Machine Learning Engineer Nanodegree Capstone Project

Vasileios A. Tatsis

May 31, 2018

Abstract

Aqueous solubility is considered to be one of the important properties in drug discovery. In this work, the performance of several machine learning methods in inferring the aqueous solubility of a small organic molecules' is evaluated.

I Definition

Project Overview

Solubility is the property of a solid, liquid, or gaseous chemical substance called solute to dissolve in a solid, liquid, or gaseous solvent to form a homogeneous solution of the solute in the solvent. The solubility of a substance fundamentally depends on the solvent used as well as on temperature and pressure. The extent of solubility of a substance in a specific solvent is measured as the saturation concentration where adding more solute does not increase its concentration in the solution.

Aqueous solubility prediction is important in drug discovery. Early identification of a compound with poor aqueous solubility in a drug discovery cascade can minimize the risk of failure. Currently, I am working in the drug discovery field and I am very interested in using predictive models for the design or optimization of small organic drug molecules.

The aqueous solubility of a drug is a significant factor for its bioavailability. Many drugs are administered via the oral route, their absorption and metabolism in organisms are closely related to its aqueous solubility.

A model capable of accurately predicting the aqueous solubility of small organic molecules will be very useful in the Drug Discovery field. In this work, I will try to compare different machine learning models in predicting the aqueous solubility of small organic molecules. As a sole descriptor I will use the compounds' structural fingerprints. Other helpful descriptors that could be used, could be the physicochemical properties (polar surface area, logD, logP) or the molecule's geometric features.

For the training and testing of the machine learning models, I will employ the Huuskonen data set (ref. 1). This data set contains 1297 organic molecules by Jarmo Huuskonen (ref. 2). Molecules are listed in smiles format along with their aqueous solubility values, expressed in log mol/L at 20-25 degrees, and their octanol-water partition coefficient values (logP).

I will use as input to the machine learning models only the structural information (2D graph) from this data set to infer the aqueous solubility of the compounds contained in the data set. *The simplified molecular-input line-entry system (SMILES) is a specification in form of a line notation for describing the structure of chemical species using short ASCII strings* (Wikipedia). For example, cyclopropene is usually written $C1 = CC1$.

For this purpose the chemical structure information will be converted into feature vectors of fixed length (fingerprints). Before training the machine learning models, the available fingerprints in RDKit will be tested in order to select the one that can represent chemical structure with a minimum loss of information. Molecular fingerprints encode molecular structure in a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule. But, fingerprints do not include full structural data (such as coordinates). The cross-validation technique will be also employed.

RDKit, an open-source cheminformatics Python library, will be used to drive the conversion of the chemical structure into feature vectors of fixed length (structural fingerprints).

Problem Statement

The aim of this project is to evaluate the performance of several machine learning models in predicting the aqueous solubility of small organic molecules. Aqueous solubility is a critical property since it is related to the absorption and metabolism of oral drugs in human body. In this study, a public data set containing the 2D structures (in SMILES format) of small organic molecules and their aqueous solubility values will be used as input. I will use as input to the machine learning models only the structural fingerprints. An analysis of the data set will be the first step in this study. I will try to map map the chemical space in terms of physicochemical properties and explore how chemical diverse is the data set. Also, I will try to highlight possible implications (if any) of the chemical diversity in the predictability of the machine learning models and if there are any outliers. Several chemical fingerprints (Estate, Morgan, RDKit, Topological torsion, Extended reduced graph approach, Avalon) will be tested and their performance will be assessed using the Bayesian Ridge model. Grid searches to tune the hyperparameters of the machine learning models like RandomForestRegressor, KernelRidge, and GaussianProcessRegressor will be employed. Simple learning models like PLS will be employed. Furthermore, models like XGBoost and LightGBM will be evaluated in this work.

