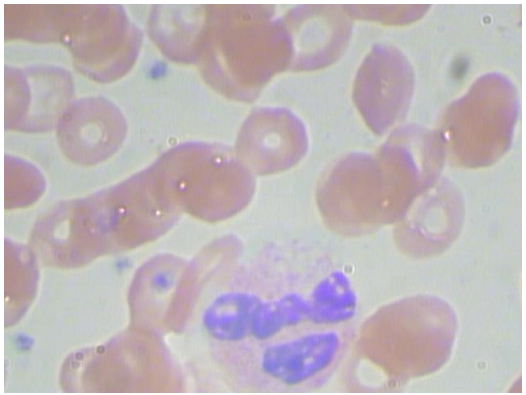
ML Final Project

Shuyu Chen, Changlin Jiang, Chaoji Zuo

Classification

Dataset Description



```
- <object>
    <name>WBC</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
  - <bndbox>
      <xmin>260</xmin>
      <ymin>177</ymin>
      <xmax>491</xmax>
      <ymax>376</ymax>
    </bndbox>
</object>
```

The White Blood Cells dataset consists of 410 images, with 4 classes: Eosinophil, Lymphocyte, Monocyte, and Neutrophil.

Reduce the data dimensionality by cropping the exact white blood cell from each images.

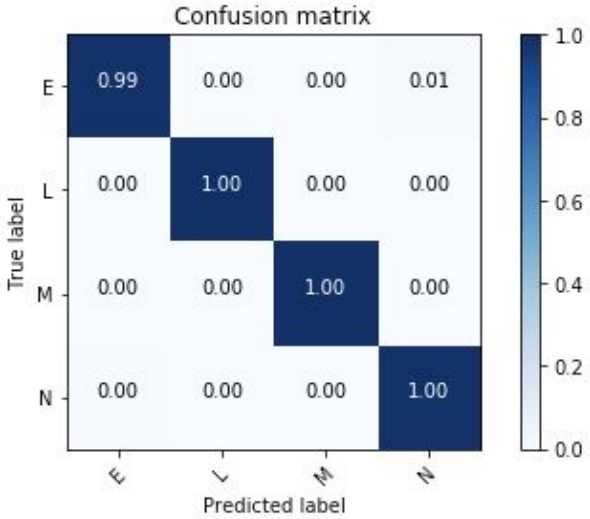
KNN approach

Determine the class of the test sample by sorting the distance of sample to its neighboring points.

Random Forest approach

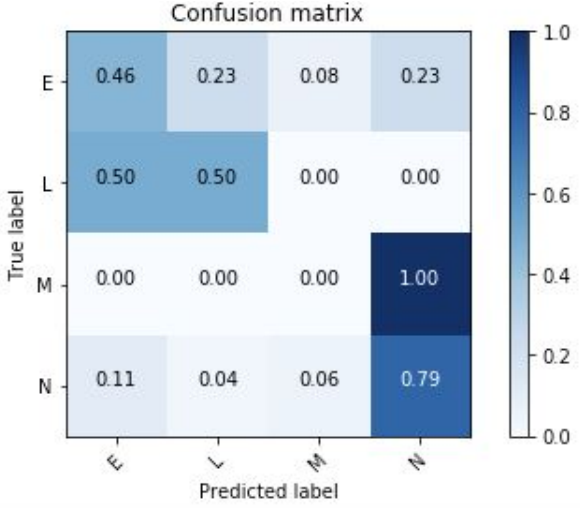
first try on whole data

```
precision score : 0.9859154929577465
```

