Carbon

Jiayu Wang

4.Factorization

1. Load data

```
#load module data
import data
#create a dataset object and read data from file sample.txt
sample = data.DataSet()
#To read nominal data, you have to add argument 'nominal',
default is 'numeric'
#Read data from 'sample.txt'
sample.read('sample.txt', 'numeric')
#create train dataset and test dataset using 1:10 hold out
train,test = data.holdOut(sample,0.1)
```

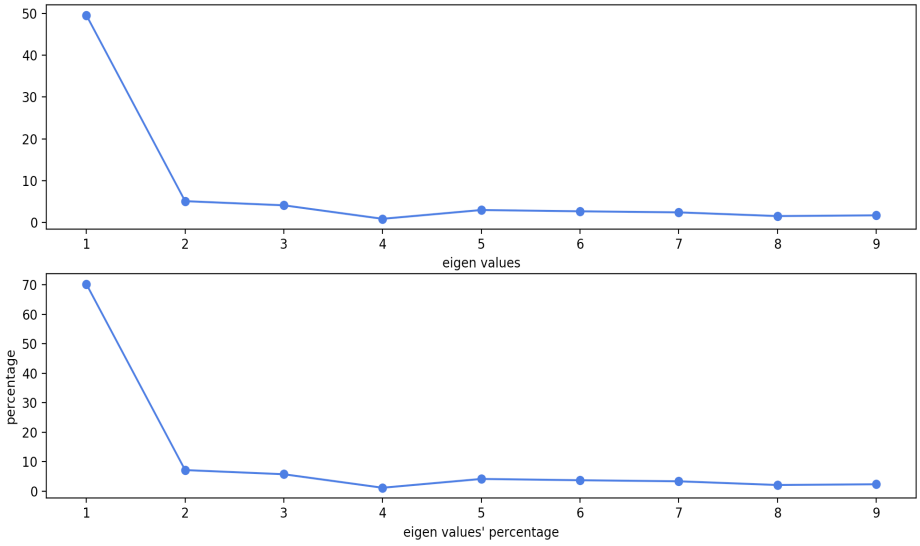
2. PCA

The algorithm only works for: numerical data

Parameters:

topNfeat: +int

```
#load pca module
from factorize import pca
#Find the frequent set with support bigger than 0.5
dataRedu, dataRecover = pca.pca(train.x, topNfeat = 5)
```




```
#dataRedu contains the reduced dimension data in new space
dataRedu
matrix([[  5.13380356e+00,   1.21247031e+00,  -1.53566333e+00,
          -1.01799592e+00,   5.90839965e-01],
        [  3.84335049e+00,  -1.99252297e-01,   1.99577675e+00,
        ...,
        [ -6.59705036e+00,   2.68400028e+00,  -1.21848752e+00,
          -9.04074040e-01,   3.13377195e+00],
        [ -7.69160051e+00,   1.47694596e+00,  -1.83005308e+00,
           1.25551269e+00,   4.03064968e+00]])
```

```
#dataInOri contains the reduced dimension data in original space
dataRecover
matrix([[  0.93352354,   1.10721433,   1.1980775 , ...,
2.34203796, 2.12665989,   0.77889049],
        ...,
        [  3.3711653 ,   8.18153978,   8.02416097, ...,
6.74482839, 4.25162364,   1.93840186]])
```

```
#fast start
import data
sample = data.DataSet()
sample.read('sample.txt')
train,test = data.holdOut(sample,0.1)
from imp import reload
```