CHAPTER

ONE

SCIPY TUTORIAL

1.1 Introduction

Contents

SciPy is a collection of mathematical algorithms and convenience functions built on the Numpy extension of Python. It adds significant power to the interactive Python session by providing the user with high-level commands and classes for manipulating and visualizing data. With SciPy an interactive Python session becomes a data-processing and system-prototyping environment rivaling sytems such as MATLAB, IDL, Octave, R-Lab, and SciLab.

The additional benefit of basing SciPy on Python is that this also makes a powerful programming language available for use in developing sophisticated programs and specialized applications. Scientific applications using SciPy benefit from the development of additional modules in numerous niche's of the software landscape by developers across the world. Everything from parallel programming to web and data-base subroutines and classes have been made available to the Python programmer. All of this power is available in addition to the mathematical libraries in SciPy.

This tutorial will acquaint the first-time user of SciPy with some of its most important features. It assumes that the user has already installed the SciPy package. Some general Python facility is also assumed, such as could be acquired by working through the Python distribution's Tutorial. For further introductory help the user is directed to the Numpy documentation.

For brevity and convenience, we will often assume that the main packages (numpy, scipy, and matplotlib) have been imported as:

```
>>> import numpy as np
>>> import scipy as sp
>>> import matplotlib as mpl
>>> import matplotlib.pyplot as plt
```

These are the import conventions that our community has adopted after discussion on public mailing lists. You will see these conventions used throughout NumPy and SciPy source code and documentation. While we obviously don't require you to follow these conventions in your own code, it is highly recommended.

1.1.1 SciPy Organization

SciPy is organized into subpackages covering different scientific computing domains. These are summarized in the following table:

Subpackage	Description
cluster	Clustering algorithms
constants	Physical and mathematical constants
fftpack	Fast Fourier Transform routines
integrate	Integration and ordinary differential equation solvers
interpolate	Interpolation and smoothing splines
io	Input and Output
linalg	Linear algebra
ndimage	N-dimensional image processing
odr	Orthogonal distance regression
optimize	Optimization and root-finding routines
signal	Signal processing
sparse	Sparse matrices and associated routines
spatial	Spatial data structures and algorithms
special	Special functions
stats	Statistical distributions and functions
weave	C/C++ integration

Scipy sub-packages need to be imported separately, for example:

```
>>> from scipy import linalg, optimize
```

Because of their ubiquitousness, some of the functions in these subpackages are also made available in the scipy namespace to ease their use in interactive sessions and programs. In addition, many basic array functions from numpy are also available at the top-level of the scipy package. Before looking at the sub-packages individually, we will first look at some of these common functions.

1.1.2 Finding Documentation

SciPy and NumPy have documentation versions in both HTML and PDF format available at http://docs.scipy.org/, that cover nearly all available functionality. However, this documentation is still work-in-progress and some parts may be incomplete or sparse. As we are a volunteer organization and depend on the community for growth, your participation - everything from providing feedback to improving the documentation and code - is welcome and actively encouraged.

Python's documentation strings are used in SciPy for on-line documentation. There are two methods for reading them and getting help. One is Python's command help in the pydoc module. Entering this command with no arguments (i.e. >>> help) launches an interactive help session that allows searching through the keywords and modules available to all of Python. Secondly, running the command *help(obj)* with an object as the argument displays that object's calling signature, and documentation string.

The pydoc method of help is sophisticated but uses a pager to display the text. Sometimes this can interfere with the terminal you are running the interactive session within. A scipy-specific help system is also available under the command sp.info. The signature and documentation string for the object passed to the help command are printed to standard output (or to a writeable object passed as the third argument). The second keyword argument of sp.info defines the maximum width of the line for printing. If a module is passed as the argument to help than a list of the functions and classes defined in that module is printed. For example:

```
>>> sp.info(optimize.fmin)
 fmin(func, x0, args=(), xtol=0.0001, ftol=0.0001, maxiter=None, maxfun=None,
      full_output=0, disp=1, retall=0, callback=None)

Minimize a function using the downhill simplex algorithm.

Parameters
----------
func : callable func(x,*args)
```

```
    The objective function to be minimized.
x0 : ndarray
    Initial guess.
args : tuple
    Extra arguments passed to func, i.e. ``f(x,*args)``.
callback : callable
    Called after each iteration, as callback(xk), where xk is the
    current parameter vector.

Returns
-------
xopt : ndarray
    Parameter that minimizes function.
fopt : float
    Value of function at minimum: ``fopt = func(xopt)``.
iter : int
    Number of iterations performed.
funcalls : int
    Number of function calls made.
warnflag : int
    1 : Maximum number of function evaluations made.
    2 : Maximum number of iterations reached.
allvecs : list
    Solution at each iteration.

Other parameters
----------------
xtol : float
    Relative error in xopt acceptable for convergence.
ftol : number
    Relative error in func(xopt) acceptable for convergence.
maxiter : int
    Maximum number of iterations to perform.
maxfun : number
    Maximum number of function evaluations to make.
full_output : bool
    Set to True if fopt and warnflag outputs are desired.
disp : bool
    Set to True to print convergence messages.
retall : bool
    Set to True to return list of solutions at each iteration.

Notes
-----
Uses a Nelder-Mead simplex algorithm to find the minimum of function of
one or more variables.
```

Another useful command is `source`. When given a function written in Python as an argument, it prints out a listing of the source code for that function. This can be helpful in learning about an algorithm or understanding exactly what a function is doing with its arguments. Also don't forget about the Python command `dir` which can be used to look at the namespace of a module or package.

1.2 Basic functions

Contents

1.2.1 Interaction with Numpy

Scipy builds on Numpy, and for all basic array handling needs you can use Numpy functions:

```
>>> import numpy as np
>>> np.some_function()
```

Rather than giving a detailed description of each of these functions (which is available in the Numpy Reference Guide or by using the `help`, `info` and `source` commands), this tutorial will discuss some of the more useful commands which require a little introduction to use to their full potential.

To use functions from some of the Scipy modules, you can do:

```
>>> from scipy import some_module
>>> some_module.some_function()
```

The top level of `scipy` also contains functions from `numpy` and `numpy.lib.scimath`. However, it is better to use them directly from the `numpy` module instead.

Index Tricks

There are some class instances that make special use of the slicing functionality to provide efficient means for array construction. This part will discuss the operation of `np.mgrid` , `np.ogrid` , `np.r_` , and `np.c_` for quickly constructing arrays.

For example, rather than writing something like the following

```
>>> concatenate(([3],[0]*5,arange(-1,1.002,2/9.0)))
```

with the `r_` command one can enter this as

```
>>> r_[3,[0]*5,-1:1:10j]
```

which can ease typing and make for more readable code. Notice how objects are concatenated, and the slicing syntax is (ab)used to construct ranges. The other term that deserves a little explanation is the use of the complex number 10j as the step size in the slicing syntax. This non-standard use allows the number to be interpreted as the number of points to produce in the range rather than as a step size (note we would have used the long integer notation, 10L, but this notation may go away in Python as the integers become unified). This non-standard usage may be unsightly to some, but it gives the user the ability to quickly construct complicated vectors in a very readable fashion. When the number of points is specified in this way, the end- point is inclusive.

The "r" stands for row concatenation because if the objects between commas are 2 dimensional arrays, they are stacked by rows (and thus must have commensurate columns). There is an equivalent command `c_` that stacks 2d arrays by columns but works identically to `r_` for 1d arrays.

Another very useful class instance which makes use of extended slicing notation is the function `mgrid`. In the simplest case, this function can be used to construct 1d ranges as a convenient substitute for arange. It also allows the use of complex-numbers in the step-size to indicate the number of points to place between the (inclusive) end-points. The real purpose of this function however is to produce N, N-d arrays which provide coordinate arrays for an N-dimensional volume. The easiest way to understand this is with an example of its usage:

```
>>> mgrid[0:5,0:5]
array([[[0, 0, 0, 0, 0],
        [1, 1, 1, 1, 1],
        [2, 2, 2, 2, 2],
        [3, 3, 3, 3, 3],
        [4, 4, 4, 4, 4]],
       [[0, 1, 2, 3, 4],
        [0, 1, 2, 3, 4],
        [0, 1, 2, 3, 4],
        [0, 1, 2, 3, 4],
        [0, 1, 2, 3, 4]]])
>>> mgrid[0:5:4j,0:5:4j]
array([[[ 0.    ,  0.    ,  0.    ,  0.    ],
        [ 1.6667,  1.6667,  1.6667,  1.6667],
        [ 3.3333,  3.3333,  3.3333,  3.3333],
        [ 5.    ,  5.    ,  5.    ,  5.    ]],
       [[ 0.    ,  1.6667,  3.3333,  5.    ],
        [ 0.    ,  1.6667,  3.3333,  5.    ],
        [ 0.    ,  1.6667,  3.3333,  5.    ],
        [ 0.    ,  1.6667,  3.3333,  5.    ]]])
```

Having meshed arrays like this is sometimes very useful. However, it is not always needed just to evaluate some N-dimensional function over a grid due to the array-broadcasting rules of Numpy and SciPy. If this is the only purpose for generating a meshgrid, you should instead use the function `ogrid` which generates an "open" grid using `newaxis` judiciously to create N, N-d arrays where only one dimension in each array has length greater than 1. This will save memory and create the same result if the only purpose for the meshgrid is to generate sample points for evaluation of an N-d function.

Shape manipulation

In this category of functions are routines for squeezing out length- one dimensions from N-dimensional arrays, ensuring that an array is at least 1-, 2-, or 3-dimensional, and stacking (concatenating) arrays by rows, columns, and "pages "(in the third dimension). Routines for splitting arrays (roughly the opposite of stacking arrays) are also available.

Polynomials

There are two (interchangeable) ways to deal with 1-d polynomials in SciPy. The first is to use the `poly1d` class from Numpy. This class accepts coefficients or polynomial roots to initialize a polynomial. The polynomial object can then be manipulated in algebraic expressions, integrated, differentiated, and evaluated. It even prints like a polynomial:

```
>>> p = poly1d([3,4,5])
>>> print p
   2
3 x + 4 x + 5
>>> print p*p
   4      3      2
9 x + 24 x + 46 x + 40 x + 25
>>> print p.integ(k=6)
 3     2
x + 2 x + 5 x + 6
```

```
>>> print p.deriv()
6 x + 4
>>> p([4,5])
array([ 69, 100])
```

The other way to handle polynomials is as an array of coefficients with the first element of the array giving the coefficient of the highest power. There are explicit functions to add, subtract, multiply, divide, integrate, differentiate, and evaluate polynomials represented as sequences of coefficients.

Vectorizing functions (vectorize)

One of the features that NumPy provides is a class `vectorize` to convert an ordinary Python function which accepts scalars and returns scalars into a "vectorized-function" with the same broadcasting rules as other Numpy functions (*i.e.* the Universal functions, or ufuncs). For example, suppose you have a Python function named `addsubtract` defined as:

```
>>> def addsubtract(a,b):
...     if a > b:
...         return a - b
...     else:
...         return a + b
```

which defines a function of two scalar variables and returns a scalar result. The class vectorize can be used to "vectorize "this function so that

```
>>> vec_addsubtract = vectorize(addsubtract)
```

returns a function which takes array arguments and returns an array result:

```
>>> vec_addsubtract([0,3,6,9],[1,3,5,7])
array([1, 6, 1, 2])
```

This particular function could have been written in vector form without the use of `vectorize` . But, what if the function you have written is the result of some optimization or integration routine. Such functions can likely only be vectorized using `vectorize`.

Type handling

Note the difference between `np.iscomplex`/`np.isreal` and `np.iscomplexobj`/`np.isrealobj`. The former command is array based and returns byte arrays of ones and zeros providing the result of the element-wise test. The latter command is object based and returns a scalar describing the result of the test on the entire object.

Often it is required to get just the real and/or imaginary part of a complex number. While complex numbers and arrays have attributes that return those values, if one is not sure whether or not the object will be complex-valued, it is better to use the functional forms `np.real` and `np.imag` . These functions succeed for anything that can be turned into a Numpy array. Consider also the function `np.real_if_close` which transforms a complex-valued number with tiny imaginary part into a real number.

Occasionally the need to check whether or not a number is a scalar (Python (long)int, Python float, Python complex, or rank-0 array) occurs in coding. This functionality is provided in the convenient function `np.isscalar` which returns a 1 or a 0.

Finally, ensuring that objects are a certain Numpy type occurs often enough that it has been given a convenient interface in SciPy through the use of the `np.cast` dictionary. The dictionary is keyed by the type it is desired to cast to and the dictionary stores functions to perform the casting. Thus, `np.cast['f'](d)` returns an array of `np.float32` from *d*. This function is also useful as an easy way to get a scalar of a certain type:

```
>>> np.cast['f'](np.pi)
array(3.1415927410125732, dtype=float32)
```

Other useful functions

There are also several other useful functions which should be mentioned. For doing phase processing, the functions `angle`, and `unwrap` are useful. Also, the `linspace` and `logspace` functions return equally spaced samples in a linear or log scale. Finally, it's useful to be aware of the indexing capabilities of Numpy. Mention should be made of the function `select` which extends the functionality of `where` to include multiple conditions and multiple choices. The calling convention is `select(condlist,choicelist,default=0)`. `select` is a vectorized form of the multiple if-statement. It allows rapid construction of a function which returns an array of results based on a list of conditions. Each element of the return array is taken from the array in a `choicelist` corresponding to the first condition in `condlist` that is true. For example

```
>>> x = r_[-2:3]
>>> x
array([-2, -1,  0,  1,  2])
>>> np.select([x > 3, x >= 0],[0,x+2])
array([0, 0, 2, 3, 4])
```

Some additional useful functions can also be found in the module `scipy.misc`. For example the `factorial` and `comb` functions compute $n!$ and $n!/k!(n-k)!$ using either exact integer arithmetic (thanks to Python's Long integer object), or by using floating-point precision and the gamma function. Another function returns a common image used in image processing: `lena`.

Finally, two functions are provided that are useful for approximating derivatives of functions using discrete-differences. The function `central_diff_weights` returns weighting coefficients for an equally-spaced N-point approximation to the derivative of order o. These weights must be multiplied by the function corresponding to these points and the results added to obtain the derivative approximation. This function is intended for use when only samples of the function are avaiable. When the function is an object that can be handed to a routine and evaluated, the function `derivative` can be used to automatically evaluate the object at the correct points to obtain an N-point approximation to the o-th derivative at a given point.

1.3 Special functions (`scipy.special`)

The main feature of the `scipy.special` package is the definition of numerous special functions of mathematical physics. Available functions include airy, elliptic, bessel, gamma, beta, hypergeometric, parabolic cylinder, mathieu, spheroidal wave, struve, and kelvin. There are also some low-level stats functions that are not intended for general use as an easier interface to these functions is provided by the `stats` module. Most of these functions can take array arguments and return array results following the same broadcasting rules as other math functions in Numerical Python. Many of these functions also accept complex numbers as input. For a complete list of the available functions with a one-line description type `>>> help(special)`. Each function also has its own documentation accessible using help. If you don't see a function you need, consider writing it and contributing it to the library. You can write the function in either C, Fortran, or Python. Look in the source code of the library for examples of each of these kinds of functions.

