Carbon

Jiayu Wang

2.Clustering

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. kMeans

A kMeans algorithm which also support bi-kMeans.

The algorithm only works for: numerical data

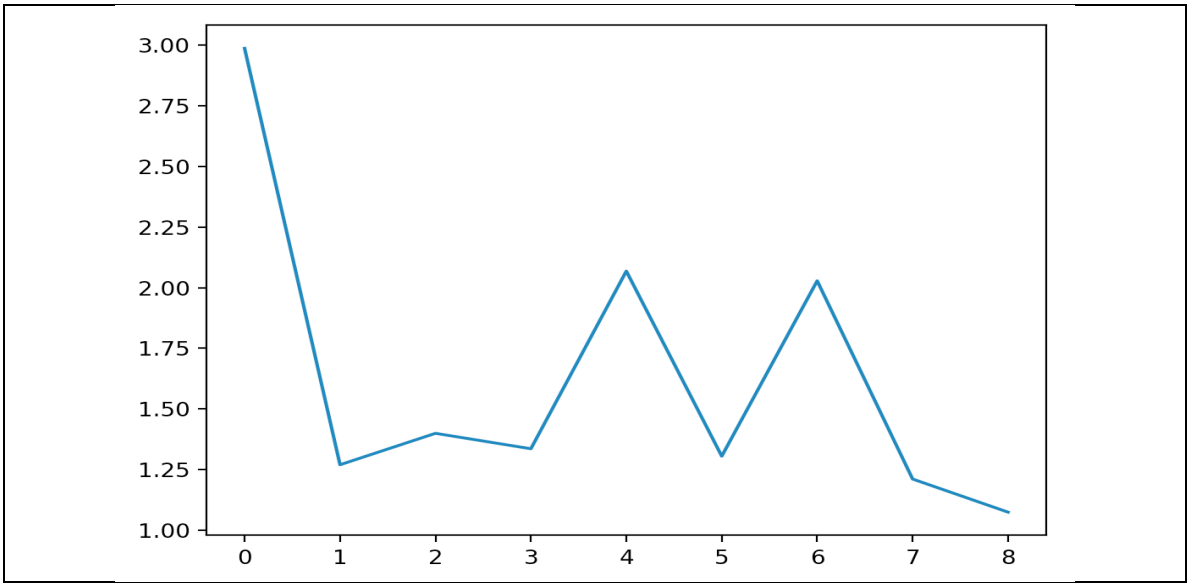
Parameters:

k: int from 1 to inf

dist: 'euclidean'

method: 'KMeans', 'biKMeans'

```
#load kMeans module
from clustering import kMeans
#create an instance with k=3, euclidean distance, KMeans
instance = kMeans.build(k=3, dist='euclidean', method='KMeans')
#train the classifier with train data and k=4
centroids,clusters = instance.cluster(train.x)
#centroids contains the centers of each cluster
centroids
matrix([[ 7.25555556,  4.87777778,  5.12222222,  4.9       ,
4.02222222, 9.04444444,  5.27777778,  3.75555556,  1.7       ],
        [ 2.96933962,  1.26650943,  1.4009434 ,  1.31603774,
2.0754717 , 1.29009434,  2.02830189,  1.22169811,  1.07075472],
        [ 7.04310345,  8.40517241,  8.05172414,  6.72413793,  6.5
, 7.17241379,  6.89655172,  7.88793103,  3.3362069 ]])
#clusters contains the cluster labels and distances from the
centers
clusters
matrix([[   1.        ,    5.85476927],
        [   1.        ,    6.9962787 ],
        [   2.        ,  125.31599287],
        ...,
        [   2.        ,   49.50564804],
        [   2.        ,   62.02288942],
        [   2.        ,   53.57461356]])
#to view the clusters by plot 2 features
kMeans.view(train, centroids, clusters)
```




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

reload(CART)
clf = CART.build()
clf.train(train,tolS=1,tolN=4,model=False)
clf.classify([5, 4, 4, 5, 7, 10, 3, 2, 1])
clf.view('Uniformity of Cell Size','Uniformity of Cell
Shape',train)
```