Metrics

I will use the following five metrics to assess and to compare the performance of the machine learning models: i) Pearson correlation coefficient (R), ii) mean absolute error (MAE), iii) root mean square error (RMSE), iv) coefficient of determination (R^2), and iv) explained variance score (EVS) defined by:

$$
\begin{aligned}
R &= \frac{\sum_{i=1}^{n}(t_i - \bar{t})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n}(t_i - \bar{t})^2}\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2}} \\
MAE &= \frac{1}{n}\sum_{i=1}^{n}|t_i - p_i| \\
RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{i}(t_i - p_i)^2} \\
EVS &= 1 - Var(p_i - t_i)/Var(t_i)
\end{aligned}
\tag{1}
$$

where t_i is the target value (experimentally measured) and p_i is the predicted value for a compound (data point) i. Furthermore, Var is biased variance, i.e. $Var(p_i - t_i) = sum(error^2 - mean(error))/n$.

II Analysis

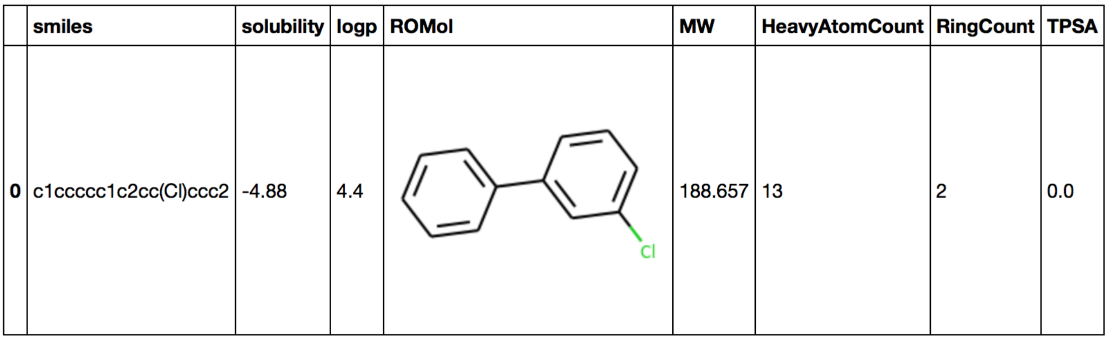
(approx. 2-4 pages)

Data Exploration

To train and test the machine learning algorithms, I use one publicly available benchmark data set widely used in the solubility prediction literature. This dataset consists of 1297 organic molecules from the AQUASOL database and the PHYSPROP database. Molecules are listed together with their aqueous solubility values, expressed in log mol/L at 20-25°C. In Figure 1, I have highlighted the first data point in the dataset.

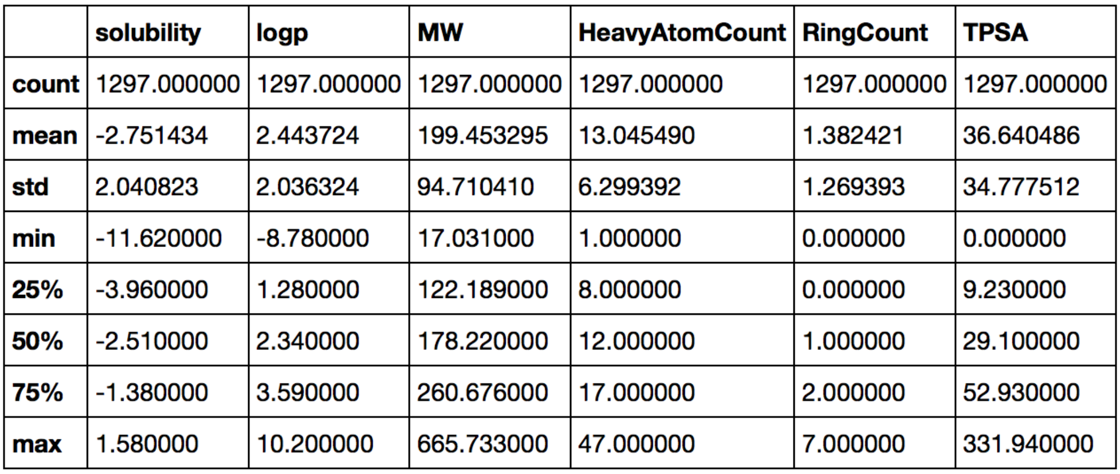
Also, in Figure 2 there is a table summarizing the statistics of the available dataset. On average, the compounds in the dataset have 13 heavy atoms, molecular weight of 200 and polar surface area of 36 $\mathring{A}$. From this table, I notice that there are compounds with high molecular weight and very high polar surface area.

