Statistical Learning June, 2019
Sharif University of Technology CE 40-956
H.R. Rabiee

Quiz 8 Solution (Interpretable Learning)

Solutions

1. We want to interpret the behavior of the model in the locality of that sample. We can select an arbitrary number of samples from this locality and run the predictive model on them to obtain the results. Now we have a set of (x, y) pairs by which we can train another model. We can select a linear model and train its weights by the train set that we constructed. Because these samples are obtained by the predictive model, these two models have a same behavior and weights of the linear model are a good estimation of the weights the black box model gives to features.