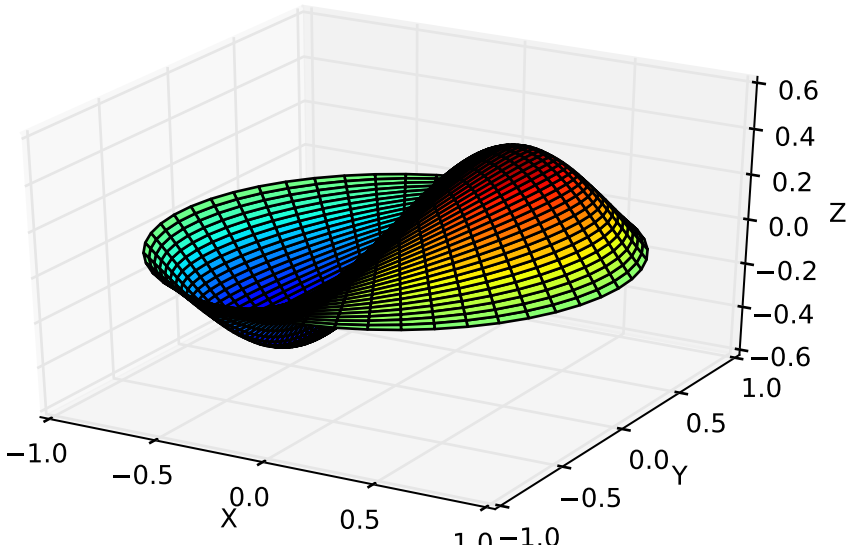
1.3.1 Bessel functions of real order(`jn`, `jn_zeros`)

Bessel functions are a family of solutions to Bessel's differential equation with real or complex order alpha:

$$x^2 \frac{d^2y}{dx^2} + x\frac{dy}{dx} + (x^2 - \alpha^2)y = 0$$

Among other uses, these functions arise in wave propagation problems such as the vibrational modes of a thin drum head. Here is an example of a circular drum head anchored at the edge:

```
>>> from scipy import *
>>> from scipy.special import jn, jn_zeros
>>> def drumhead_height(n, k, distance, angle, t):
...     nth_zero = jn_zeros(n, k)
...     return cos(t)*cos(n*angle)*jn(n, distance*nth_zero)
>>> theta = r_[0:2*pi:50j]
>>> radius = r_[0:1:50j]
>>> x = array([r*cos(theta) for r in radius])
>>> y = array([r*sin(theta) for r in radius])
>>> z = array([drumhead_height(1, 1, r, theta, 0.5) for r in radius])

>>> import pylab
>>> from mpl_toolkits.mplot3d import Axes3D
>>> from matplotlib import cm
>>> fig = pylab.figure()
>>> ax = Axes3D(fig)
>>> ax.plot_surface(x, y, z, rstride=1, cstride=1, cmap=cm.jet)
>>> ax.set_xlabel('X')
>>> ax.set_ylabel('Y')
>>> ax.set_zlabel('Z')
>>> pylab.show()
```



We get the results we were hoping for, with much less computational time. Note that the transformation from $\nu \to \lambda$ takes place entirely in the background. The user need not worry about the details.

The shift-invert mode provides more than just a fast way to obtain a few small eigenvalues. Say you desire to find internal eigenvalues and eigenvectors, e.g. those nearest to $\lambda = 1$. Simply set `sigma = 1` and ARPACK takes care of the rest:

```
>>> evals_mid, evecs_mid = eigsh(X, 3, sigma=1, which='LM')
>>> i_sort = np.argsort(abs(1. / (1 - evals_all)))[-3:]
>>> print evals_all[i_sort]
[ 1.16577199  0.85081388  1.06642272]
>>> print evals_mid
[ 0.85081388  1.06642272  1.16577199]
>>> print np.dot(evecs_mid.T, evecs_all[:,i_sort])
[[-0.  1.  0.]
 [-0. -0.  1.]
 [ 1.  0.  0.]]
```

The eigenvalues come out in a different order, but they're all there. Note that the shift-invert mode requires the internal solution of a matrix inverse. This is taken care of automatically by `eigsh` and `eigs`, but the operation can also be specified by the user. See the docstring of `scipy.sparse.linalg.eigsh` and `scipy.sparse.linalg.eigs` for details.

1.10.5 References

1.11 Compressed Sparse Graph Routines `scipy.sparse.csgraph`

1.11.1 Example: Word Ladders

A Word Ladder is a word game invented by Lewis Carroll in which players find paths between words by switching one letter at a time. For example, one can link "ape" and "man" in the following way:

$$\text{ape} \to \text{apt} \to \text{ait} \to \text{bit} \to \text{big} \to \text{bag} \to \text{mag} \to \text{man}$$

Note that each step involves changing just one letter of the word. This is just one possible path from "ape" to "man", but is it the shortest possible path? If we desire to find the shortest word ladder path between two given words, the sparse graph submodule can help.

First we need a list of valid words. Many operating systems have such a list built-in. For example, on linux, a word list can often be found at one of the following locations:

```
/usr/share/dict
/var/lib/dict
```

Another easy source for words are the scrabble word lists available at various sites around the internet (search with your favorite search engine). We'll first create this list. The system word lists consist of a file with one word per line. The following should be modified to use the particular word list you have available:

```
>>> word_list = open('/usr/share/dict/words').readlines()
>>> word_list = map(str.strip, word_list)
```

We want to look at words of length 3, so let's select just those words of the correct length. We'll also eliminate words which start with upper-case (proper nouns) or contain non alpha-numeric characters like apostrophes and hyphens. Finally, we'll make sure everything is lower-case for comparison later:

```
>>> word_list = [word for word in word_list if len(word) == 3]
>>> word_list = [word for word in word_list if word[0].islower()]
>>> word_list = [word for word in word_list if word.isalpha()]
>>> word_list = map(str.lower, word_list)
>>> len(word_list)
586
```

Now we have a list of 586 valid three-letter words (the exact number may change depending on the particular list used). Each of these words will become a node in our graph, and we will create edges connecting the nodes associated with each pair of words which differs by only one letter.

There are efficient ways to do this, and inefficient ways to do this. To do this as efficiently as possible, we're going to use some sophisticated numpy array manipulation:

```
>>> import numpy as np
>>> word_list = np.asarray(word_list)
>>> word_list.dtype
dtype('|S3')
>>> word_list.sort()  # sort for quick searching later
```

We have an array where each entry is three bytes. We'd like to find all pairs where exactly one byte is different. We'll start by converting each word to a three-dimensional vector:

```
>>> word_bytes = np.ndarray((word_list.size, word_list.itemsize),
...                         dtype='int8',
...                         buffer=word_list.data)
>>> word_bytes.shape
(586, 3)
```

Now we'll use the Hamming distance between each point to determine which pairs of words are connected. The Hamming distance measures the fraction of entries between two vectors which differ: any two words with a hamming distance equal to $1/N$, where N is the number of letters, are connected in the word ladder:

```
>>> from scipy.spatial.distance import pdist, squareform
>>> from scipy.sparse import csr_matrix
>>> hamming_dist = pdist(word_bytes, metric='hamming')
>>> graph = csr_matrix(squareform(hamming_dist < 1.5 / word_list.itemsize))
```

When comparing the distances, we don't use an equality because this can be unstable for floating point values. The inequality produces the desired result as long as no two entries of the word list are identical. Now that our graph is set up, we'll use a shortest path search to find the path between any two words in the graph:

```
>>> i1 = word_list.searchsorted('ape')
>>> i2 = word_list.searchsorted('man')
>>> word_list[i1]
'ape'
>>> word_list[i2]
'man'
```

We need to check that these match, because if the words are not in the list that will not be the case. Now all we need is to find the shortest path between these two indices in the graph. We'll use dijkstra's algorithm, because it allows us to find the path for just one node:

```
>>> from scipy.sparse.csgraph import dijkstra
>>> distances, predecessors = dijkstra(graph, indices=i1,
...                                    return_predecessors=True)
>>> print distances[i2]
5.0
```

So we see that the shortest path between 'ape' and 'man' contains only five steps. We can use the predecessors returned by the algorithm to reconstruct this path:

```
>>> path = []
>>> i = i2
>>> while i != i1:
>>>     path.append(word_list[i])
>>>     i = predecessors[i]
>>> path.append(word_list[i1])
>>> print path[::-1]
['ape', 'apt', 'opt', 'oat', 'mat', 'man']
```

This is three fewer links than our initial example: the path from ape to man is only five steps.

Using other tools in the module, we can answer other questions. For example, are there three-letter words which are not linked in a word ladder? This is a question of connected components in the graph:

```
>>> from scipy.sparse.csgraph import connected_components
>>> N_components, component_list = connected_components(graph)
>>> print N_components
15
```

In this particular sample of three-letter words, there are 15 connected components: that is, 15 distinct sets of words with no paths between the sets. How many words are in each of these sets? We can learn this from the list of components:

```
>>> [np.sum(component_list == i) for i in range(15)]
[571, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

There is one large connected set, and 14 smaller ones. Let's look at the words in the smaller ones:

```
>>> [list(word_list[np.where(component_list == i)]) for i in range(1, 15)]
[['aha'],
 ['chi'],
 ['ebb'],
 ['ems', 'emu'],
 ['gnu'],
 ['ism'],
 ['khz'],
 ['nth'],
 ['ova'],
 ['qua'],
 ['ugh'],
 ['ups'],
 ['urn'],
 ['use']]
```

These are all the three-letter words which do not connect to others via a word ladder.

We might also be curious about which words are maximally separated. Which two words take the most links to connect? We can determine this by computing the matrix of all shortest paths. Note that by convention, the distance between two non-connected points is reported to be infinity, so we'll need to remove these before finding the maximum:

```
>>> distances, predecessors = dijkstra(graph, return_predecessors=True)
>>> np.max(distances[~np.isinf(distances)])
13.0
```

So there is at least one pair of words which takes 13 steps to get from one to the other! Let's determine which these are:

```
>>> i1, i2 = np.where(distances == 13)
>>> zip(word_list[i1], word_list[i2])
```

```
[('imp', 'ohm'),
 ('imp', 'ohs'),
 ('ohm', 'imp'),
 ('ohm', 'ump'),
 ('ohs', 'imp'),
 ('ohs', 'ump'),
 ('ump', 'ohm'),
 ('ump', 'ohs')]
```

We see that there are two pairs of words which are maximally separated from each other: 'imp' and 'ump' on one hand, and 'ohm' and 'ohs' on the other hand. We can find the connecting list in the same way as above:

```
>>> path = []
>>> i = i2[0]
>>> while i != i1[0]:
>>>     path.append(word_list[i])
>>>     i = predecessors[i1[0], i]
>>> path.append(word_list[i1[0]])
>>> print path[::-1]
['imp', 'amp', 'asp', 'ask', 'ark', 'are', 'aye', 'rye', 'roe', 'woe', 'woo', 'who', 'oho', 'ohm']
```

This gives us the path we desired to see.

Word ladders are just one potential application of scipy's fast graph algorithms for sparse matrices. Graph theory makes appearances in many areas of mathematics, data analysis, and machine learning. The sparse graph tools are flexible enough to handle many of these situations.

1.12 Spatial data structures and algorithms (`scipy.spatial`)

`scipy.spatial` can compute triangulations, Voronoi diagrams, and convex hulls of a set of points, by leveraging the Qhull library.

Moreover, it contains `KDTree` implementations for nearest-neighbor point queries, and utilities for distance computations in various metrics.

1.12.1 Delaunay triangulations

The Delaunay triangulation is a subdivision of a set of points into a non-overlapping set of triangles, such that no point is inside the circumcircle of any triangle. In practice, such triangulations tend to avoid triangles with small angles.

Delaunay triangulation can be computed using `scipy.spatial` as follows:

```
>>> from scipy.spatial import Delaunay
>>> points = np.array([[0, 0], [0, 1.1], [1, 0], [1, 1]])
>>> tri = Delaunay(points)
```

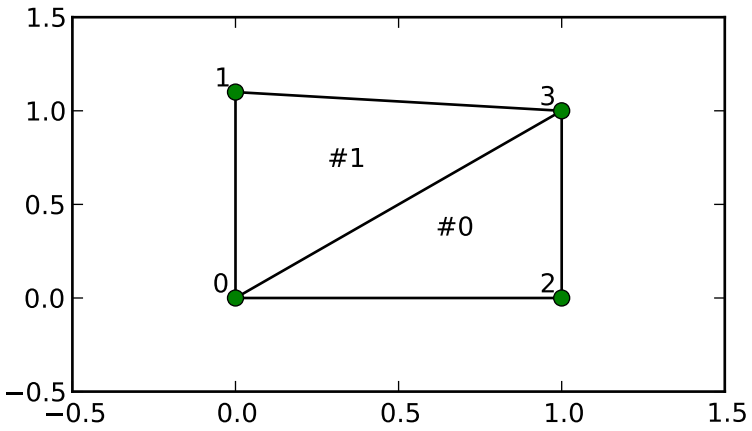
We can visualize it:

```
>>> import matplotlib.pyplot as plt
>>> plt.triplot(points[:,0], points[:,1], tri.simplices.copy())
>>> plt.plot(points[:,0], points[:,1], 'o')
```

And add some further decorations:

```
>>> for j, p in enumerate(points):
...     plt.text(p[0]-0.03, p[1]+0.03, j, ha='right') # label the points
```

```
>>> for j, s in enumerate(tri.simplices):
...     p = points[s].mean(axis=0)
...     plt.text(p[0], p[1], '#%d' % j, ha='center') # label triangles
>>> plt.xlim(-0.5, 1.5); plt.ylim(-0.5, 1.5)
>>> plt.show()
```



The structure of the triangulation is encoded in the following way: the `simplices` attribute contains the indices of the points in the `points` array that make up the triangle. For instance:

```
>>> i = 1
>>> tri.simplices[i,:]
array([3, 1, 0], dtype=int32)
>>> points[tri.simplices[i,:]]
array([[ 1. ,  1. ],
       [ 0. ,  1.1],
       [ 0. ,  0. ]])
```

Moreover, neighboring triangles can also be found out:

```
>>> tri.neighbors[i]
array([-1,  0, -1], dtype=int32)
```

What this tells us is that this triangle has triangle #0 as a neighbor, but no other neighbors. Moreover, it tells us that neighbor 0 is opposite the vertex 1 of the triangle:

```
>>> points[tri.simplices[i, 1]]
array([ 0. ,  1.1])
```

Indeed, from the figure we see that this is the case.

Qhull can also perform tesselations to simplices also for higher-dimensional point sets (for instance, subdivision into tetrahedra in 3-D).

Coplanar points

It is important to note that not *all* points necessarily appear as vertices of the triangulation, due to numerical precision issues in forming the triangulation. Consider the above with a duplicated point:

```
>>> points = np.array([[0, 0], [0, 1], [1, 0], [1, 1], [1, 1]])
>>> tri = Delaunay(points)
>>> np.unique(tri.simplices.ravel())
array([0, 1, 2, 3], dtype=int32)
```

Observe that point #4, which is a duplicate, does not occur as a vertex of the triangulation. That this happened is recorded:

```
>>> tri.coplanar
array([[4, 0, 3]], dtype=int32)
```

This means that point 4 resides near triangle 0 and vertex 3, but is not included in the triangulation.

Note that such degeneracies can occur not only because of duplicated points, but also for more complicated geometrical reasons, even in point sets that at first sight seem well-behaved.

However, Qhull has the "QJ" option, which instructs it to perturb the input data randomly until degeneracies are resolved:

```
>>> tri = Delaunay(points, qhull_options="QJ Pp")
>>> points[tri.simplices]
array([[[1, 1],
        [1, 0],
        [0, 0]],
       [[1, 1],
        [1, 1],
        [1, 0]],
       [[0, 1],
        [1, 1],
        [0, 0]],
       [[0, 1],
        [1, 1],
        [1, 1]]])
```

Two new triangles appeared. However, we see that they are degenerate and have zero area.

1.12.2 Convex hulls

Convex hull is the smallest convex object containing all points in a given point set.

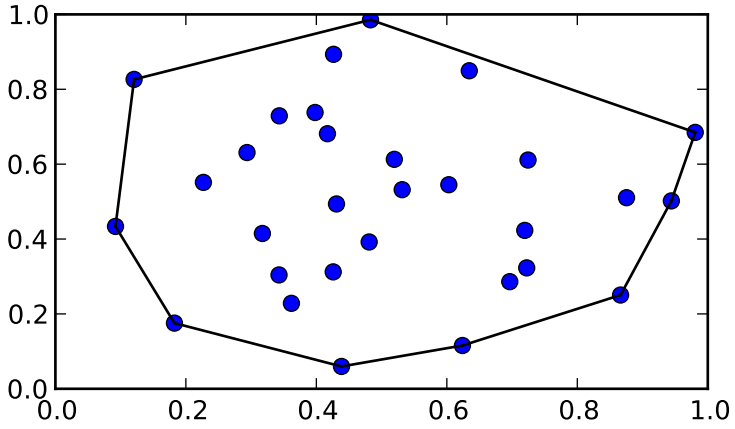
These can be computed via the Qhull wrappers in `scipy.spatial` as follows:

```
>>> from scipy.spatial import ConvexHull
>>> points = np.random.rand(30, 2)   # 30 random points in 2-D
>>> hull = ConvexHull(points)
```

The convex hull is represented as a set of N-1 dimensional simplices, which in 2-D means line segments. The storage scheme is exactly the same as for the simplices in the Delaunay triangulation discussed above.

We can illustrate the above result:

```
>>> import matplotlib.pyplot as plt
>>> plt.plot(points[:,0], points[:,1], 'o')
>>> for simplex in hull.simplices:
>>>     plt.plot(points[simplex,0], points[simplex,1], 'k-')
>>> plt.show()
```



The same can be achieved with `scipy.spatial.convex_hull_plot_2d`.

1.12.3 Voronoi diagrams

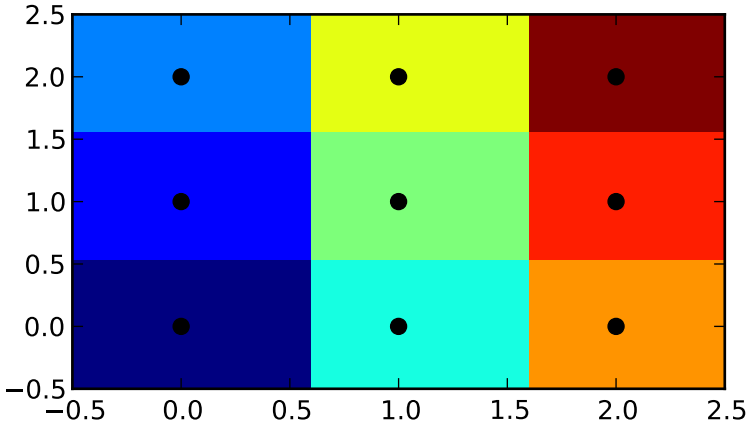
A Voronoi diagram is a subdivision of the space into the nearest neighborhoods of a given set of points.

There are two ways to approach this object using `scipy.spatial`. First, one can use the `KDTree` to answer the question "which of the points is closest to this one", and define the regions that way:

```
>>> from scipy.spatial import KDTree
>>> points = np.array([[0, 0], [0, 1], [0, 2], [1, 0], [1, 1], [1, 2],
...                    [2, 0], [2, 1], [2, 2]])
>>> tree = KDTree(points)
>>> tree.query([0.1, 0.1])
(0.14142135623730953, 0)
```

So the point `(0.1, 0.1)` belongs to region `0`. In color:

```
>>> x = np.linspace(-0.5, 2.5, 31)
>>> y = np.linspace(-0.5, 2.5, 33)
>>> xx, yy = np.meshgrid(x, y)
>>> xy = np.c_[xx.ravel(), yy.ravel()]
>>> import matplotlib.pyplot as plt
>>> plt.pcolor(x, y, tree.query(xy)[1].reshape(33, 31))
>>> plt.plot(points[:,0], points[:,1], 'ko')
>>> plt.show()
```



This does not, however, give the Voronoi diagram as a geometrical object.

The representation in terms of lines and points can be again obtained via the Qhull wrappers in `scipy.spatial`:

```
>>> from scipy.spatial import Voronoi
>>> vor = Voronoi(points)
>>> vor.vertices
array([[ 0.5,  0.5],
       [ 1.5,  0.5],
       [ 0.5,  1.5],
       [ 1.5,  1.5]])
```

The Voronoi vertices denote the set of points forming the polygonal edges of the Voronoi regions. In this case, there are 9 different regions:

```
>>> vor.regions
[[-1, 0], [-1, 1], [1, -1, 0], [3, -1, 2], [-1, 3], [-1, 2], [3, 1, 0, 2], [2, -1, 0], [3, -1, 1]]
```

Negative value `-1` again indicates a point at infinity. Indeed, only one of the regions, `[3, 1, 0, 2]`, is bounded. Note here that due to similar numerical precision issues as in Delaunay triangulation above, there may be fewer Voronoi regions than input points.

The ridges (lines in 2-D) separating the regions are described as a similar collection of simplices as the convex hull pieces:

```
>>> vor.ridge_vertices
[[-1, 0], [-1, 0], [-1, 1], [-1, 1], [0, 1], [-1, 3], [-1, 2], [2, 3], [-1, 3], [-1, 2], [0, 2], [1,
```

These numbers indicate indices of the Voronoi vertices making up the line segments. `-1` is again a point at infinity — only four of the 12 lines is a bounded line segment while the others extend to infinity.

The Voronoi ridges are perpendicular to lines drawn between the input points. Which two points each ridge corresponds to is also recorded:

```
>>> vor.ridge_points
array([[0, 3],
       [0, 1],
       [6, 3],
       [6, 7],
       [3, 4],
```

```
       [5, 8],
       [5, 2],
       [5, 4],
       [8, 7],
       [2, 1],
       [4, 1],
       [4, 7]], dtype=int32)
```

This information, taken together, is enough to construct the full diagram.

We can plot it as follows. First the points and the Voronoi vertices:

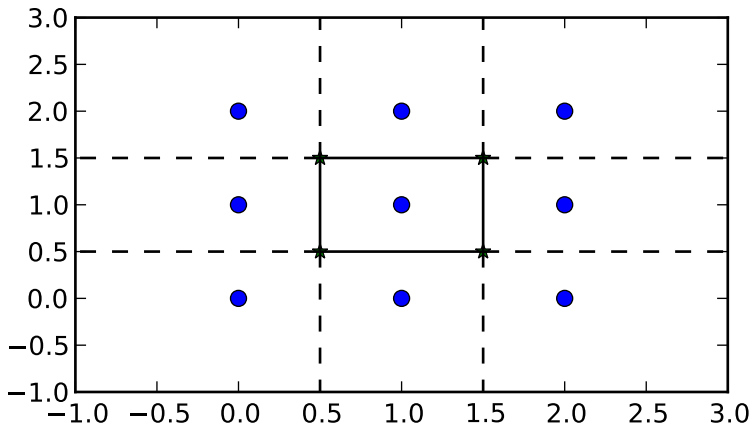
```
>>> plt.plot(points[:,0], points[:,1], 'o')
>>> plt.plot(vor.vertices[:,0], vor.vertices[:,1], '*')
>>> plt.xlim(-1, 3); plt.ylim(-1, 3)
```

Plotting the finite line segments goes as for the convex hull, but now we have to guard for the infinite edges:

```
>>> for simplex in vor.ridge_vertices:
>>>     simplex = np.asarray(simplex)
>>>     if np.all(simplex >= 0):
>>>         plt.plot(vor.vertices[simplex,0], vor.vertices[simplex,1], 'k-')
```

The ridges extending to infinity require a bit more care:

```
>>> center = points.mean(axis=0)
>>> for pointidx, simplex in zip(vor.ridge_points, vor.ridge_vertices):
>>>     simplex = np.asarray(simplex)
>>>     if np.any(simplex < 0):
>>>         i = simplex[simplex >= 0][0] # finite end Voronoi vertex
>>>         t = points[pointidx[1]] - points[pointidx[0]] # tangent
>>>         t /= np.linalg.norm(t)
>>>         n = np.array([-t[1], t[0]]) # normal
>>>         midpoint = points[pointidx].mean(axis=0)
>>>         far_point = vor.vertices[i] + np.sign(np.dot(midpoint - center, n)) * n * 100
>>>         plt.plot([vor.vertices[i,0], far_point[0]],
...                  [vor.vertices[i,1], far_point[1]], 'k--')
>>> plt.show()
```



This plot can also be created using `scipy.spatial.voronoi_plot_2d`.

1.13 Statistics (`scipy.stats`)

1.13.1 Introduction

In this tutorial we discuss many, but certainly not all, features of `scipy.stats`. The intention here is to provide a user with a working knowledge of this package. We refer to the reference manual for further details.

Note: This documentation is work in progress.

1.13.2 Random Variables

There are two general distribution classes that have been implemented for encapsulating *continuous random variables* and *discrete random variables* . Over 80 continuous random variables (RVs) and 10 discrete random variables have been implemented using these classes. Besides this, new routines and distributions can easily added by the end user. (If you create one, please contribute it).

All of the statistics functions are located in the sub-package `scipy.stats` and a fairly complete listing of these functions can be obtained using `info(stats)`. The list of the random variables available can also be obtained from the docstring for the stats sub-package.

In the discussion below we mostly focus on continuous RVs. Nearly all applies to discrete variables also, but we point out some differences here: *Specific Points for Discrete Distributions*.

Getting Help

First of all, all distributions are accompanied with help functions. To obtain just some basic information we can call

```
>>> from scipy import stats
>>> from scipy.stats import norm
>>> print norm.__doc__
```

To find the support, i.e., upper and lower bound of the distribution, call:

```
>>> print 'bounds of distribution lower: %s, upper: %s' % (norm.a,norm.b)
bounds of distribution lower: -inf, upper: inf
```

We can list all methods and properties of the distribution with `dir(norm)`. As it turns out, some of the methods are private methods although they are not named as such (their name does not start with a leading underscore), for example `veccdf`, are only available for internal calculation.

To obtain the *real* main methods, we list the methods of the frozen distribution. (We explain the meaning of a *frozen* distribution below).

```
>>> rv = norm()
>>> dir(rv) #reformatted
    ['__class__', '__delattr__', '__dict__', '__doc__', '__getattribute__',
    '__hash__', '__init__', '__module__', '__new__', '__reduce__', '__reduce_ex__',
    '__repr__', '__setattr__', '__str__', '__weakref__', 'args', 'cdf', 'dist',
    'entropy', 'isf', 'kwds', 'moment', 'pdf', 'pmf', 'ppf', 'rvs', 'sf', 'stats']
```

Finally, we can obtain the list of available distribution through introspection:

```
>>> import warnings
>>> warnings.simplefilter('ignore', DeprecationWarning)
>>> dist_continu = [d for d in dir(stats) if
...                 isinstance(getattr(stats,d), stats.rv_continuous)]
>>> dist_discrete = [d for d in dir(stats) if
...                  isinstance(getattr(stats,d), stats.rv_discrete)]
>>> print 'number of continuous distributions:', len(dist_continu)
number of continuous distributions: 84
>>> print 'number of discrete distributions:  ', len(dist_discrete)
number of discrete distributions:   12
```

Common Methods

The main public methods for continuous RVs are:

- rvs: Random Variates
- pdf: Probability Density Function
- cdf: Cumulative Distribution Function
- sf: Survival Function (1-CDF)
- ppf: Percent Point Function (Inverse of CDF)
- isf: Inverse Survival Function (Inverse of SF)
- stats: Return mean, variance, (Fisher's) skew, or (Fisher's) kurtosis
- moment: non-central moments of the distribution

Let's take a normal RV as an example.

```
>>> norm.cdf(0)
0.5
```

To compute the `cdf` at a number of points, we can pass a list or a numpy array.

```
>>> norm.cdf([-1., 0, 1])
array([ 0.15865525,  0.5       ,  0.84134475])
>>> import numpy as np
>>> norm.cdf(np.array([-1., 0, 1]))
array([ 0.15865525,  0.5       ,  0.84134475])
```

Thus, the basic methods such as *pdf*, *cdf*, and so on are vectorized with `np.vectorize`.

Other generally useful methods are supported too:

```
>>> norm.mean(), norm.std(), norm.var()
(0.0, 1.0, 1.0)
>>> norm.stats(moments = "mv")
(array(0.0), array(1.0))
```

To find the median of a distribution we can use the percent point function `ppf`, which is the inverse of the `cdf`:

```
>>> norm.ppf(0.5)
0.0
```

To generate a set of random variates:

```
>>> norm.rvs(size=5)
array([-0.35687759,  1.34347647, -0.11710531, -1.00725181, -0.51275702])
```

Don't think that `norm.rvs(5)` generates 5 variates:

```
>>> norm.rvs(5)
7.131624370075814
```

This brings us, in fact, to the topic of the next subsection.

Shifting and Scaling

All continuous distributions take `loc` and `scale` as keyword parameters to adjust the location and scale of the distribution, e.g. for the standard normal distribution the location is the mean and the scale is the standard deviation.

```
>>> norm.stats(loc = 3, scale = 4, moments = "mv")
(array(3.0), array(16.0))
```

In general the standardized distribution for a random variable `X` is obtained through the transformation `(X - loc) / scale`. The default values are `loc = 0` and `scale = 1`.

Smart use of `loc` and `scale` can help modify the standard distributions in many ways. To illustrate the scaling further, the `cdf` of an exponentially distributed RV with mean $1/\lambda$ is given by

$$F(x) = 1 - \exp(-\lambda x)$$

By applying the scaling rule above, it can be seen that by taking `scale = 1./lambda` we get the proper scale.

```
>>> from scipy.stats import expon
>>> expon.mean(scale = 3.)
3.0
```

The uniform distribution is also interesting:

```
>>> from scipy.stats import uniform
>>> uniform.cdf([0,1,2,3,4,5], loc = 1, scale = 4)
array([ 0.  ,  0.  ,  0.25,  0.5 ,  0.75,  1.  ])
```

Finally, recall from the previous paragraph that we are left with the problem of the meaning of `norm.rvs(5)`. As it turns out, calling a distribution like this, the first argument, i.e., the 5, gets passed to set the `loc` parameter. Let's see:

```
>>> np.mean(norm.rvs(5, size=500))
4.983550784784704
```

Thus, to explain the output of the example of the last section: `norm.rvs(5)` generates a normally distributed random variate with mean `loc=5.

I prefer to set the `loc` and `scale` parameter explicitly, by passing the values as keywords rather than as arguments. This is less of a hassle as it may seem. We clarify this below when we explain the topic of *freezing a RV*.

Shape Parameters

While a general continuous random variable can be shifted and scaled with the `loc` and `scale` parameters, some distributions require additional shape parameters. For instance, the gamma distribution, with density

$$\gamma(x, n) = \frac{\lambda(\lambda x)^{n-1}}{\Gamma(n)} e^{-\lambda x} ,$$

requires the shape parameter n. Observe that setting λ can be obtained by setting the `scale` keyword to $1/\lambda$.

Let's check the number and name of the shape parameters of the gamma distribution. (We know from the above that this should be 1.)