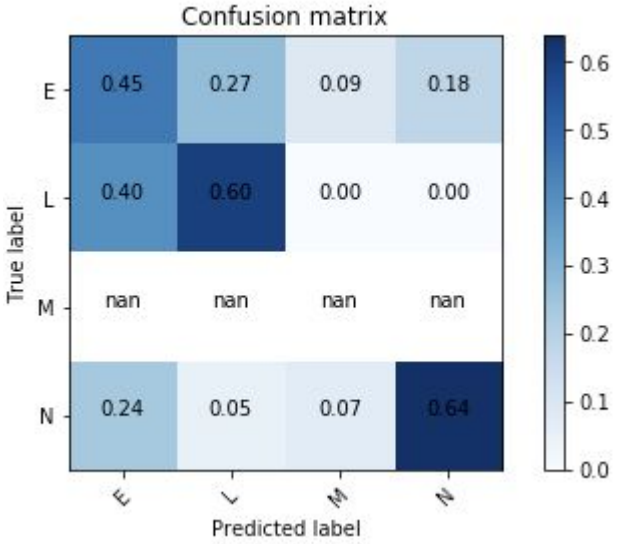


try on test set

```
average precision: 0.6225352112676057
```



Test with optimized parameters




```
Best parameters:
{'max_depth': 30, 'max_features': 'auto', 'n_estimators': 100}
```

```
average precision: 0.6535211267605633
```

Naive Bayes approach

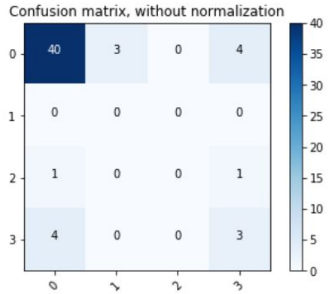
Gaussian Naive Bayes approach

- Discussion:Image data input is likely to be iid.The model is derived by real world so it's probably a Gaussian model.
- Result trained with first 300 samples and tested by the remaining samples(n=56):

  0-1 Loss= 13 Accuracy= 0.7678571428571428

- Cross validation(k=5),confusion matrix:

  Average accuracy= 0.5679434159053987



Regression

Dataset description

- True values of concentrations of some substance
- Concentrations of some substance responsed by a sensor
- Temperature,humidity
- Measured along with time

Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)
3/10/2004	18:00:00	2.6	1360	150	11.9	1046
3/10/2004	19:00:00	2	1292	112	9.4	955

NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
166	1056	113	1692	1268	13.6	48.9	0.7578
103	1174	92	1559	972	13.3	47.7	0.7255

Challenge

- Dataset attributes not explicitly tagged
- Unknown model
- Effort

  - Go back to the paper to find out which attribute is the label(by Changlin Jiang)

  - Implement random forest algorithm to learn the dataset regardless of model(by Chaoji Zuo)

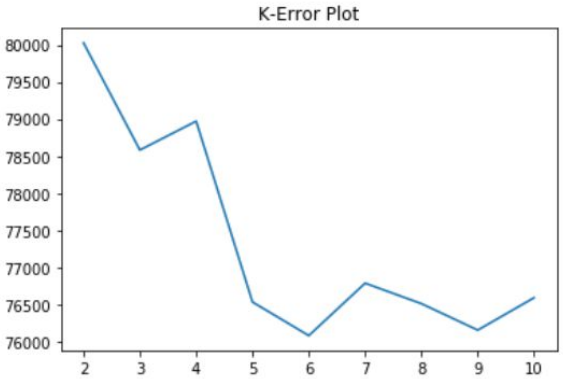
Least Square Regression approach

Assumption:Linear model

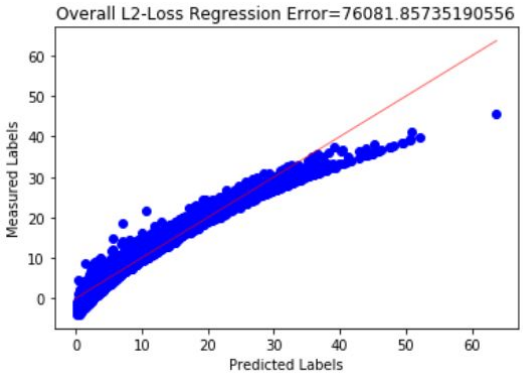
- Input:Measured concentrations of CO and 4 other substance,temperature,humidity
- Labels:True benzene concentration
- Purpose:Estimate concentrations with measured concentrations of CO,NMHC,NOx,NO2,O3,temperature,humidity.

