submca.qc.elifeasex.subook.featslxn_gene_clustering_K1520_202000625

October 13, 2020

[1]:

```
from plotnine import *
import matplotlib

%matplotlib notebook
matplotlib.rcParams['figure.figsize'] = [8, 5]
```

[2]:

```
import pandas as pd
import numpy as np
import re
#from pip._internal import main as pipmain
#pipmain(['install', 'colour'])
from colour import Color
import operator
from sklearn.neighbors import NearestNeighbors
#spectral clustering
from sklearn.cluster import SpectralClustering
import scipy
```

[3]:

```
import numpy
```

[4]:

```
data = pd.read_csv("genegraph/submca.qc.counts.elifeasex.subook.
 ↪featslxn_forgenegraph.csv")
data = data.set_index("Unnamed: 0")
print(data.shape)
#data = data.transpose()
data = data[list(data.sum(axis=1)>=200)]
print(data.shape)
# get rid of gene name column
#data = data.drop(columns=["Unnamed: 0"])

genes = list(data.index)
data = data.transpose()

matrix = data.values

print(type(matrix))
```

```
(2000, 623)
(1797, 623)
<class 'numpy.ndarray'>
```

[5]:
```
len(genes)
```

[5]:
```
1797
```

[6]:
```
#expresses =
#genes = [x for x in genes[[index for index, x in enumerate(data.
 ↪sum(axis=0)>=200) if x == True]]]
#expresses = data
normalized = np.log(data+1)
normalized = normalized/normalized.mean(axis=0)
normalized = normalized.transpose()

normalized.shape
```

[6]:
```
(1797, 623)
```

[7]:
```
nbrs = NearestNeighbors(n_neighbors=6,metric="manhattan").fit(normalized.values)
```

[8]:
```
adj_mat = nbrs.kneighbors_graph(normalized.values)
```

[9]:
```
distances, indices = nbrs.kneighbors(normalized.values)
```

[10]:
```
sc = SpectralClustering(15, affinity='precomputed', n_init=1000,␣
 ↪assign_labels='discretize')
sc.fit(adj_mat)
print('spectral clustering')
print(sc.labels_)
```

```
/Users/vh3/miniconda3/envs/genegraph/lib/python3.8/site-
packages/sklearn/manifold/_spectral_embedding.py:212: UserWarning: Array is not
symmetric, and will be converted to symmetric by average with its transpose.
  adjacency = check_symmetric(adjacency)
```

```
spectral clustering
[11  1  1 …  9  1  1]
```

[11]:
```
with open("genegraph/subclusters.submca.elifeasex.subook.featslxn.k15.csv",'w')␣
 ↪as out:
    for index, gene in enumerate(genes):
        out.write(",".join([gene,str(sc.labels_[index])])+"\n")
```

[12]:
```
with open("genegraph/subgraph_submca.elifeasex.subook.featslxn.k15.dot",'w') as␣
 ↪graph:
    graph.write("graph genes{\n")
    for i, edges in enumerate(indices):
        for edge in edges:
            graph.write(genes[i]+" -- "+genes[edge]+";\n")
    graph.write("}")
```

[13]:

```
sc = SpectralClustering(20, affinity='precomputed', n_init=1000,␣
 ↪assign_labels='discretize')
sc.fit(adj_mat)
print('spectral clustering')
print(sc.labels_)
```

```
/Users/vh3/miniconda3/envs/genegraph/lib/python3.8/site-
packages/sklearn/manifold/_spectral_embedding.py:212: UserWarning: Array is not
symmetric, and will be converted to symmetric by average with its transpose.
  adjacency = check_symmetric(adjacency)
```

```
spectral clustering
[6 9 9 … 2 9 9]
```

[14]:

```
with open("genegraph/subclusters.submca.subook.elifeasex.featslxn.k20.csv",'w')␣
 ↪as out:
    for index, gene in enumerate(genes):
        out.write(",".join([gene,str(sc.labels_[index])])+"\n")
```

[ ]: