Permeability prediction

Bochkarev Artem

Moscow Institute of Physics and Technology
Faculty of Control and Applied Mathematics
Department of Intellectual Systems
Supervisors: Sofronov I., Strijov V.

17 May 2016

Purpose

Permeability is one of the crucial parameters of core samples. It's measurements are relatively expensive, take much time and require special equipment. In our study we try to predict permeability using other core information, such as porosity and density. We will show and compare different approaches to its prediction. One of the main goals of this research is to find interpretable model for this problem.

Problem statement

Data consists of measurements from approximately 2km depth, with 230 samples of core and 30 features for each of them. As we want to predict only permeability, we take 83 samples on which it is defined. Also we use only 4 main features, namely horizontal thermal diffusivity, horizontal thermal conductivity, density and porosity. All of them are measured on the dry rock.

Formal statement

Let $\mathbf{X} = x_{ij}$ be the given data, $\mathbf{Y}$ - vector of permeabilities. We want to find function $f : \mathbf{X} \to \mathbf{Y}$, which minimizes

$$\min_f \sum_{i,j} ||f(x_{ij}) - y_i||_2$$

Experiment

In order to solve this problem, we tested several standart machine learning algoritms. Before all experiments we preprocessed our data: normalizing all features (mean equals 0 and standart deviation equals 1). We will also be working with logarithm of permeability, which has mean of -0.14 and standart deviation of 0.88.

Algoritms

The algoritms we are testing: 2-layer neural network, ridge regression, random forest, gradient boosting, kNN

Results

Algorithm	Mean error	Standart deviation
2-Layer neural network	0.47	0.08
Ridge regression	0.71	0.11
Random forest	0.73	0.18
Gradient boosting	0.77	0.22
kNN	0.67	0.15

Genetic algorithm

Next, we try to find our model using symbolic regression. We try to build different trees, representing mathematical expression, next we search the space of all functions using genetic algorithm.

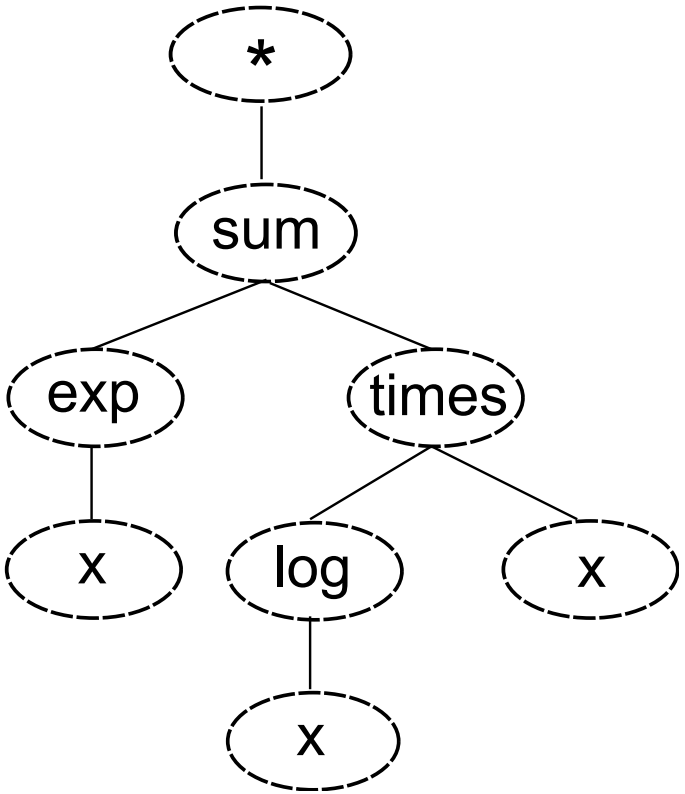


Figure: Tree example

Genetic algorithm

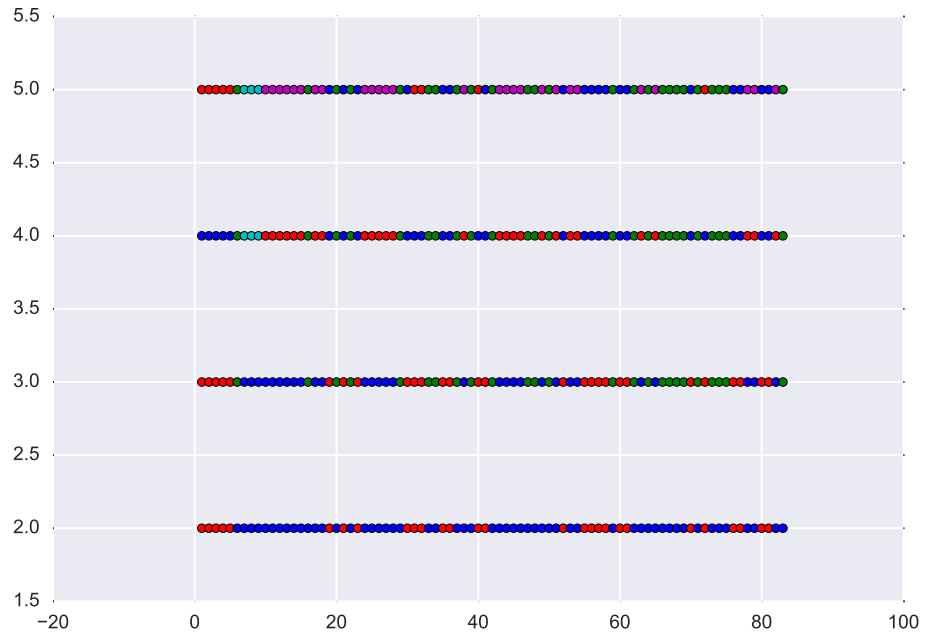
In our experiment, we used following parameters:

- Population size — 10000
- Number of generations – 30
- Crossover probability – 0.5
- Mutation probability – 0.4
- Selected parents – 500

The algorithm with these parameters correctly recognizes functions such as $y = x_0 + x_1 \cdot x_2 - 3 \cdot e^{x_3}$. On our data it gives error of 0.69, which is greater that we want to. The next idea is to make clusterisation of data first.

Clusterisation

We made clusterisation using three different approaches: k-means, hierarchical clustering and spectral clustering. The result is very similar.



We manually selected one cluster which is consistently present using different parameters.

Selected cluster

We run genetic algorithm on this cluster 5 times, with the same parameters which we used in our previous experiment. You can see the results in the table.

Function number	Function	Error
1	$\sin({x_3}^4) - \sin(x_0 - 1)$	0.36
2	$sin(4 \cdot {x_2}^9)$	0.54
3	$\sin(\sqrt{7}^{(x_0+x_2)})$	0.54
4	$\cos(\log(x_1) \cdot (x_0 - 8))$	0.56
5	$\sin(\exp x_0 \cdot {x_2}^4)$	0.59

Conclusion

- Conducted several experiments with different algorithms
- Made clustering, using different approaches
- Found interpretable model for one of the clusters