A list of topics for a Google summer of code (GSOC) 2011

Online learning

Mentor : O. Grisel

Possible candidate: MB (Averaged Perceptron, MIRA, Structured MIRA, ...)

Goal : Devise an intuitive yet efficient API dedicated to the incremental fitting of some scikit-learn estimators (on an infinite stream of samples for instance).

See this thread on the mailing list for a discussion of such an API. Design decision will be taken by implementing / adapting three concrete models:

- text feature extraction
- online clustering with sequential k-means
- generalized linear model fitting with Stochastic Gradient Descent (both for regression and classification)

Dictionary Learning a.k.a. Sparse Coding

Mentor : Gael Varoquaux, Alex Gramfort

The objective is to bring to the scikit some recent yet very popular methods known as Dictionary Learning or Sparse Coding. It involves heavy numerical computing and has many applications from general signal/image processing to very applied topics such as biomedical imaging. The project will start from existing code snippets (see below) and will require to make some design decision to keep the API simple yet powerful as the rest of the scikit.

Some useful ressources with compatible License:

- Sparse PCA gist
- Another gist by A. Passos
- NMF + Hoyer method in milk

Ensemble methods: Boosting, Bagging, Random Forests, Super Learners, ...

Focus : Boosting

Mentor : Satrajit Ghosh

Quote: [from ESL - Chapter 10] > Boosting is one of the most powerful learning ideas introduced in the last twenty years. It was originally designed for classification problems, but as will be seen in this chapter, it can profitably be extended to regression as well. The motivation for boosting was a procedure that combines the outputs of many "weak" classifiers to produce a powerful

"committee." From this perspective boosting bears a resemblance to bagging and other committee-based approaches (Section 8.8). However we shall see that the connection is at best superficial and that boosting is fundamentally different.

Objective: The goal would be to implement boosting algorithms, but with constantly keeping the general domain of ensemble learning in mind. The specific aims of the project are to implement:

1. loss functions for classification and regression that are not already there in the package
2. general boosting algorithm that can use off-the shelf classifiers
3. gradient boosting

Manifold learning

Mentor : [[Fabian Pedregosa]]

Locality Sensitive Hashing

Mentor : ?

There is an LSH implementation in pybrain (pybrain/supervised/knn/lsh)

This should be combined with implementing hash kernels, to be able to use LSH for a larger purpose than nearest neighbors searches.

Consider this work as a possible basis for a kernelised hashing implementation.

Command line interface

Mentor : ?

Interaction with mldata.org

Mentor : Vincent Michel?