Model and Features Selection by Data Driven Analysis in Linear Prediction

Renfei Sun

May 8, 2019

1 Abstract / Introduction

With the high speed development of Machine Learning technology, many of models – from linear regression to Bayes probability, from one layer perceptron to neural network – were discovered or produced by Computer, Math and Engineering Scientists. Especially, after people started to use neural network models, the speed of model developing get a critical increasing. Based on analysis of the development of Convolutional Neural Networks, Recurrent Neural Networks, AutoEncoders, and Deep Learning, we could say that nowadays, Scientists tried to find more powerful model driven by data, instead of driven by the math or statistic logic. Linear Model versus neural network is a good example. When people use linear model, such as linear regression, Bayes probability, or perceptron, they knew the details of the models. People understand that the models are using less square error, max likelyhood, or perceptron to find the appropriate weights. Due to people know the details of the model, they pay much attention on tuning the model by decided a loss function, kernel function, and regularization, but people did not get high accuracy prediction result. Compared with CNN, people do not 100% focus on the logic of model (they call the model a black box), but pay much more attention on the relations in each feature with other features. Therefore, when we select a model to predict our data, we could try to analyze much more on training data. By analysis of the relationship between features and labels, features and features, we could find out a appropriate prediction model.

2 Describe your project in one sentence

Through analysis the data deeply, find out a appropriate model and do a suitable features selection.

3 Who is the audience for this project? How does it meet their needs? What happens if their needs remain unmet?

The outcomes of this project will have great impact to both the academic community and commercial industry, especially the commercial industry. Using data driven model features selection, could improve the accuracy and reasonable of models.

4 What is your approach and why do you think it will be successful?

1. Using python matplotlib to draw the distribution images of enegery usage of data
2. Using python pandas to draw the correlation images of features and labels of data
3. Using python Machine learning library such as Tensorflow or scikit-learn, find a approximate relationship between features land labels.
4. Draw the Residuals Plot to draw the difference between prediction value and actual label value (Residual $= \hat{y} - y$).
5. Analysis the Residuals Plot to check if the model or features are good for using to do the estimation.
6. Based on the analysis, then select the best model and select some good features.

5 In the best-case scenario, what would be the impact statement (conclusion statement) for this project?

The good scenario is that after the Data Driven Analysis, the prediction accuracy could get a significant increment.

6 List all major milestones for this project, and how you intend to spread them throughout the course.

1. Data collections: Find out a regression data.
2. Using python library do estimation without data analysis.
3. Do the data analysis
4. Find out the best model and useful features
5. Re-use the python library to do estimation, and check there is a improvement

7 What obstacles do you anticipate, and how do you plan to address them?

7.1 Major obstacles

(Major obstacles are ones which threaten the viability of the entire project)

Hard to come up with a data useful analysis method.

7.2 Minor obstacles

(Minor obstacles are ones which threaten the timeline estimates)

Might be hard to find out the data.

8 What additional resources do you need to complete this project?

- Some python library to help data relationship analysis

9 List 5 major publications that are most relevant to this project, and how they are related.

- Henri create a system, called Simgrid [2], used for predict the execution time of multiple application by simulation methodology, so we could use them as a baseline in our comparison.

- Farrukh Nadeem set up a model for prediction workflow by machine learning [1]. The model is linear regression model, and his target is grid workflow (applications executed on same time with different nodes). Although our target is applications executed on same time with same node, Farrukh's paper could give up good ideas.

- As for memory usage feature, Williams' roofline modeling [4] could be used for representing memory bandwidth in each application.

- As for usage of power feature, a good starting point is to collect information on last-level cache misses, which has been shown to be useful in predicting effects of changing CPU power on performance when considering only CPU and memory.[5]

- As for the interference by network and I/O be the feature to effect the execution time, Smith built a model of network interference[3] is good to use.

10 When / How do you know if you have succeeded in this project?

When the prediction reach around 80% correctness, we could assume our model is built succeeded. Our prediction does not need to get really high accuracy, such as 95% correctness like face detection programs because the prediction is just a suggestion for application/nodes scheduling of high-performance system. In our new model of the real high-performance system, there are others factor will guide applications/nodes scheduling, such as power save (green computer), network interference, etc.

11 References

[1] Luis, Candanedo M. Data Driven Prediction Models of Energy Use of Appliances in a Low-Energy House. Energy and Buildings, Elsevier, 31 Jan. 2017

[2] THEO GASSER, LOTHAR SROKA, CHRISTINE JENNEN-STEINMETZ, Residual variance and residual pattern in nonlinear regression, Biometrika, Volume 73, Issue 3, December 1986, Pages 625633, https://doi.org/10.1093/biomet/73.3.625

[3] Hamidieh, Kam. A Data-Driven Statistical Model for Predicting the Critical Temperature of a Superconductor. Computational Materials Science, Elsevier, 10 Aug. 2018.

[4] Wang, Zeyu. A Review of Artificial Intelligence Based Building Energy Use Prediction: Contrasting the Capabilities of Single and Ensemble Prediction Models. Renewable and Sustainable Energy Reviews, Pergamon, 10 Nov. 2016.