HW6

March 18, 2024

Problem 6.1 (California Housing Price Prediction)

a. Load the California Housing dataset, which is part of `sklearn.datasets` module. We wish to use all features to predict the target `MedHouseVal` and to do so, we train and evaluate CART and AdaBoost (with CART as the estimator) using the standard 5-fold cross-validation with shuffling. To create a CART regressor anywhere that is needed, only set the minimum number of samples required at each leaf to 10 and leave all other hyperparameters as default. For the AdaBoost, set the number of regressors to 50 (`n_estimators`). In the cross-validation (`KFold`), CART, and AdaBoost, set `random_state=42`. Compute and report the estimates of the MAE (mean absolute error) and the R^2 using CV for the CART and the AdaBoost. Which case is showing the best performance in terms of $\hat{R}^2$?

b. Use the same dataset but this time randomly split the dataset to 80% training and 20% test. Use cross-validation grid search to find the best set of hyperparameters for random forest to predict the target based on all features. To define the space of hyperparameters of random forest assume:

1. `max_features` is 1/3;
2. `min_samples_leaf` $\in \{2, 10\}$;
3. `n_estimators` $\in \{10, 50, 100\}$.

In the grid search, use 10-fold CV with shuffling and compute and report both the estimates of both the MAE and R^2. Then evaluate the model that has the highest $\hat{R}^2$ determined from CV on the held-out test data and record its test $\hat{R}^2$.