Before applying the machine learning methods, I tried to scan the available chemical space to check if there are any outliers, in terms of physicochemical properties (compounds that may have high molecular weight, high polar surface area) Figure 3. Indeed, there are some very interesting compounds with many polar groups and quite rigid 3D structure.



	smiles	solubility	logp	ROMol	MW	HeavyAtomCount	RingCount	TPSA
0	c1ccccc1c2cc(Cl)ccc2	-4.88	4.4		188.657	13	2	0.0

Figure 1: Data point example from dataset.



	solubility	logp	MW	HeavyAtomCount	RingCount	TPSA
count	1297.000000	1297.000000	1297.000000	1297.000000	1297.000000	1297.000000
mean	-2.751434	2.443724	199.453295	13.045490	1.382421	36.640486
std	2.040823	2.036324	94.710410	6.299392	1.269393	34.777512
min	-11.620000	-8.780000	17.031000	1.000000	0.000000	0.000000
25%	-3.960000	1.280000	122.189000	8.000000	0.000000	9.230000
50%	-2.510000	2.340000	178.220000	12.000000	1.000000	29.100000
75%	-1.380000	3.590000	260.676000	17.000000	2.000000	52.930000
max	1.580000	10.200000	665.733000	47.000000	7.000000	331.940000

Figure 2: Dataset's statistics.

```
There are 15 compounds with high MW (>550) and TPSA (>130) values!
```

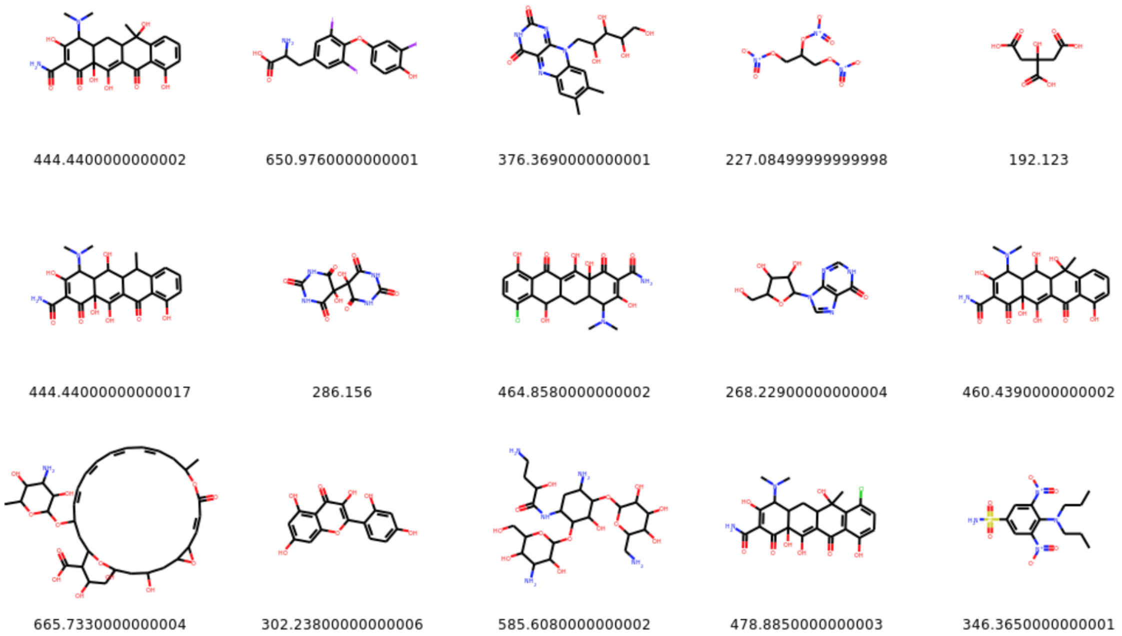


Figure 3: 15 compounds having either high molecular weight (>550) or high polar surface area (>130) were found in the available dataset.

Exploratory Visualization

In this section, I provide a form of visualization that summarizes the chemical space covered by the compounds found in the dataset used in this work. Figure 4 presents the distribution of solubility, logP, molecular weight, polar surface area, and the number of heavy atoms for the compounds used in this work. From this figure, we can notice that there are some apparent correlations between logP and solubility, molecular weight and number of heavy atoms. These correlations are well expected. The target property, solubility, has a quite wide distribution from -11.6 to 1.58 log mol/L and a peak at -2.75 log mol/L. In terms of molecular weight, most compounds seem to quite small with molecular weight close to 200 and spanning from 17 to 665.

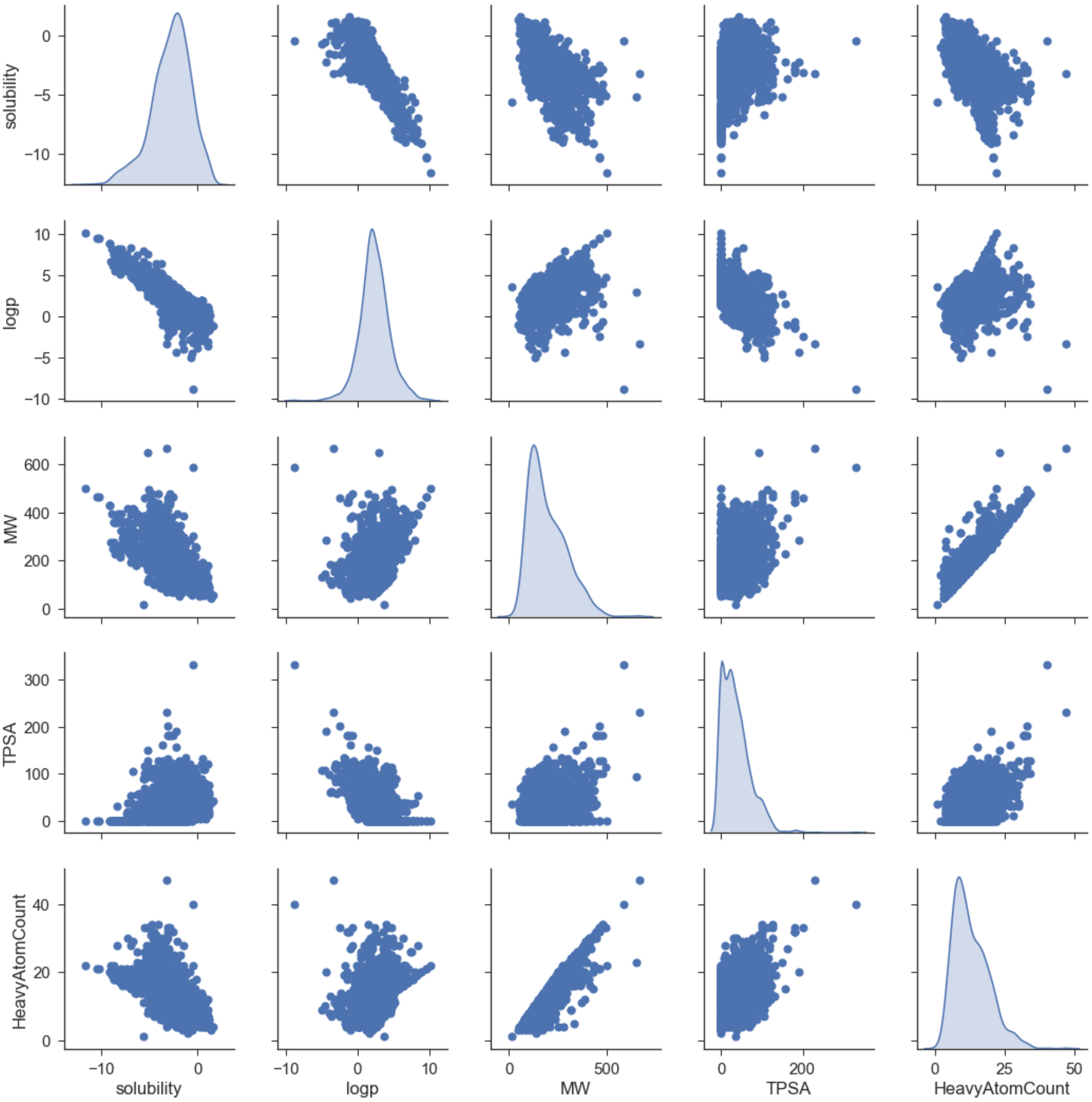


Figure 4: Mapping chemical space in terms of physicochemical properties.

