A list of topics for a Google summer of code (GSOC) 2012

Important: Expectations for prospective students

Also important: A letter from Gaël

Hi folks,

The deadline for applications is nearing (April 6th). 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 entousiastic 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 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 EasyFix 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 enthousiastic community,

Gaël

Improvements and extensions to the Decision Tree Implementation

Possible Mentor: Andreas Mueller?, Peter Prettenhofer (backup)

Possible Candidate: Vikram Kamath

Goal: The C5.0 is an algorithm used to construct m-ary decision trees. It is a successor to the C4.5 algorithm (which in turn is an extension of the ID3 algorithm), all of which were developed by Ross Quinlan. The C5.0 source (implemented in C) has been released under the GNU General Public License (GPL). The aim is to port it and hence make it a feature of sklearn. Additionally, documentation/examples can be created (I have learned from my interaction with Ross Quinlan that the documentation of the C5.0 has not been released under the GPL and is in fact, proprietary).

References:

- 1. http://www.rulequest.com/see5-info.html
- 2. http://www.rulequest.com/see5-unix.html
- 3. http://rulequest.com/see5-comparison.html.
- 4. http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html
- 5. ai.stanford.edu/~ronnyk/treesHB.pdf

Online Low Rank Matrix Completion

Possible mentor: Olivier Grisel, Peter Prettenhofer (backup)

Possible candidate: Vlad Niculae, ?

Goal: Online or Minibatch SGD or similar on a squared l2 reconstruction loss + low rank penalty (nuclear norm) on scipy.sparse matrix: the implicit components of the sparse input representation would be interpreted by the algorithms as missing values rather than zero values.

Application: Build a scalable recommender system example, e.g. on the movielens dataset.

TODO: find references in the literature. Matrix Factorization Jungle

Online Non Negative Matrix Factorization

Possible mentor: Olivier Grisel

Possible candidate: Vlad Niculae, ?

Goal: Online or Minibatch NMF using SGD + positive projections (or any other out-of-core algorithms) accepting both dense and sparse matrix as input (decomposition components can be dense array only).

Application: Build a scalable topic model e.g. on million of Wikipedia abstracts for instance using this script.

References:

- http://research.microsoft.com/apps/pubs/default.aspx?id=143211

Note: it is possible that we will combine the two Online Non Negative Matrix Factorization + Matrix Completion ideas in a single project. Please prospective students feel free to write proposals on one or the other or both ideas at the same time.

Robust PCA

Algorithms for decomposing a design matrix into a low rank + sparse components.

Possible mentor: ?

Possible candidate: Kerui Min (Minibio: "I'm a graduate student at UIUC who is currently pursuing the research work related to low-rank matrices recovery & Robust PCA.")

Applications: ?

References:

- http://perception.csl.uiuc.edu/matrix-rank/home.html
- http://www.icml-2011.org/papers/41_icmlpaper.pdf (randomized algorithm supposedly scalable to larg-ish datasets)

Multilayer Perceptron / Neural Network

Possible mentor: Andreas Mueller, David Warde-Farley

Possible candidate: David Marek

Goal: Implement a stochastic gradient descent algorithm to learn a multi-layer perceptron, starting from https://gist.github.com/2061456.

References:

- http://en.wikipedia.org/wiki/Multi-layer_perceptron
- http://yann.lecun.com/exdb/publis/pdf/lecun-98b.pdf

SVM with low rank kernel approximation

Possible mentor: Andreas Mueller

Goal: Implement a stochastic gradient descent SVM using a low-rank kernel approximation.

References:

- http://pages.cs.wisc.edu/~swright/papers/sncss_tpami.pdf

Generalized Additive Models

Possible mentor: Paolo Losi, Alex Gramfort, (others?)

Goal: Implement one of the state of art methods for Generalized Additive Models
Sparse Version of it is SpAM

References:

- arxiv.org/pdf/0711.4555
- http://code.google.com/p/google-summer-of-code-2011-r/downloads/detail?name=Juemin_Yang.tar.gz
- http://en.wikipedia.org/wiki/Generalized_additive_model
- http://arxiv.org/abs/0806.4115
- http://www.stats.ox.ac.uk/~meinshau/liso.pdf

Coordinated descent in linear models beyond squared loss (eg Logistic)

Possible mentors: Alex Gramfort, Gael Varoquaux

Possible candidate: Immanuel Bayer

Goal: Implement state of art methods for optimizing sparse linear models using coordinate descent.

One objective to avoid the dependency on LibLinear for the LogisticRegression model in order to allow warm restart and Elastic-Net regularization (L1 + L2)

A second objective is to improve the Lasso coordinate descent using strong rules to automatically discard features.

References:

- http://www.jmlr.org/papers/volume11/yuan10c/yuan10c.pdf
- http://www-stat.stanford.edu/~jbien/jrssb2011strong.pdf

Improve GMM

Possible mentors: Gael Varoquaux

- Refurbish the current GMM code to put it to the scikit's standards
- Implement a core-set strategy for GMM

http://las.ethz.ch/files/feldman11scalable-long.pdf http://videolectures.net/nips2011_faulkner_coresets/