Bio 723: Clustering I

Paul Magwene

November 3, 2014

Contents

1 Dissimilarity measures

1.1 Dissimilarity measures in R

R includes a function, `dist()`, for calculating some of the most basic dissimilarity measures including Euclidean, Minkowski, and Manhattan metrics among others. The typical input to `dist()` is a data frame or matrix and a `method` argument specifying the type of distance measure to use. The `upper` argument specifies whether the upper diagonal of the calculated distance matrix should be printed (by default only the lower diagonal is printed).

To start with let's create a small 4×3 matrix where we can easily calculate the distances between the 4 points by pencil and paper.

```
# create a 4 x 3 matrix
z <- matrix(c(0,0,0,
              1,0,0,
              0,1,0,
              0,0,1), 4, 3, byrow=T)
dist(z)
```

```
         1        2        3
2 1.000000
3 1.000000 1.414214
4 1.000000 1.414214 1.414214
```

The default distance measure is Euclidean distance. Let's apply Manhattan distance to the same matrix.

```
dist(z, method='manhattan')
```

```
  1 2 3
2 1
3 1 2
4 1 2 2
```

1.2 Dissimilarity Measures in Python

```
import numpy as np # import numpy

z = np.matrix([[0,0,0],
               [1,0,0],
               [0,1,0],
               [0,0,1]], dtype=np.float)
z
```