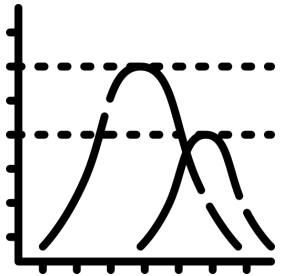
Classifying Pump Sizes with Gaussian Process Regression

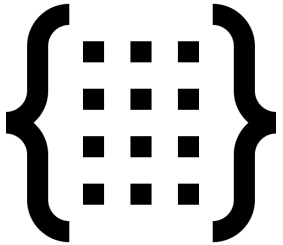
PARISlab Research Project
Nick Monozon

Overview



Gaussian Process Regression

Field Dataset



Model Performance

What is Gaussian Process Regression?

Gaussian Process Regression

- Defines a probability distribution of possible functions
- Each observation follows a Gaussian (normal) distribution
  - Every subset of of n observations is an n-variate normal distribution
- Kernel is an n-by-n covariance matrix for each pair of observations
  - Each $(i,j)^{th}$ entry is the influence between the i^{th} and j^{th} points
- Train to determine joint probability density function $\mathbb{P}_Y$
  - Uses Bayesian inference to construct Gaussian posterior $\mathbb{P}_{X|Y}$ from Gaussian prior $\mathbb{P}_{X,Y}$ with testing data
  - Possible functions determined by kernel

Visualization of Kernel Hyperparameters

https://distill.pub/2019/visual-exploration-gaussian-processes/

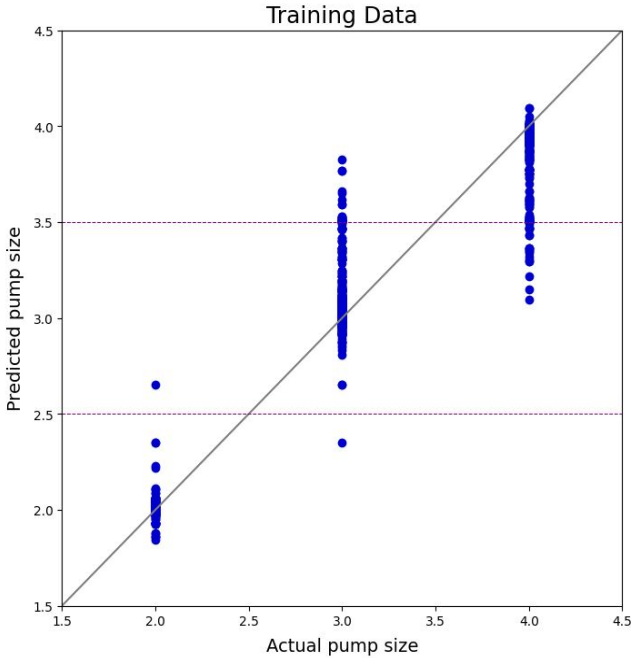
The dataset

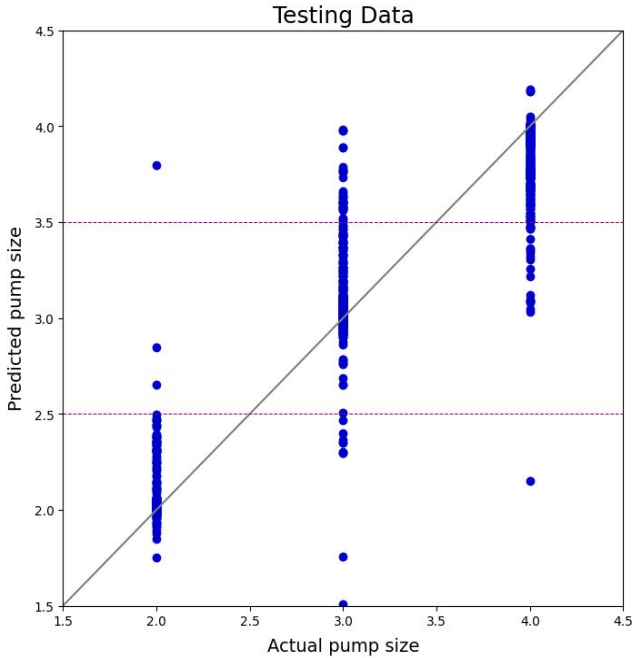
	w/cm	cementitious	Coarse agg total	Fine agg total	Coarse agg (>1/2")	Coarse agg (<1/2")	Pump size
0	0.47	602.0	1525.0	1721.88	0.0	1525.0	2
1	0.47	635.0	1250.0	1701.90	0.0	1250.0	2
2	0.40	752.0	1250.0	1565.18	0.0	1250.0	2
3	0.52	611.0	1250.0	1649.95	0.0	1250.0	2
4	0.41	750.0	1050.0	1919.73	0.0	1050.0	2
...	...	...	...	...	...	...	...
5871	0.54	510.0	1700.0	1683.30	1700.0	0.0	4
5872	0.55	500.0	1700.0	1692.29	1700.0	0.0	4
5873	0.55	484.0	1700.0	1740.93	1700.0	0.0	4
5874	0.55	500.0	1675.0	1556.35	1675.0	0.0	4
5875	0.55	486.0	1700.0	1538.18	1700.0	0.0	4

5747 rows × 7 columns

Model details

- Linear combination of 3 different kernels
  - Radial basis kernel (RBF)
    - Stationary (covariance based on relative position)
    - Optimized length scale and variance
  - Linear kernel
    - Preserves (predicted) linear trend in observations
  - White noise kernel
    - Introduces error term to reduce overfitting
- SciPy optimizer to minimize training MSE





Accuracy: 95.70%

Accuracy: 92.92%

Conclusions

- High accuracy for training data and comparable accuracy for testing data
- Higher accuracy than with previous neural network
- Pump sizes of 2 are imbalanced compared to 3 and 4
  - Potential impact on model performance that could be addressed in the model construction
- SHAP analysis