Stockholm University
DSV dept.
Academic year 2017/18

Big Data with NoSQL

Final report of the third assignment

Presented by:

- Giorgos Ntymenos
- Giacomo Bartoli

Part I: Data Exploration [2 points]

[Part I-A] What is the size of the vocabulary, i.e. the total number of unique tokens in all documents? 38192
[Part I-B] What are the ten most frequent tokens and their associated frequencies? [(',', 42615), ('the', 41997), ('', 38644), ('.', 37952), ('a', 21342), ('and', 19026), ('of', 18632), ('to', 18161), ('is', 13636), ('in', 11992)]
[Part I-C] What are the ten least frequent tokens and their associated frequencies? [('casseus', 1), ('tune-meister', 1), ('charcoal', 1), ("hall's", 1), ('fortune-hunter', 1), ('signatures', 1), ('bailey', 1), ('anecdotal', 1), ('ever-bemused', 1), ('bickle', 1)]

Part II: Feature Engineering [3 points]

[Part II-A] How many instances are there in the (i) training set, (ii) development set, (iii) test set?
train 685
dev 261
test 235
[Part II-B] What is the class distribution in the (i) training set, (ii) development set, (iii) test set? The class distribution should be described as the percentage of positive examples
train 15,18%
test 17,44%
dev 13,79%

[Part III-C] What is the size of the feature set after feature selection?
14092

Part III: Model Selection [4 points]

[Part III-A] What is the predictive performance, measured using AUC, of the various models? Which model generalizes best? Provide possible explanations for the results.

- DT, maxDepth = 5 AUC = 0.46111111111111114
- DT, maxDepth = 3 AUC = 0.4707407407407407
- DT, maxDepth = 4 AUC =0.46265432098765435

- RF #20 AUC = 0.5207407407407407
- RF #20 (training set) AUC = 0.9393121938302659 [overfitting]
- RF #100 AUC = 0.718148148148148
- RF #100 (union) AUC = 0.702665325622328

Random Forests achieve better performances rather than Decision Trees. To be specific, the best results are given by Random Forest with 100 trees, which has AUC = 0.71.

[Part III-B] Are there indications that the models built with default parameters are overfitting the training data? Provide evidence for your conclusions.

Models tend to overfit when the performance on the training set is better than the validation set. In our example, Random forest made by 20 trees and evaluated using the training set has AUC = 0.93 while the same random forest, but when trained on the development set, had AUC = 0.52. This model clearly overfits because 0.93 > 0.52.
In general, good performances are always between 70% and 80%. We should always be suspicious when we reach more than 90% in performance evaluation.

Part IV: Model Evaluation [1 points]

[Part IV-A] What is the AUC of the selected model on the test set?
AUC = 0.702665325622328
[Part IV-B] Does this performance estimate differ much from the one obtained on the dev set and, if so, what could be an explanation for that?

Performances are a little bit different and we expected this because we evaluate our model on the test set instead of the dev set. Moreover, we trained the data using the union between train and dev set instead of only the train set.