Prediction using fold sampling

David Banh - AskExplain

04/01/2022

Linear regression using folded samples runs faster, retains accuracy

Fold sampling is a way of reducing the total number of samples to a manageable number in order to run prediction algorithms on modern day machines. The folded samples are then unfolded to predict the full dataset.

A way to fold the total number of samples while retaining the original sample structure is done via Generative Encoding (gcode):

https://github.com/AskExplain/gcode/tree/alpha__test__v2022.1

```r
# Removes one feature at a time and uses it as the variable to be predicted (y variable)



# Total permutations :
# total indicates the total number of features
# in the Satellite dataset (p = 36) multipled by the permutation number

permutation_test_number <- 1000







# Run SVD decomposition of samples to a reduced sample space

source("./decompose_sample_space.R")









# Run gcode encoding of samples to a reduced sample space

source("./encode_sample_space.R")
```

Boxplots of total mean absolute error and total runtime are plotted. Here, total indicates the sum of all permutations across all runs when each feature is chosen as the prediction variable.

Of great importance, the total runtime does not include the running of the SVD or gcode algorithms.

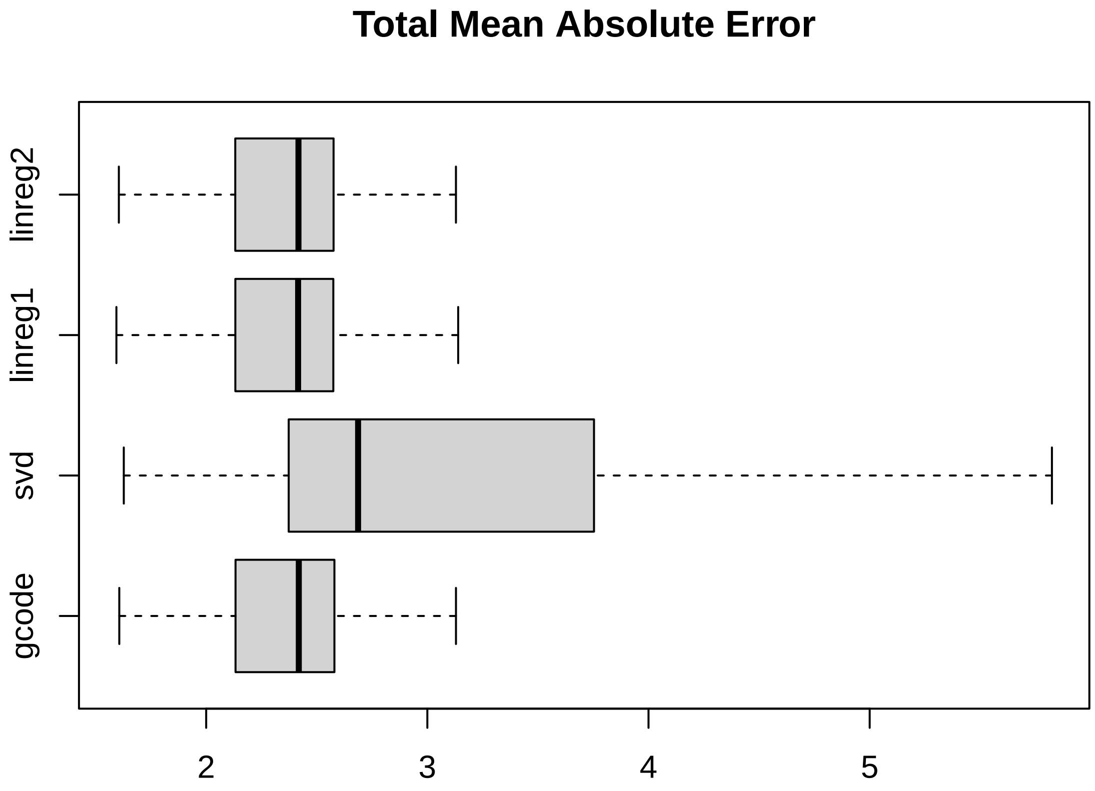
```
# Total Mean Absolute Error







boxplot(data.frame(gcode = total_mae.gcode,
                   svd = total_mae.svd,
                   linreg1 = total_mae.lreg.1,
                   linreg2 = total_mae.lreg.2),
        outline = F,
        horizontal = T,
        main = "Total Mean Absolute Error")
```



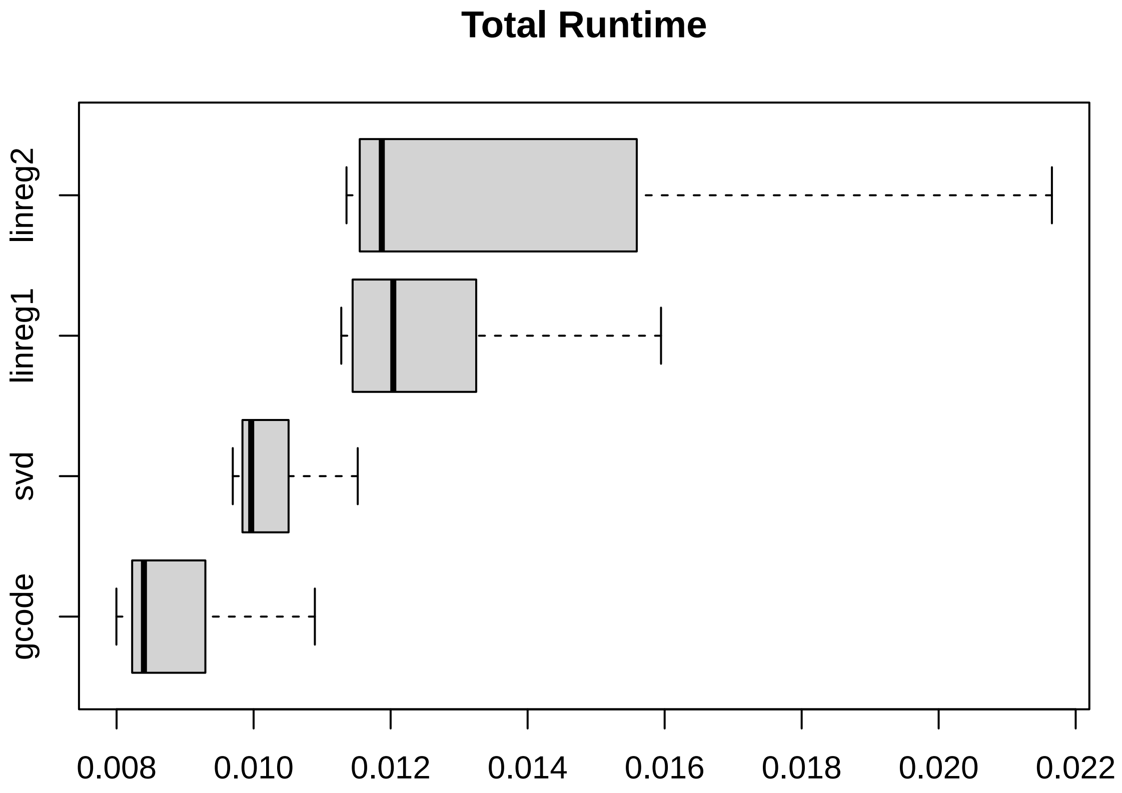
```
# Total Runtime







boxplot(data.frame(gcode = total_time.gcode,
                   svd = total_time.svd,
                   linreg1 = total_time.lreg.1,
                   linreg2 = total_time.lreg.2),
        outline = F,
        horizontal = T,
        main = "Total Runtime")
```


Total Runtime
linreg2
linreg1
svd
gcode