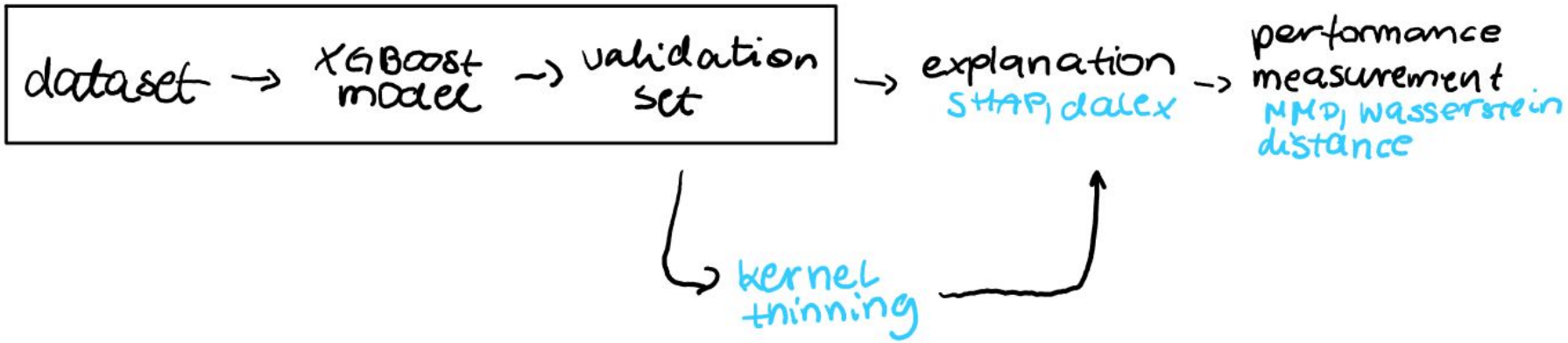
Data compression for improved explanation estimation

Mateusz Biesiadowski, Paulina Kaczyńska, Ania Semik

Pipeline


dataset
XGBoost model
validation set
explanation
SHAP, dalex
performance measurement
MMD, Wasserstein distance
kernel thinning

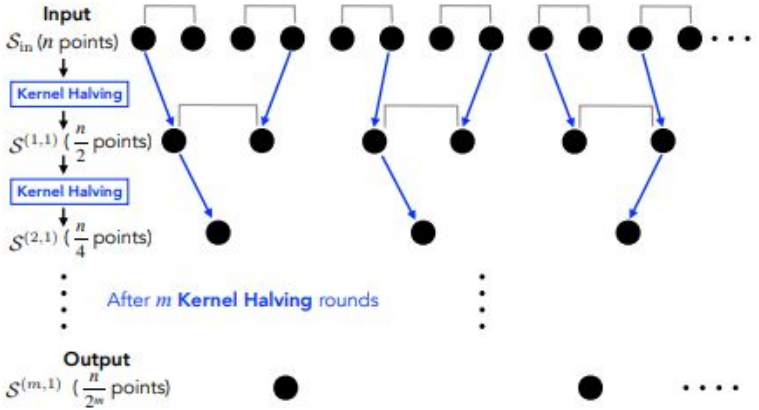
Problem: i.i.d sampling or Monte Carlo Chain Sampling takes too long on big datasets

Earlier solution: standard thinning - taking only every *k*th point - is too inaccurate

Goal: Better thinning - having more accurate distribution that takes less time to sample from

Kernel Thinning

1. Kernel Split



2. Kernel Swap

Take the best coreset $S^{(m,l)}$ according to the Maximum Mean Discrepancy with the original set of points

Iteratively swap the point in it for the best point in the original set of points S_{in} according to the Maximum Mean Discrepancy

Dwivedi, R. & Mackey, L.. (2021). Kernel Thinning. Proceedings of Thirty Fourth Conference on Learning Theory, in Proceedings of Machine Learning Research 134:1753-1753 Available from https://proceedings.mlr.press/v134/dwivedi21a.html.