SCIENTIFIC PUBLICATIONS OF THE STATE UNIVERSITY OF NOVI PAZAR
SER. A: APPL. MATH. INFORM. AND MECH. vol. 10, 2 (2018), 75-78.

Machine Learning Models Overfitting and Generalization in Very Big Datasets

M. Koroteev

Abstract: In recent years some big shifts had happened in the field of machine learning due to increasing capabilities in computing power and possibility of big datasets processing. Generative adversarial networks, increasing use of reinforcement learning, artificially generated/expanded datasets (AED), neuroevolutionary methods throw a new light on problems of estimating generalization capabilities. Model selection becomes more subtle process and model interpretation becomes more complicated when not unnecessary. Here we would like to address the question of generalization capabilities of complicated machine learned models due to possible overfitting risk with respect to very big training datasets and computer-assisted model selection process.

1 Introduction

Machine learning is a very big thing in computer science now. It's applications are very diverse. We use it to built different models for many purposes in hope that result will give us some advantage in describing, modeling and predicting real world phenomena.

In order to use any machine learning technique, we need a special set of data, relevant to our goal, to use it to tune the parameters of our model. This training dataset in fact defines how relevant, efficient and correct our learned model will and even can be. So the quality, size, relevance and completeness of data are crucial to the whole machine learning process.

Depending on a specific modeling purpose, our model can end up performing task on the training dataset, being regression, classification, clustering, dimensionality reduction, anomaly detection etc. But the ultimate point of machine learning, being a brunch of artificial intelligence field, is the premise of ability to generalize models, capable of generalization, i.e. the ability to perform the same task on different but similar datasets with comparable efficiency metric values.

Manuscript received June 12 2017; accepted October 21, 2018.

M. Koroteev is with the Department of Data Analysis, Decision Making, and Financial Technology, Financial university under the government of the Russian Federation

2 Traditional generalization metrics

In order to estimate the generalization capabilities of any model M in performing the task T according to some efficiency metric E, we need to test its productivity on the data, that was not used during learning process. So we validate the results that model had shown on the training set (hence the name "validation set", sometimes called "holdout set" [1]. Efficiency on the training and validation sets may vary due to two different reasons:

- systematic factor caused by high bias or high variance of the model;
- sampling error - variance in efficiency as the result of different samples getting in test or train set.

If we generate train and test sets from the original dataset randomly, then contribution of the sampling error will be normal random variable with zero mean. So we can average this effect out by repeating many times the division process and averaging the results.

But often we have metaparameters in our model that need separate training procedure. Or we have many different models to select the best one from [2]. In this case, accuracy score computed on the validation set may end up positively biased exactly like train accuracy is positively biased due to training process [3].

For this case one may use a third portion of data or the test set to estimate model's efficiency through cross-validation after hyperparameter learning and model selection is done. There are many different procedures how to split initial dataset into three parts and how to randomize between iterations [4]. But as a standard the k-fold cross-validation is used.

So, as a golden standard in traditional model evaluation, the test score is used computed with k-fold cross-validation method.

3 Overfitting diagnosis

The main purpose of cross-validation - is fighting overfitting of the model [5]. Addressing overfitting is an important job of an expert, but diagnosing it may be problematic in some cases. The main indicator of overfitting is a big gap between train and test (or validation depending on the method) efficiency estimates: high train accuracy and low test accuracy that is not affected by learning process duration.

This can be shown on a learning curve plot. It shows dynamic change in train and test accuracy with increasing number of used train datapoints with fixed test dataset. This method can give out important information about bias/variance relation in the model and is very useful for diagnosing overfitting. The only disadvantage of this method is that it takes many learning cycles to plot a feasible curve. In the case of a very big dataset this may become an obstacle for deep exploration of model, especially in business applications.

