This is the page for coordination of the GSoC for scikit-learn.

Scikit-learn is a machine learning module in Python. See http://scikit-learn.org for more details.

Scikit-learn is taking part of the GSoC through the Python Software Foundation: http://wiki.python.org/moin/SummerOfCode

Instructions to student: achieving a good proposal

Difficulty: Scikit-learn is a technical project. Contributing via a GSoC requires a number of expertise in Python coding as well as numerical and machine learning algorithms.

Important: Read: Expectations for prospective students

Application template: https://wiki.python.org/moin/SummerOfCode/ApplicationTemplate2015 Please follow this template.

Also important: A letter from Gaël to former applicants. His suggestions are just as relevant this year.

Hi folks,

The deadline for applications is nearing. I'd like to stress that the scikit-learn will only be accepting high-quality application: it is a challenging, though rewarding, project to work with. To maximize the quality of your application, here are a few advice:

1. First discuss on the mailing list a pre-proposal. Make sure that both the scikit-learn team and yourself are enthusiastic about the idea. Try to have one or two possible mentors that hold a dialog with you.
2. Satisfy the PSF requirements (http://wiki.python.org/moin/SummerOfCode/Expectations) briefly:
    - Demonstrate to your prospective mentor(s) that you are able to complete the project you've proposed
    - Blog for your GSoC project.
    - Contribute at least one patch to the project

I'd add the patch should be somewhat substantial, not just fixing typos.

To contribute patch, please have a look at the [contribution guide] (http://scikit-learn.org/dev/developers/index.html#contributing-code) and the Easy issues in the tracker.

3. In parallel with 2, start a online document (google doc, for instance) to elaborate your final proposal, and if you manage to convince mentors, you can get feedback on it.

As a final note, I want to stress that GSOC projects are ambitious: we are talking about a few months of full time work. Thus the ideas proposed are idea challenging, and the students are supposed to draw a battle plan, with difficult variants and less difficult variants. The GSOC is a full major set of contributions, not a single pull request.

Good luck, I am looking forward to seeing the proposals. You'll see, the scikit is a big friendly and enthusiastic community,

Gaël

A list of topics for a Google summer of code (GSOC) 2017

If you want to apply for the GSoC, you should prepare an application based on one of the ideas below. I good idea is to prepare this application in a public place (a google doc, or a Gist) and to get some feedback on the application from core contributors and potential mentors.

Disclaimer: We are planning to take very few student this time for GSoC 2017 as we have very little time to mentor. Please e-mail the list to see if mentors are available for a project.

Improve online learning for linear models

- Add the multinomial logistic loss to the SGD
- Can we do better than standard /averaging SGD, for instance by adding adagrad? This will require doing a lot of benching.
- A tool to set the learning rate on a few epochs would be nice too

Parallel Decision Tree Building

- Currently parallelization occurs by building separate trees each in parallel
- We would like to add in parallel building of a single decision tree, specifically to allow for parallelized gradient boosting
- This will likely involve refactoring the underlying tree code significantly and so those who are familiar with tree building or the tree code base as it exists are preferred
- Benchmarks will focus on speed improvements, and accuracy for a given amount of time
- The ideal implementation will provide near-linear improvements with an increasing number of jobs and work with all existing code, simply providing users with an addition parameter `n_jobs` at the decision tree level

NOTE:

Please **don't** propose a new feature as we already have a lot of work/PRs in the pipeline.

An impressive proposal would be one which takes up for instance a moderately sized project (need not be a really difficult project, simple tractable ones are preferred) that has been stalled and to define clear goals/milestones and an **achievable** timeline for the milestones.

Some ideas for this include (Please **don't** restrict yourself to the very much incomplete list of projects, if these do not interest you) -

- Adding post pruning support to decision trees. ()

NOTE Brownie points for making a good pass at stalled works and picking the one that you think is most likely to be included (based on previous discussions) and more importantly the one that you think is possible for you within the GSoC time frame.