# Differential Nearest Neighbors Regression

Youssef Nader
Leon Sixt
Tim Landgraf

Freie Universität Berlin, Department of Mathematics and Computer Science,
Institute of Computer Science

# What is DNNR?

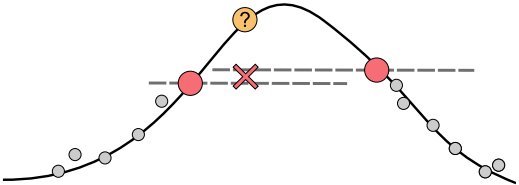


**Figure 1:** Illustration of KNN Regression

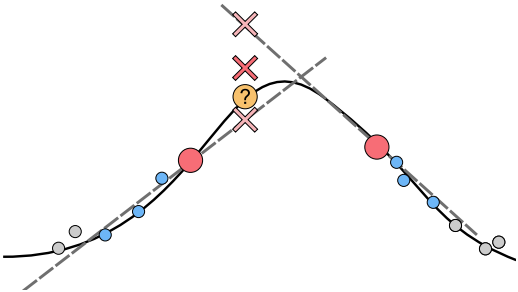


**Figure 2:** Illustration of DNNR

$$\eta_{\text{KNN}}(x) = \frac{1}{k} \sum_{X_m \in B_{x,\#k}} Y_m.$$

$$\eta_{\text{DNNR}}(x) = \frac{1}{k} \sum_{X_m \in B_{x,\#k}} (Y_m + \hat{\gamma}_m(x - X_m))$$

# DNNR Training

- DNNR Learns a matrix W

$$d(x_i, x_j) = (x_i - x_j)^T W (x_i, x_j)$$

- Optimization Objective:

$$W^* = \arg\min_W \sum_{i,j \in I} \text{cossim}\left(d(X_i, X_j), \left| Y_i - \hat{\eta}_{\text{DNNR}}(X_{\text{nn}(i,k')}) \right| \right),$$

- Optimize using SGD

# DNNR Theoretical Bounds

with probability at least $1 - \delta$:

## DNNR Pointwise error

$$\varepsilon_{\text{DNNR}} = h_{\text{DNNR}}^2 \vartheta_{\max} (1 + \tau)$$

## KNN Pointwise error

$$\varepsilon_{\text{KNN}} = 2\vartheta_{\max} h_{\text{KNN}}$$

- where $\tau$ represents the error for approximating the gradient in DNNR

$$\tau = \left[ \frac{\sqrt{\sum_{i=1}^{m} ||\nu_i||_1^{2\mu}}}{\sigma_1} \mid X \in B_{x,h} \right]$$

- the error tolerance of DNNR will be lower than for KNN As long $\tau < \frac{2}{h_{\max}} - 1$,

# Base Datasets Benchmark

- 8 real world small and medium sized datasets
- evaluation against 11 models, including Catboost and Tabnet
- DNNR and DNNR 2nd order achieve best performance on 3 datasets
- DNNR is within 5% difference of the best performing on 2 other datasets

# Feynman Benchmark

- 119 datasets sampled from classical and quantum physics equations
- 100k datapoints each

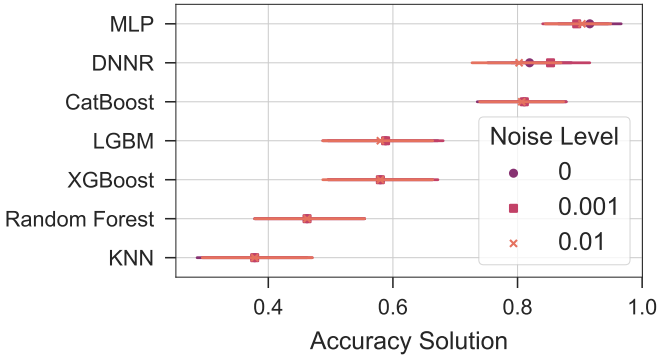


**Figure 3:** Accuracy on the Feynman Symbolic Regression Database under three levels of noise. The marks show the percentage of solutions with $R^2 > 0.999$ . The bars denote $95\%$ confidence intervals.

# PMLB Benchmark

- 133 datasets
- mixture of real world and synthetic datasets

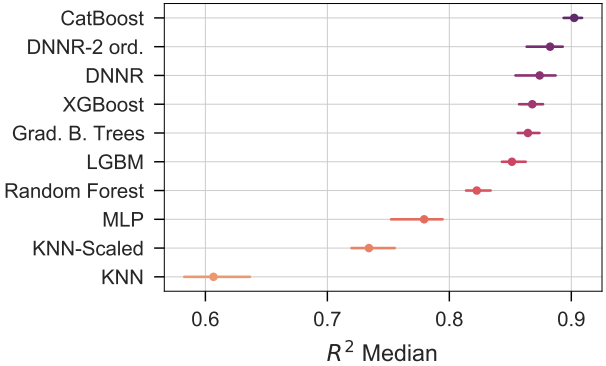


**Figure 4:** Results on the PMLB benchmark. The markers show the median $R^2$ performance over all datasets runs. Horizontal bars indicate the 95% bootstrapped confidence interval.

# Qualitative studies

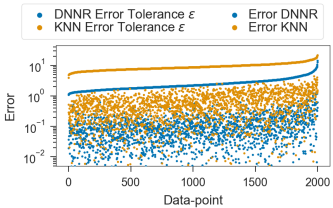


**Figure 5:** Application of theoretical bounds, comparison between the error bound of KNN (yellow) and DNNR (blue).

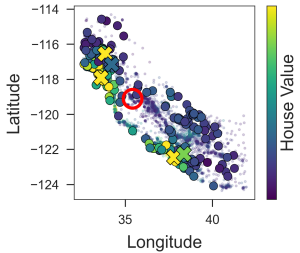


**Figure 6:** Analysis of Failure in DNNR prediction

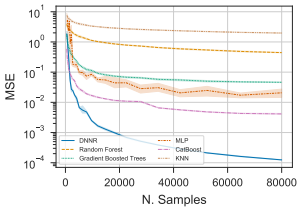


**Figure 7:** Effect of the number of samples on the different models