Result

Best cross validation K=6

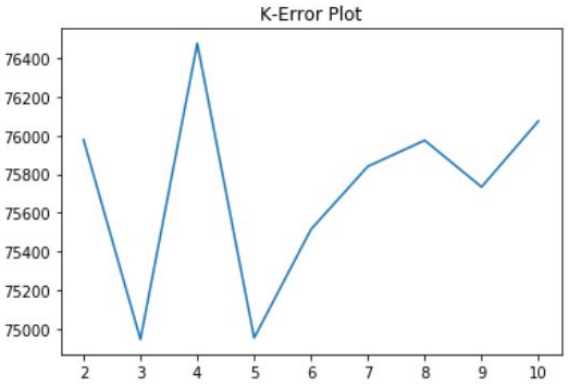


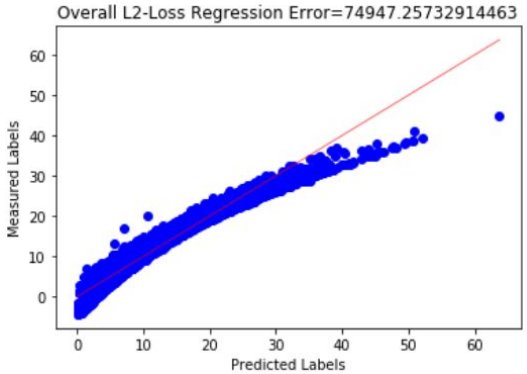
Predicted versus true scatter plot



Result

Get rid of temperature and humidity in input data,now best CV k=3,smaller error derived

