Progress Week 9

Mentors:
Alexandre Gramfort, Marine Le Morvan
Intern:
Manon Rivoire
manon.rivoire@inria.fr

INRIA - Parietal Team

July 7, 2020

Overview

- Lasso Solver
  - On the simulated data
  - Tests on the real data
    - Results
    - Further Works
    - Issue
- Safe Pattern Pruning
  - Progress
  - Further works

Lasso solver

Results on the simulated data

- **Dense with screening**
  Our Lasso : 0.0845980644226074 2
  Lasso sklearn : 0.11719298362731934
  Dual gap : $3.469446951953614e-18$
- **Sparse with screening**
  Our Lasso : 0.23732471466064453
  Lasso sklearn : 4.026190519332886
  Dual gap : 0.0
- **Dense without screening**
  Our Lasso : 0.1467876434326172
  Lasso sklearn : 0.1140127182006836
  Dual gap : $3.469446951953614e-18$

Lasso solver (cont.)

- **Sparse without screening**
  Our Lasso : 2.997429609298706
  Lasso sklearn : 4.001614093780518
  Dual gap : 0.0
- Our Lasso is less computationally expensive in every case except the one for which the lasso is dense and without screening.

Results on the real data

- cv scores obtained for the 5 models on the housing prices dataset without tuning of parameters.

Objective

Compare our Lasso solver to other regression solvers such as random forest or xgboost and show that our lasso solver is less computationally expensive and more accurate on different datasets. Proceed to a benchmark work.

Difficulties encountered

Lasso solver (cont.)

- Conclude on the influence of lambda and n_epochs on the crossval score.
- Tuning of the parameters thanks to the gridsearch : how to include gridsearch within the crossval function to only retain for each model the parameters which allow to obtain the best crossval score ?
- Once the hyperparameters are tuned, compute for each model the crossval score with the best hyperparameters and compare the crossval scores of the different models.
- Are the barplots sufficient to compare the crossval scores between the different models and the influence of the hyperparameters on the crossval score ?
- λ seems to have no influence on the crossval score of our Lasso which seems strange.
- How many features at least do the datasets have to contain ?

Issues

Lasso solver (cont.)

- We obtain a lower cross validation score with our lasso than the one obtained with xgboost and random forest, does this phenomenon only depend on the tuning of the hyperparameters or is this also due to a bad implementation of the solver ?
- In the last case, how to improve the performance of our solver ?

Safe Pattern Pruning

Progress

- The function max val and max val rec which allow to compute the maximal inner product between the feature in the root of the given subtree and the residuals are almost finished.
- Tests ok concerning the functions which compute the inner product both in a dense and in sparse ways. Tests ok for the function which compute the interaction features.
- Issue of shape on the *max_val_rec* function : debugging in progress with ipython.

Further works

- Implement the pruning function
- Include the pruning to our already implemented solver.
- Compare the performances with and without pruning process.
- Make a benchmark.

Bibliography

- Safe Pattern Pruning Paper