```
>>> from scipy.stats import gamma
>>> gamma.numargs
1
>>> gamma.shapes
'a'
```

Now we set the value of the shape variable to 1 to obtain the exponential distribution, so that we compare easily whether we get the results we expect.

```
>>>  gamma(1, scale=2.).stats(moments="mv")
(array(2.0), array(4.0))
```

Notice that we can also specify shape parameters as keywords:

```
>>> gamma(a=1, scale=2.).stats(moments="mv")
(array(2.0), array(4.0))
```

Freezing a Distribution

Passing the `loc` and `scale` keywords time and again can become quite bothersome. The concept of *freezing* a RV is used to solve such problems.

```
>>> rv = gamma(1, scale=2.)
```

By using `rv` we no longer have to include the scale or the shape parameters anymore. Thus, distributions can be used in one of two ways, either by passing all distribution parameters to each method call (such as we did earlier) or by freezing the parameters for the instance of the distribution. Let us check this:

```
>>> rv.mean(), rv.std()
(2.0, 2.0)
```

This is indeed what we should get.

Broadcasting

The basic methods `pdf` and so on satisfy the usual numpy broadcasting rules. For example, we can calculate the critical values for the upper tail of the t distribution for different probabilites and degrees of freedom.

```
>>> stats.t.isf([0.1, 0.05, 0.01], [[10], [11]])
array([[ 1.37218364,  1.81246112,  2.76376946],
       [ 1.36343032,  1.79588482,  2.71807918]])
```

Here, the first row are the critical values for 10 degrees of freedom and the second row for 11 degrees of freedom (d.o.f.). Thus, the broadcasting rules give the same result of calling `isf` twice:

```
>>> stats.t.isf([0.1, 0.05, 0.01], 10)
array([ 1.37218364,  1.81246112,  2.76376946])
>>> stats.t.isf([0.1, 0.05, 0.01], 11)
array([ 1.36343032,  1.79588482,  2.71807918])
```

If the array with probabilities, i.e, `[0.1, 0.05, 0.01]` and the array of degrees of freedom i.e., `[10, 11, 12]`, have the same array shape, then element wise matching is used. As an example, we can obtain the 10% tail for 10 d.o.f., the 5% tail for 11 d.o.f. and the 1% tail for 12 d.o.f. by calling

```
>>> stats.t.isf([0.1, 0.05, 0.01], [10, 11, 12])
array([ 1.37218364,  1.79588482,  2.68099799])
```

Specific Points for Discrete Distributions

Discrete distribution have mostly the same basic methods as the continuous distributions. However `pdf` is replaced the probability mass function `pmf`, no estimation methods, such as fit, are available, and `scale` is not a valid keyword parameter. The location parameter, keyword `loc` can still be used to shift the distribution.

The computation of the cdf requires some extra attention. In the case of continuous distribution the cumulative distribution function is in most standard cases strictly monotonic increasing in the bounds (a,b) and has therefore a unique inverse. The cdf of a discrete distribution, however, is a step function, hence the inverse cdf, i.e., the percent point function, requires a different definition:

```
ppf(q) = min{x : cdf(x) >= q, x integer}
```

For further info, see the docs here.

We can look at the hypergeometric distribution as an example

```
>>> from scipy.stats import hypergeom
>>> [M, n, N] = [20, 7, 12]
```

If we use the cdf at some integer points and then evaluate the ppf at those cdf values, we get the initial integers back, for example

```
>>> x = np.arange(4)*2
>>> x
array([0, 2, 4, 6])
>>> prb = hypergeom.cdf(x, M, n, N)
>>> prb
array([ 0.0001031991744066,  0.0521155830753351,  0.6083591331269301,
        0.9897832817337386])
>>> hypergeom.ppf(prb, M, n, N)
array([ 0.,  2.,  4.,  6.])
```

If we use values that are not at the kinks of the cdf step function, we get the next higher integer back:

```
>>> hypergeom.ppf(prb+1e-8, M, n, N)
array([ 1.,  3.,  5.,  7.])
>>> hypergeom.ppf(prb-1e-8, M, n, N)
array([ 0.,  2.,  4.,  6.])
```

Fitting Distributions

The main additional methods of the not frozen distribution are related to the estimation of distribution parameters:

- ***fit: maximum likelihood estimation of distribution parameters, including location*** and scale
- fit_loc_scale: estimation of location and scale when shape parameters are given
- nnlf: negative log likelihood function
- expect: Calculate the expectation of a function against the pdf or pmf

Performance Issues and Cautionary Remarks

The performance of the individual methods, in terms of speed, varies widely by distribution and method. The results of a method are obtained in one of two ways: either by explicit calculation, or by a generic algorithm that is independent of the specific distribution.

Explicit calculation, on the one hand, requires that the method is directly specified for the given distribution, either through analytic formulas or through special functions in `scipy.special` or `numpy.random` for `rvs`. These are usually relatively fast calculations.

The generic methods, on the other hand, are used if the distribution does not specify any explicit calculation. To define a distribution, only one of pdf or cdf is necessary; all other methods can be derived using numeric integration and root finding. However, these indirect methods can be *very* slow. As an example, `rgh = stats.gausshyper.rvs(0.5, 2, 2, 2, size=100)` creates random variables in a very indirect way and takes about 19 seconds for 100 random variables on my computer, while one million random variables from the standard normal or from the t distribution take just above one second.

Remaining Issues

The distributions in `scipy.stats` have recently been corrected and improved and gained a considerable test suite, however a few issues remain:

- skew and kurtosis, 3rd and 4th moments and entropy are not thoroughly tested and some coarse testing indicates that there are still some incorrect results left.
- the distributions have been tested over some range of parameters, however in some corner ranges, a few incorrect results may remain.
- the maximum likelihood estimation in *fit* does not work with default starting parameters for all distributions and the user needs to supply good starting parameters. Also, for some distribution using a maximum likelihood estimator might inherently not be the best choice.

1.13.3 Building Specific Distributions

The next examples shows how to build your own distributions. Further examples show the usage of the distributions and some statistical tests.

Making a Continuous Distribution, i.e., Subclassing `rv_continuous`

Making continuous distributions is fairly simple.

```
>>> from scipy import stats
>>> class deterministic_gen(stats.rv_continuous):
...     def _cdf(self, x ):
...         return np.where(x<0, 0., 1.)
...     def _stats(self):
...         return 0., 0., 0., 0.

