Capstone Proposal - Machine Learning Engineer Nanodegree

Prarit Agarwal

Friday 12th July, 2019

Contents

1 Introduction and Domain Background

This proposal is based on the Kaggle competition called “Predicting Molecular Properties” [1] hosted by the CHemistry and Mathematics in Phase Space (CHAMPS) group of researchers. The aim is to develop an algorithm to predict strength of interatomic interactions called 'scalar couplings' in any given molecule. The knowledge of such scalar couplings is highly sought after and is extremely helpful in understanding the physical and chemical properties of these molecules [2]. In principle, it is possible to solve these scalar couplings through quantum mechanical computations. These computations involve solving the Schrondinger's equation for a 'many-body-system', a problem that is known to be hard and time consuming owing to its computational intensiveness. At the same time, the scalar couplings are highly constrained by the requirement of invariance under translation and rotation of the molecule. It therefore follows that they can only depend upon interatomic distances, the relative orientation of the atoms and the overall geometry of the molecule along with the different physical properties of the atoms themselves. It therefore follows that it should be possible to develop numerical models for these couplings which would greatly enhance our computational ability without a significant loss in precision when compared to an honest 'quantum mechanical computation'. This is also evident from the fact that chemist have already developed a reliable set of rules (for e.g. see [3] and [4]) which can be used to predict the coupling constants as long as the required precision is not too high. However, in applications such those in the pharmaceutical industry and material science, one often desires a higher precision than might be possible using these rules. Given the nature of the problem above, it is therefore natural to try to apply machine learning tools to this area and see if one can develop better models with more precise predictions.

2 Problem Statement

The CHAMPS group has provided explicit interatomic scalar couplings for 85003 molecules along with information about their 3d molecular structure. Using these to train models, one then has to make predictions for 45772 new molecules. Given that these scalar couplings can in principle take any real value, this is clearly a regression task. As is the case with any regression task, the goodness of the solution can be easily quantified in terms of the error. Additionally, if desired, one can also use the R_2-score to gauge the predictability of the model.

3 Datasets and Inputs

The dataset provided by the CHAMPS group is publicly available through this link. As mentioned in the previous section, it consists of a training set with 85003 molecules and a test set consisting of 45772 other molecules. It has also been ensured that there is no overlap between the training and the test set. For all these molecules, CHAMPS has also provided the spatial coordinates for all the atoms in each molecule. This is to be treated as the only input to the models.

CHAMPS has also provided additional data namely, *dipole moments*, *magnetic shielding tensors*, *mulliken charges*, *potential energy* and *scalar coupling contributions* for the molecules in the training set only. This data is not available for the molecules in the test set. Thus, if one intends to use this information, then they will have to separately model them first and then use them as meta-features for the molecules in the test set.

Additionally, the scalar couplings to be predicted are of 8 distinct types: 1JHC, 1JHN, 2JHC, 2JHN, 2JHH, 3JHC, 3JHN, 3JHH. This information is also provided explicitly for both the training and test molecules. The values for scalar couplings of each type seem to have slightly distributions.

4 Solution Statement

Given that the explicit values for the scalar couplings of different types have different distributions, I propose to train an individual model for each type. I also wish to use neural networks to create my models. As mentioned in section 1, translational and rotational invariance requires that the molecular properties should only depend upon the interatomic distances and angles. This can easily be extracted given the spatial coordinates of all the atoms in each molecules. Having done this, I will feed this information into my model and train them appropriately, using regularization techniques such as batch-normalization, dropout and early stopping.

At the same time, we wish to apply ensembling techniques to improve my predictions. To this end, for each type of scalar coupling, I will train a couple of different neural networks (with different configurations/hyper-parameters) and take the weighted average of their predictions as my final prediction. The weights to be applied to each individual model will be based on their R_2-score for the validation set i.e. the predictions from the model with a higher R_2-score will be given a proportionately higher weight.

5 Benchmark Model

A benchmark model has already been suggested by the competition sponsors. This can be found here. It is based on engineering a single feature i.e. the distance vector between the atoms involved in the coupling.

6 Evaluation Metric

As mentioned here, submission are evaluated based on the log of the mean absolute error for each scalar coupling type type, which is then averages over all types. More explicitly, it is given by

$$\text{score} = \frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{1}{n_t}\sum_{i=1}^{n_t}|y_i - \hat{y}_i|\right) \tag{6.1}$$

where t runs over the different types of scalar couplings, i runs over the number of instances in the corresponding type and

- T = number of different scalar coupling types
- n_t = number of instances of type t
- y_i = prediction for the i-th instance
- $\hat{y}_i$ = true value for the i-th instance

7 Project Design

The project will proceed according to the following workflow:

1. EDA: A preliminary analysis of the data
2. Feature Engineering: Use the spatial coordinates of the atoms to extract their relative distance and spatial orientation
3. Benchmarking: Train the benchmark model mentioned in section 5 to establish a baseline.
4. Modeling: For each type of scalar coupling, construct several different DNNs that attempt to the predict the scalar coupling constants. These models will differ from each other in the number of their hidden layers, type of regularization techniques (such as batch-normalization vs dropout) etc. We will also use early stopping to prevent overfitting the training set.

5. Ensembling: Having developed a couple of different models for each type, we will take a weighted average of their predictions to obtain our final predictions.

6. Submission: We will then use our model to predict the scalar couplings for the test set provided by CHAMPS and submit these to Kaggle to obtain a final evaluation and score.

References

[1] CHAMPS-Kaggle, "Predicting Molecular Properties." https://www.kaggle.com/c/champs-scalar-coupling.

[2] P. E. Hansen, *Carbon—hydrogen spin—spin coupling constants*, *Progress in Nuclear Magnetic Resonance Spectroscopy* **14** (1981) 175 – 295.

[3] H. J. Reich, "Spin-Spin Splitting: J-Coupling." https://www.chem.wisc.edu/areas/reich/nmr/05-hmr-03-jcoupl.htm/.

[4] Y. Rubin, "Coupling constants for ^{1}H and ^{13}C NMR." https://yvesrubin.files.wordpress.com/2011/03/coupling-constants-for-1h-and-13c-nmr.pdf.