Algorithms and Techniques

The following machine learning algorithms were used:

- Linear Regression
- Random forest
- Kernel Ridge Regression

- Guassian Process Regressor
- Support Vector Regression
- Regression based on k-nearest neighbors
- Multi-layer Perceptron regressor
- Gradient Boosted Trees
- Another flavour of gradient boosting: LightGBM
- More gradient boosting: XGBoost

I selected linear regression and results from literature (ref. 4) as benchmarks for the performance of the machine learning models used in this work. Frohlich et al. (ref 4), using the same data set, report a squared correlation coefficient of 0.90 for an 8-fold cross-validation, using support vector machines with a radial basis function kernel.

III Methodology

(approx. 3-5 pages)

Data Preprocessing

Input data consists of smiles (see ref. 5) strings, which encode molecular graphs into a string of characters. We would like to transform these molecular graphs in a form that a machine learning model could "understand". In this work I used the fingerprint approach. Fingerprints were first used in molecular substructure search: finding a compound with a certain substructure in a molecular database. Fingerprinting creates an efficient representation of the molecular graph. The basic process of fingerprinting is as follows, first the algorithm generates a set of patterns. For instance, enumeration of different paths is common: For example (taken from Daylight's site), the molecule OC=CN would generate the following patterns:

- 0-bond paths: C / O / N
- 1-bond paths: OC / C=C / CN
- 2-bond paths: OC=C / C=CN
- 3-bond paths: OC=CN

Fingerprints having a length of 1024 bits were generated. Each bit helps encode a part of the molecule, such as atom environments along a path with Morgan fingerprints for instance, or number and type of functionalities with MACCS fingerprints. If the number of bits is reduced in amount we risk two different atom environments being encoded with the same bits. Thus two molecules could be identified as more similar than they actually are. More info about chemical fingerprints can be found in ref. 6. RDKIT (rdkit.org), an open source cheminformatics library, is used for generating the fingerprints.

The performance of the selected fingerprints were compared employing a Bayesian Ridge regression model (http://scikit-learn.org/stable/modules/linear_model.html#bayesian-ridge-regression) and 20 fold cross validation (https://en.wikipedia.org/wiki/Cross-validation).

Implementation

All the algorithms and metrics used in this study were imported from scikit-learn and RDKit (rdkit.org). No complications were observed in their implementation.

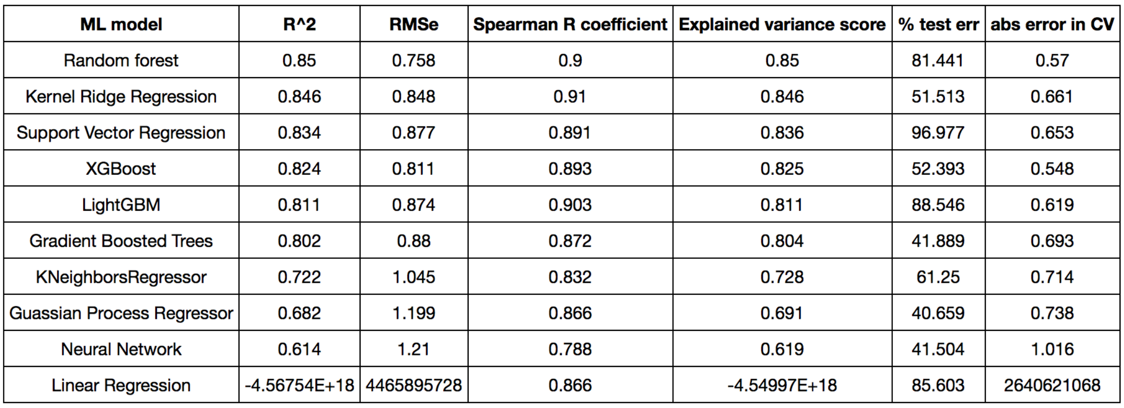
Refinement

Before training and testing the machine learning models, I applied grid search to explore the hyperparameters of RandomForestRegressor, LightGBMRegressor, GaussianProcessRegressor, and Kernel Ridge models. The hyperparameters that resulted from the grid search were used in the production phase.

