Scikits and stuff

Extra bits for scientific computing in python

Neil Muller

October 29, 2011

Talk Overview

- The 30 second numpy / scipy overview
- The scikits project
- scikits-learn
- scikits-image
- scikits-statsmodels
- Other scikits
- Caveats

The 30s numpy / scipy overview

Or: Where do scikits fit in?

- Numpy
  - Provides array / matrix operations
  - Significant members include: numpy.ndarray, numpy.matrix, numpy.dtype, numpy.ftt, numpy.linalg.svd, numpy.linalg.eig, numpy.polynomial, numpy.ma
  - C-API is designed to make it easy to move data between python and common array / matrix representations
    - Thus tracking flags such c-contiguous (C data layout), f-contiguous (Fortran data layout) and methods for manipulating these flags
  - Provides little actual numeric code beyond basic linear algebra

- Scipy
  - Builds on numpy
  - Integrates various well-known scientific codes (usually by wrapping existing FORTRAN libraries)
  - "glue numpy to existing libraries" philosophy
  - Adds support for optimization (minpack, minpack2, Cobyla), Fancier FFT support (fftpack), numerical integration (quadpack), special functions (specfun), ode integration (odepack), orthogonal distance regression (odrpack), sparse matrixes, numerical computation of probability distributions, interpolation (fitpack), spatial queries and algorithms (qhull), clustering and so forth
- Scipy very broad, but aims to be "generically useful tools for scientific computing in python", rather than "all things to all people"
  - Scipy / Numpy also don't allow GPL code, to allow easy reuse of projects elsewhere

The Scikits project

- Similar concept to "MATLAB toolboxes"
- Provide specialized tools for specific tasks
- Assumes Scipy/Numpy available, and can specific dependencies as required
- Provide unified namespace for finding useful tools
- Not bound by same licensing restrictions as Scipy, so GPL-code can be wrapped
- Currently, everything exists under the scikits namespace package, but this is changing
  - Mainly due to the complexity of several packages needing to manage the same namespace package
  - Most probable result is move to sk- prefix, so sklean, skimage, etc.

scikits-learn

- Toolbox for machine learning / classification problems
- Unified interface to a range of machine learning algorithms
- For labeled data (supervised learning)
  - algo = method(...)
  - algo.fit(x, y)
  - algo.predict(new_x)
- For unlabeled data (unsupervised learning)
  - algo = method(...)
  - algo.fit(x)
  - algo.predict(new_x)
- Numerous algorithms implemented
  - SVM, Linear models (Linear regression, ridge regression, Least Angle Regression, etc.), RFE, Nearest neighbor approaches, naive Bayes, HMM, GMM, etc.

scikits-image

- Aims to match the Matlab Image processing toolbox
- Images are either integer (pixel represented by values in range 0..255) or floating point (pixels in range 0..1)
- Design concept is functions handle both as input and return whichever is most convenient
  - Pipeline tools together - sobel(greyscale_close(image, <structure>)) and so forth
- Integrates with the opencv python bindings

scikits-statsmodels

- Provides functionality for more in-depth statistical modeling than scipy
  - Originally part of scipy, but split off for separate development
- Variety of useful tools, including:
  - Mainly regression models (Ordinary Least Squares, Generalized Least Squares, etc.)
  - Various robust methods
  - Some timeseries analysis options
    - Overlaps a bit with pandas and scikits-timeseries here - some plans to unify this work
  - Several additional statistical tests, mainly aimed at validating models (Cox, J-Test, etc.)

Other scikits

- Various others exists, in varying stages of development and activity
  - see http://scikits.appspot.com and http://projects.scipy.org/scikits
- Some interesting ones
  - audiolab & samplerate - audio manipulation using numpy arrays (io and resampling)
  - talkbox - speech and signal processing (still fairly feature light)
  - optimization - numerical optimization
  - sparse - wrapper for GPL code to play nicely with scipy.sparse
  - odes - development playground for ode solvers to (to be integrated into scipy when mature)
  - cuda - integrate pycuda into scipy framework
  - timeseries - Support for timeseries data
    - Somewhat different focus to pandas
    - Ongoing discussion about merging pandas and scikits-timeseries

Caveats

- Most scikits are still maturing, so features / api often in flux
- Usually small development communties
  - Pitch in if you the feature you want aren't there
- Tendancy to live on the bleeding edge
  - Latest numpy, scipy & cython development braches often needed