One can argue that big datasets by themselves reduce overfitting risks, so this problem is insignificant. Indeed, model variance can be compensated with more data (which is actually the best advice to avoid overfitting in general). But in the same time, big datasets and increasing computational power give rise to usage of much more variable models. Contemporary deep neural nets can have dozens of layers, hundreds if not thousands of neurons and millions of free parameters. Models like this can express extremely high variance and in the same time can be pretty slow to learn.

4 Generalization in AED

We all know traditional rule of thumb in machine learning: less data lead to overfitting as well as usage of complicated models. In general, one doesn't want to train very variable, complicated models with a lot of free parameters on a small dataset. Of course, this is all relative, and estimating which level of model complexity is appropriate for a particular dataset is completely up to the human decision.

Later advances in reinforcement learning, semi-supervised learning has led to expanding usage of artificially generating, augmenting and expanding datasets for problems that experience high bias due to lack of sufficient data amount. This is nowadays the most popular method for increasing complexity of models in use [6].

With data augmentation there is an open question of its influence on the generalization capabilities. Studies on this topic only begin to emerge [7].

We know for a certain time that enhancing the dataset in particular way can result in training more robust models. For example, if we talk about image recognition problem, introducing gaussian noise to initial image, applying rotational and translational transformations may end up with the classificator, persistent to this kind of transformations in test datasets.

In particular, in the present time we witness great interest in generative models, like generative adversarial neural networks (GANs), that can be used not only to perform traditional machine learning tasks, like classification, but also, provide an instrument to generate life-like data based on special adversarial learning procedure. Recent advances in bidirectional and conditional GANs lead to growing use of enhancing initial real-life datasets with generated samples. This give rise to an open question: how expanding datasets influence generalization capabilities of trained models.

After all that being said, we propose several guidelines for estimating the generalization capabilities of one's models when using any technique for artificially expanding initial dataset:

1. Always use cross-validation for estimating any efficiency metrics. Currently you have no real reason not to. If you have to few data to separate validation and test set, then you don't have enough data for machine learning at all.

2. Never use enhancing methods in your validation or test sets. You need accurate measures that represent how well model will generalize on real data. If you have trained well over one variation of the datapoint, model will likely perform well on other variations. This is not generalization, this is overfitting in its worst.

3. Previous advise implies that you need to perform any enhancing of the data after the separation into train, test and validation sets is done.

5 Conclusion

In this paper we have put the question of how does artificially expanding machine learning datasets influence generalization capabilities of the resulting models.

We expect growing interest in methodology development for avoiding overfitting, and creating guidelines for building more robust models based on the descriptive dataset analysis, expected to result in models less prone to overfitting.

Due to growing complexity of learning models, volumes of datasets and, as a result, bigger computational demand in machine learning, we predict growing significance of overfitting pre-diagnostic methods, based on indirect measurement of overfitting and generalization capabilities of machine learning models especially in big datasets.

References

[1] *Correctly Validating Machine Learning Models – Data science and AI solutions for Fortune 3000*, Data science and AI solutions for Fortune 3000, 02-Oct-2017. [Online]. Available: https://wildfireforce.com/correctly-validating-machine-learning-models/. [Accessed:06-Apr-2018].

[2] K. P. BURNHAM, D. R. ANDERSON, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer Science & Business Media, 2003.

[3] H. LEEB, B. M. PÖTSCHER, *Model Selection, Handbook of Financial Time Series* (T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen, Eds.) Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 889 925.

[4] S. ARLOT, A. CELISSE, *A survey of cross-validation procedures for model selection*, 27-Jul-2009.

[5] T. SHAH, ABOUT TRAIN, *Validation and Test Sets in Machine Learning, Towards Data Science*, [Online]. Available: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7 [Accessed: 06-Apr-2018]

[6] K. Y. YIP, M. GERSTEIN, *Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions*, Bioinformatics vol. 25, no. 2, pp. 243 250, Jan. 2009.

[7] S. C. WONG, A. GATT, V. STAMATESCU, M. D. MCDONNELL, *Understanding data augmentation for classification: when to warp?*, 28-Sep-2016.