>>> deterministic = deterministic_gen(name="deterministic")
>>> deterministic.cdf(np.arange(-3, 3, 0.5))
array([ 0.,  0.,  0.,  0.,  0.,  0.,  1.,  1.,  1.,  1.,  1.,  1.])
```

Interestingly, the `pdf` is now computed automatically:

```
>>> deterministic.pdf(np.arange(-3, 3, 0.5))
array([  0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
         0.00000000e+00,   0.00000000e+00,   0.00000000e+00,
         5.83333333e+04,   4.16333634e-12,   4.16333634e-12,
         4.16333634e-12,   4.16333634e-12,   4.16333634e-12])
```

Be aware of the performance issues mentions in *Performance Issues and Cautionary Remarks*. The computation of unspecified common methods can become very slow, since only general methods are called which, by their very nature, cannot use any specific information about the distribution. Thus, as a cautionary example:

```
>>> from scipy.integrate import quad
>>> quad(deterministic.pdf, -1e-1, 1e-1)
(4.163336342344337e-13, 0.0)
```

But this is not correct: the integral over this pdf should be 1. Let's make the integration interval smaller:

```
>>> quad(deterministic.pdf, -1e-3, 1e-3) # warning removed
(1.000076872229173, 0.0010625571718182458)
```

This looks better. However, the problem originated from the fact that the pdf is not specified in the class definition of the deterministic distribution.

Subclassing `rv_discrete`

In the following we use `stats.rv_discrete` to generate a discrete distribution that has the probabilities of the truncated normal for the intervals centered around the integers.

General Info

From the docstring of rv_discrete, i.e.,

```
>>> from scipy.stats import rv_discrete
>>> help(rv_discrete)
```

we learn that:

> "You can construct an aribtrary discrete rv where P{X=xk} = pk by passing to the rv_discrete initialization method (through the values= keyword) a tuple of sequences (xk, pk) which describes only those values of X (xk) that occur with nonzero probability (pk)."

Next to this, there are some further requirements for this approach to work:

- The keyword *name* is required.
- The support points of the distribution xk have to be integers.
- The number of significant digits (decimals) needs to be specified.

In fact, if the last two requirements are not satisfied an exception may be raised or the resulting numbers may be incorrect.

An Example

Let's do the work. First

```
>>> npoints = 20   # number of integer support points of the distribution minus 1
>>> npointsh = npoints / 2
>>> npointsf = float(npoints)
>>> nbound = 4   # bounds for the truncated normal
>>> normbound = (1+1/npointsf) * nbound   # actual bounds of truncated normal
>>> grid = np.arange(-npointsh, npointsh+2, 1)   # integer grid
>>> gridlimitsnorm = (grid-0.5) / npointsh * nbound   # bin limits for the truncnorm
>>> gridlimits = grid - 0.5   # used later in the analysis
>>> grid = grid[:-1]
>>> probs = np.diff(stats.truncnorm.cdf(gridlimitsnorm, -normbound, normbound))
>>> gridint = grid
```

And finally we can subclass `rv_discrete`:

```
>>> normdiscrete = stats.rv_discrete(values=(gridint,
...              np.round(probs, decimals=7)), name='normdiscrete')
```

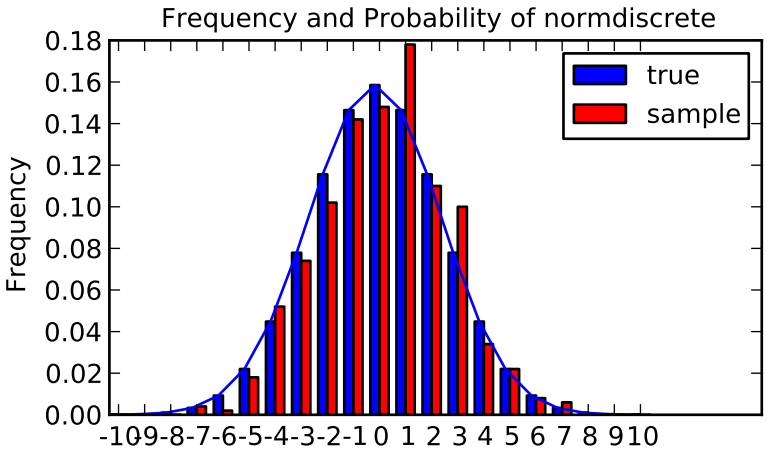
Now that we have defined the distribution, we have access to all common methods of discrete distributions.

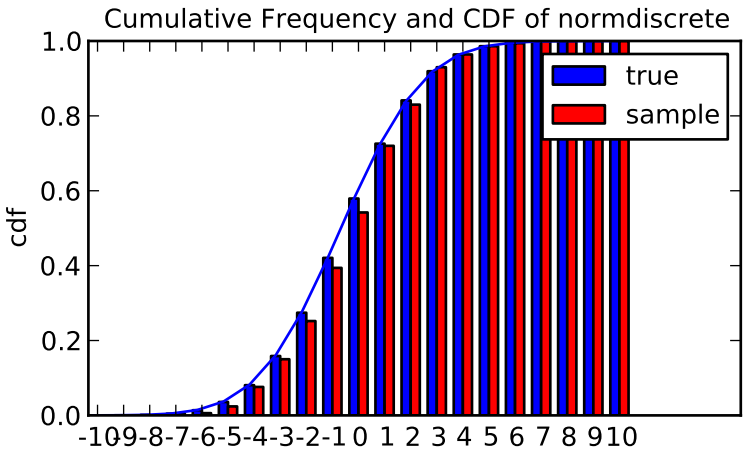
```
>>> print 'mean = %6.4f, variance = %6.4f, skew = %6.4f, kurtosis = %6.4f'% \
...       normdiscrete.stats(moments =  'mvsk')
mean = -0.0000, variance = 6.3302, skew = 0.0000, kurtosis = -0.0076

>>> nd_std = np.sqrt(normdiscrete.stats(moments='v'))
```

Testing the Implementation

Let's generate a random sample and compare observed frequencies with the probabilities.

```
>>> n_sample = 500
>>> np.random.seed(87655678)   # fix the seed for replicability
>>> rvs = normdiscrete.rvs(size=n_sample)
>>> rvsnd = rvs
>>> f, l = np.histogram(rvs, bins=gridlimits)
>>> sfreq = np.vstack([gridint, f, probs*n_sample]).T
>>> print sfreq
[[ -1.00000000e+01   0.00000000e+00   2.95019349e-02]
 [ -9.00000000e+00   0.00000000e+00   1.32294142e-01]
 [ -8.00000000e+00   0.00000000e+00   5.06497902e-01]
 [ -7.00000000e+00   2.00000000e+00   1.65568919e+00]
 [ -6.00000000e+00   1.00000000e+00   4.62125309e+00]
 [ -5.00000000e+00   9.00000000e+00   1.10137298e+01]
 [ -4.00000000e+00   2.60000000e+01   2.24137683e+01]
 [ -3.00000000e+00   3.70000000e+01   3.89503370e+01]
 [ -2.00000000e+00   5.10000000e+01   5.78004747e+01]
 [ -1.00000000e+00   7.10000000e+01   7.32455414e+01]
 [  0.00000000e+00   7.40000000e+01   7.92618251e+01]
 [  1.00000000e+00   8.90000000e+01   7.32455414e+01]
 [  2.00000000e+00   5.50000000e+01   5.78004747e+01]
 [  3.00000000e+00   5.00000000e+01   3.89503370e+01]
 [  4.00000000e+00   1.70000000e+01   2.24137683e+01]
 [  5.00000000e+00   1.10000000e+01   1.10137298e+01]
 [  6.00000000e+00   4.00000000e+00   4.62125309e+00]
 [  7.00000000e+00   3.00000000e+00   1.65568919e+00]
 [  8.00000000e+00   0.00000000e+00   5.06497902e-01]
 [  9.00000000e+00   0.00000000e+00   1.32294142e-01]
 [  1.00000000e+01   0.00000000e+00   2.95019349e-02]]
```





Next, we can test, whether our sample was generated by our normdiscrete distribution. This also verifies whether the random numbers are generated correctly.

The chisquare test requires that there are a minimum number of observations in each bin. We combine the tail bins into larger bins so that they contain enough observations.

```
>>> f2 = np.hstack([f[:5].sum(), f[5:-5], f[-5:].sum()])
>>> p2 = np.hstack([probs[:5].sum(), probs[5:-5], probs[-5:].sum()])
>>> ch2, pval = stats.chisquare(f2, p2*n_sample)

>>> print 'chisquare for normdiscrete: chi2 = %6.3f pvalue = %6.4f' % (ch2, pval)
chisquare for normdiscrete: chi2 = 12.466 pvalue = 0.4090
```

The pvalue in this case is high, so we can be quite confident that our random sample was actually generated by the distribution.