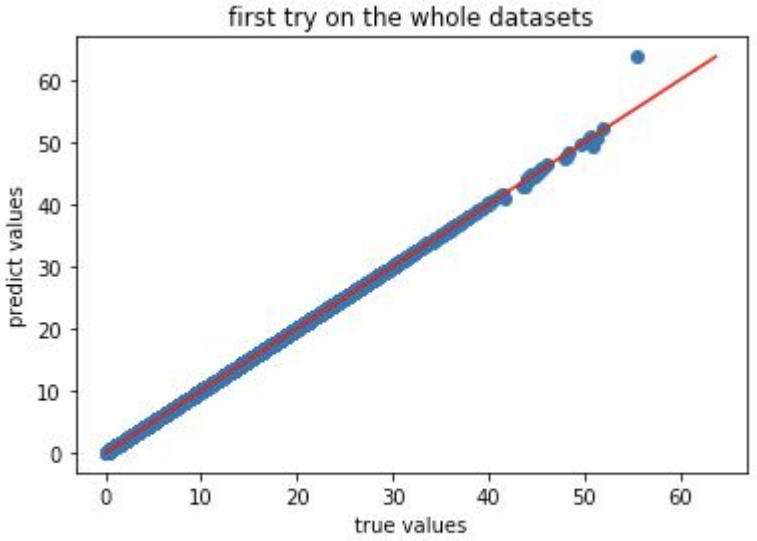




LASSO Regression approach

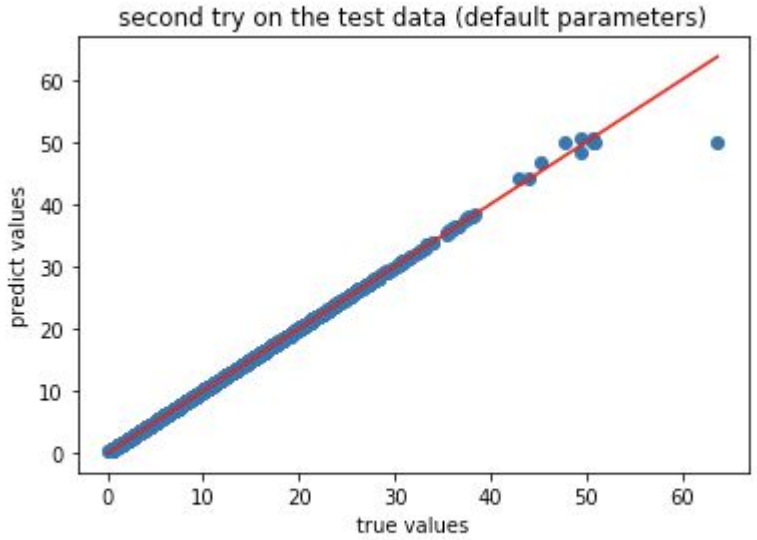
Random Forest approach

first try on whole datasets



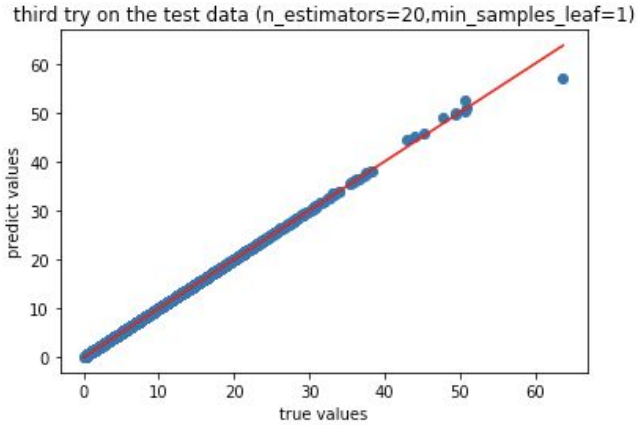

```
2-norm error: 8.681710660923919
R^2 score of predict values: 0.9998489365496641
```

predict on test data




```
2-norm error: 14.214309507124208
R^2 score of predict values: 0.9988393238910085
```

third try by parameters optimization




```
2-norm error: 7.34916772622886
R^2 score of predict values: 0.9996897328273662
```

I really got a better solution using the best parameters of "min_samples_leaf" and "n_estimators".

But the progress was not very significant, because the former soultion is great enough.

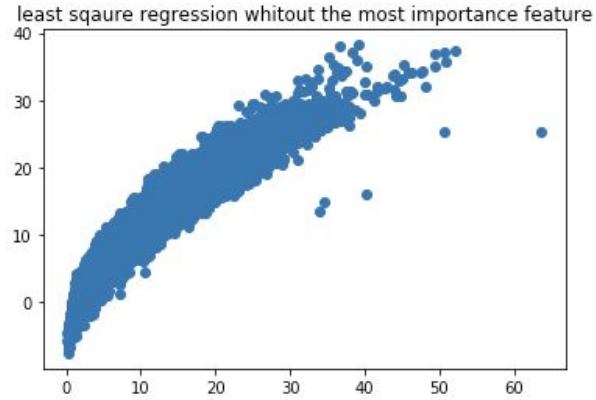
model review

```
shape of decision path: (3357, 15240)
```

```
CO measured,NMHC measured,NOx measured,NO2 measured,O3 measured,temp,RH,AH
```

```
feature importances:
[4.00605965e-04 9.99040707e-01 8.42519545e-05 1.10125513e-04
 1.10173806e-04 1.47280047e-05 1.46982777e-04 9.24247248e-05]
```

```
error: 226.72020689446524
```



Clustering

Dataset: Black-Friday

raw-data

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	\
324155	1001891	P00345742	M	46-50	1	C	
384692	1005193	P00084842	M	36-45	12	B	
34262	1005282	P00183642	F	18-25	4	B	
328396	1002590	P00128342	M	18-25	4	A	
270293	1005650	P00367042	F	36-45	12	B	

	Stay_In_Current_City_Years	Marital_Status	Product_Category_1	\
324155	3	1	1	
384692	2	1	8	
34262	1	1	4	
328396	0	0	5	
270293	2	1	8	

	Product_Category_2	Product_Category_3	Purchase
324155	2.0	15.0	11700
384692	16.0	NaN	6051
34262	5.0	9.0	766
328396	12.0	14.0	3480
270293	NaN	NaN	6164

Pre-processing

mean and mode data

User_ID	Occupation	Age	City_Category	Marital_Status	Product_CateGory_1	\
1004956	15	36-45	B	1	8	
1000839	0	26-35	A	0	8	
1003510	4	18-25	B	1	5	
1003016	12	18-25	A	0	1	
1005555	10	0-17	B	0	1	

User_ID	Stay_In_Current_City_Years	times	Gender_M	Purchase
1004956	1	120	1	9324.600000
1000839	2	435	1	10761.390805
1003510	1	32	0	9913.406250
1003016	1	18	1	11067.111111
1005555	2	276	1	9055.329710