IV Results

Model Evaluation and Validation

I used several machine learning models in order to test which model could perform better in infering the aqeuous solubility of small organic drug-like molecules. Estate fingerprints were used as the only feature to train the model and the target value was the aqueous solubility of the compounds found in the dataset. In Figure 5 the scores and metrics of the different machine learning models that were used are summarized.



ML model	R^2	RMSe	Spearman R coefficient	Explained variance score	% test err	abs error in CV
Random forest	0.85	0.758	0.9	0.85	81.441	0.57
Kernel Ridge Regression	0.846	0.848	0.91	0.846	51.513	0.661
Support Vector Regression	0.834	0.877	0.891	0.836	96.977	0.653
XGBoost	0.824	0.811	0.893	0.825	52.393	0.548
LightGBM	0.811	0.874	0.903	0.811	88.546	0.619
Gradient Boosted Trees	0.802	0.88	0.872	0.804	41.889	0.693
KNeighborsRegressor	0.722	1.045	0.832	0.728	61.25	0.714
Guassian Process Regressor	0.682	1.199	0.866	0.691	40.659	0.738
Neural Network	0.614	1.21	0.788	0.619	41.504	1.016
Linear Regression	-4.56754E+18	4465895728	0.866	-4.54997E+18	85.603	2640621068

Figure 5: Scores and metrics of the ML models used for predicting aqueous solubility.

Random forest is the best performing model for predicting aqueous solublity with an R^2 value of 0.85 and the lowest root mean squared error (0.76). Random forest algorithm has been used widely in chemoinformatics and it is a well established model in this field. Kernel Ridge regression follows very closely with high R^2 value (0.846) and slightly higher rmse value (0.85). Kernel Ridge regression outputs a low mean absolute percentage error value (51.5) for the test set. Gradient Boost trees reports the lowest mean absolute percentage error value (41.9).

Furthermore, Random Forest yields similar R^2 value compared to a support vector machine model reported in literature. Frohlich et al. [4], using the same data set, report a squared correlation coefficient of 0.90 for an 8-fold cross-validation, using support vector machines with a radial basis function kernel.

Figure 6 displays the experimental and predicted aqueous solubility values for the training and test using the Random Forest model. Comparing the Random Forest plot with the second best model, which is Kernel Ridge (Figure 7), I notice that Random Forest appears to overfit in the training data.

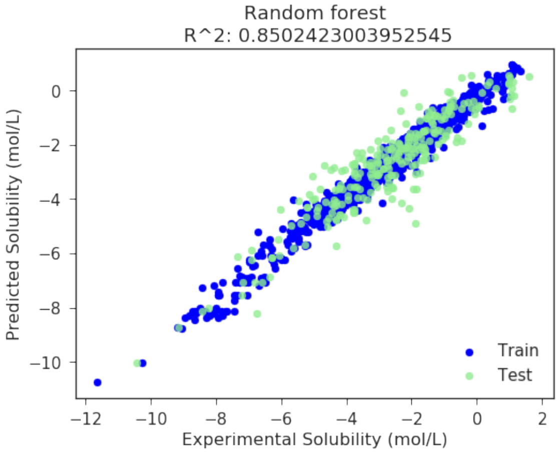


Figure 6: Results of training and test set applying the Random Forest model.

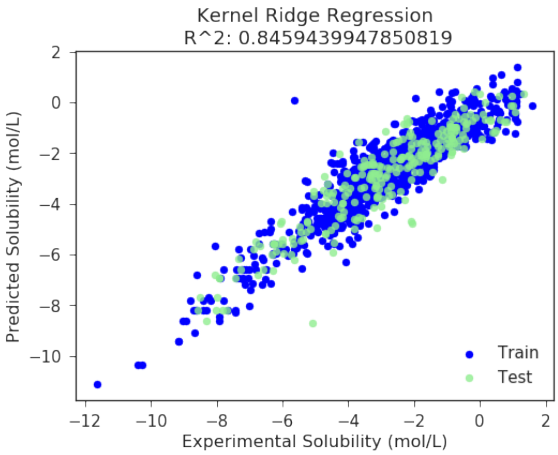


Figure 7: Results of training and test set applying the Kernel Ridge Regression model.

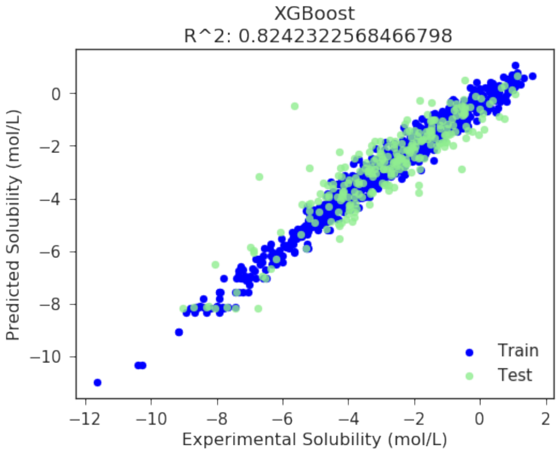


Figure 8: Results of training and test set applying the XGBoost model.

Furthermore, altough Gradient Boosted trees and XGBoost belong to the same family of algorithms, they appear to perform very differently (see Figures 9 and 8).

Also, I find very interesting the fact that linear regression performs very poorly in this case. Possibly, the use of fingerprints as descriptor is the cause of this poor performance.

Justification

I selected to use linear regression as a benchmark model, but unfortunately its performance was really low. An issue that needs to be examined and I would like to study this low yield in future work. Also, I found in literature that Frohlich et al. [4], using the same data set, report a squared correlation coefficient of 0.90 for an 8-fold cross-validation, using support vector machines with a radial basis function kernel. Random Forest yields an R^2 value of 0.85 which is close to the squared correlation coeffient reported in literature. I have to note also that there were five more models that resulted similar R^2 value to Random Forest.

V Conclusion

Reflection

I have assessed ten machine learning models in inferring the aqueous solubility of a dataset containing small organic molecules. Estate fingerprints were used to describe the chemical space covered by the dataset's compounds. Random Forest was picked as the best performing model

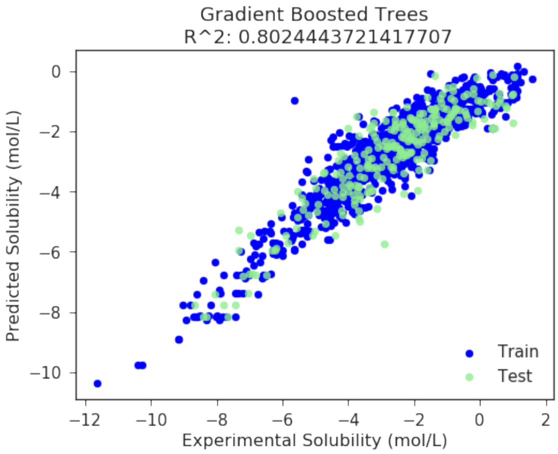


Figure 9: Results of training and test set applying the Gradient Boosted trees model.

with five other models yielding quite similar results. Random Forest also exhibits the same behaviour with another model found in literature [4]. In a real-life project scenario I will probably use the top six performing models and not only Random Forest to predict the aqueous solubility of small organic molecules. Another important issue is the domain of applicability, it is evident that these models can predict the solubility of a compound only if they have been trained with similar compounds of functional groups (hydroxyl, carbonyl, furan, etc.). It would be quite hard and extremely interesting project to build a global solubility model.

Improvement

For future work, I would like to examine the fine-tuning of the machine learning models' parameters. Also, I would be interested in comparing the performance of deep learning models and shallow learning models. Another aspect of this project that seeks thorough investigation is the linear regression's poor performance. Could it be improved by using more (2D and 3D) descriptors and/or by using longer fingerprints ($>$ 1024 bits)? How would these changes impact the performance of the other models that were used in this work? Furthermore, I would like to employ a larger and more diverse dataset like the one found in ChEMBL (https://www.ebi.ac.uk/chembl/) or in other datasources.

References

1. http://cheminformatics.org/data sets/

2. SAR and QSAR in Environmental Research, 2008, 19 (3-4), 191-212

3. J. Chem. Inf. Model., 2013, 53 (7), 1563–1575

4. QSAR Comb. Sci., 2004, 23, 311-218

5. https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

6. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html

7. http://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics