	0.64	0.64
24	regression_randomforest	featureset_logs 1_7_8_16_18_17_14	0.66	0.65
25	regression_randomforest	featureset_logs 1_7_8_16_18_17_15	0.65	0.64
26	regression_randomforest	featureset_logs 1_7_8_16_18_17_19	0.66	0.65
27	regression_randomforest	featureset_logs 1_7_8_16_18_17_20	0.64	0.63
28	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_14	0.63	0.63
29	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_15	0.63	0.62
30	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_19	0.63	0.62
31	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_20	0.63	0.63
32	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_15_14	0.62	0.62
33	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_15_19	0.61	0.61
34	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_15_20	0.62	0.62
35	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_15_19_14	0.60	0.60
36	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_15_19_20	0.60	0.60
37	regression_randomforest	featureset_logs 1_7_8_16_18_17_13_15_19_20_14	0.59	0.59

Table 3: Execution of the forward selection algorithm

In this table we can observe how we add features to the dataset until the validation error stops decreasing. At that point the best feature set is found. We can plot the evolution of the error measure (NRMSE of the cross validation) during the execution of the forward selection algorithm. In Figure.2 We can see the error in each iteration space (between the dashed-orange-lines), in every iteration we choose the feature that decrease the error the most.

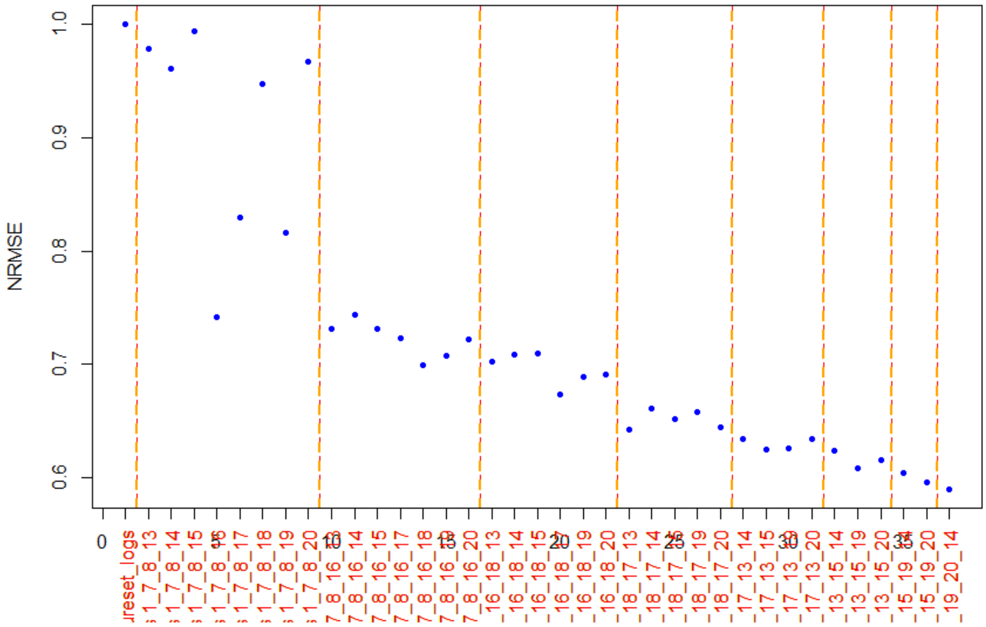

Figure 2: Evolution of the NRMSE for the Sequential forward selection

We used SFS on the two best models, decision tree and lasso. **Decision tree** - Using feature-set without correlated features gives us good results, knowing how decision tree this can be explained by the fact that every split will try to maximize the differences between it's nodes. When we are using features that are not correlate with each other it should help the model do better in it's splitting decision. We start the SFS with 8 features and add features iteratively. Eventually, the validation error dropped from 0.66 to 0.51. **Lasso** - We used dataset with large range of features we extract. Start with the small number of features having validation error of 0.68, after 351 iteration of SFS algorithm, the final number of feature grow significantly and so the validation error dropped to 0.54.

Once we have selected our best 3 models and performed feature selection over the dataset, we can fit the best model with the feature set and see what is its validation error and how it generalizes by the testing error.

6 Conclusion

We start with manual data exploration using unsupervised techniques, checking for outliers and structures in the dataset. In the next step we did feature engineering for creating new features using ratio, log and other transformations. Facing the million dollar question which combination of feature set and model is the best for our problem, we decided to implement sequential forward selection. After many steps trail and error, we came up with the results in Table.4. The most interesting to see, in our point of view, is the comparison between the decision tree and the Random forest. As we can see in the decision tree preform better than random tree. However, this is base on the subset we used from SFS, looking on Table.?? we see that the Random forest model that trained with the complete data set (featureset_logs) preform much better. But when we used SFS not all the features were chosen, that show us that doing SFS is good, but not enough. On the other hand, we see that the decision tree model didn't suffer from that, we assume it because it have less randomness in it's process. About the Lasso regression model, we expected it be worse than Random Forest, but using the best featureset (by forward selection) we found that this is not the case. This is a bit surprising, we can also say the difference is not so significant.

We proposed a series of improvements over the current work:

- test more advanced models: Splines, RVM, Bayesian approach to regression,

- perform a deeper exploration of the possible features and combinations of features
- Explore better feature selection method.

	Model	Feature set	Validation.NRMSE	Training.NRMSE	Testing.NRMSE
1	regression_tree_rpartlib	featureset_regression_rpart_tree_fitting_sfs	0.51	0.45	0.56
2	lasso regression GLMNET	featureset_glmnet_lasso_sfs	0.54	0.53	0.52
3	regression_randomforest	featureset_regression_randomforest_sfs	0.59	0.59	0.59

Table 4: Best 3 models

References

[1]

[2]

[3]

Appendices

A Implementation details

The contents...