Challenge

- ***how to handle the discrete attributes?***

  categorical data

  continuous data

- ***how to evaluate?***

  distance calculation

  show difference

$$\mathrm{JC} = \frac{a}{a+b+c} \ .$$

$$\mathrm{VDM}_p\left(a,b\right) = \sum_{i=1}^{k} \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p \ .$$

try some things

Jaccard coefficient

one-hot encoding

```
categorical data:
        Gender  Occupation    Age City_Category  Marital_Status  \
User_ID
1004322      F           6  51-55             B               0
1004518      M           0  26-35             C               0

         Product_CateGory_1 Stay_In_Current_City_Years
User_ID
1004322                   8                         1
1004518                   3                        4+
one-hot encoding data:
         0-17  18-25  26-35  36-45  46-50  51-55  55+  A  B  C  \
User_ID
1002359     0      0      0      0      0      0    1  0  0  1
1005850     0      0      1      0      0      0    0  0  0  1

              ...         8  10  11  12  13  15  16  18  Gender_M  \
User_ID       ...
1002359       ...         0   0   0   0   0   0   0   0         1
1005850       ...         0   0   0   0   0   0   0   0         0

         Marital_Status
User_ID
1002359               1
1005850               0

[2 rows x 53 columns]
```

K-mode approach

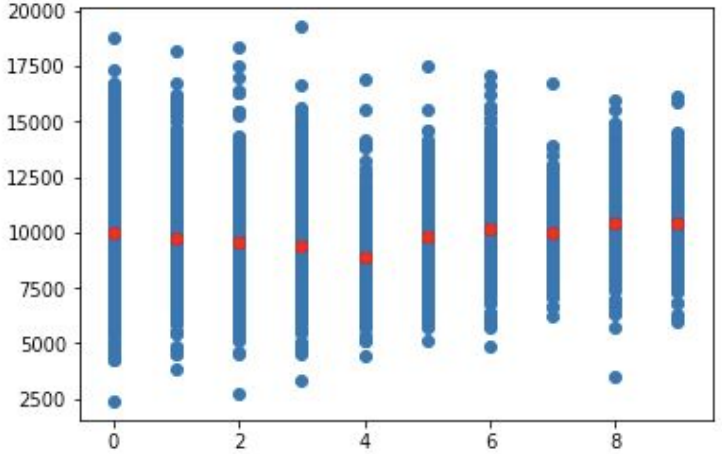
```
n_cluster = 2 : 0.5000269875392017
n_cluster = 3 : 0.5294875041896924
n_cluster = 4 : 0.5534965418820955
n_cluster = 5 : 0.5645384012323531
n_cluster = 6 : 0.5720855312627525
n_cluster = 7 : 0.5890558734522645
n_cluster = 8 : 0.6010495871774273
n_cluster = 9 : 0.6128901655705694
```

part features

User_ID	Gender	Occupation	City_Category
1005314	F	0	A
1002289	F	1	C

```
average jc distance in selected features: 0.5
average ec distance in all one-hot features: 1.3071067811865476
```

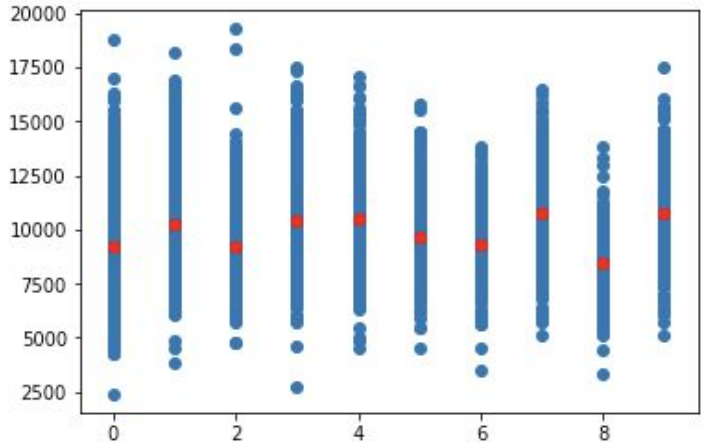
1.2846433962579515



All features

```
average jc distance in selected features: 0.35
average ec distance in all one-hot features: 1.8432220400206423
```

User_ID	Gender	Occupation	Age	City_Category	Marital_Status	Product_CateGory_1
1004639	M	11	36-45	B	0	5
1004422	M	5	26-35	A	1	5



ROCK approach

Challenges and effort

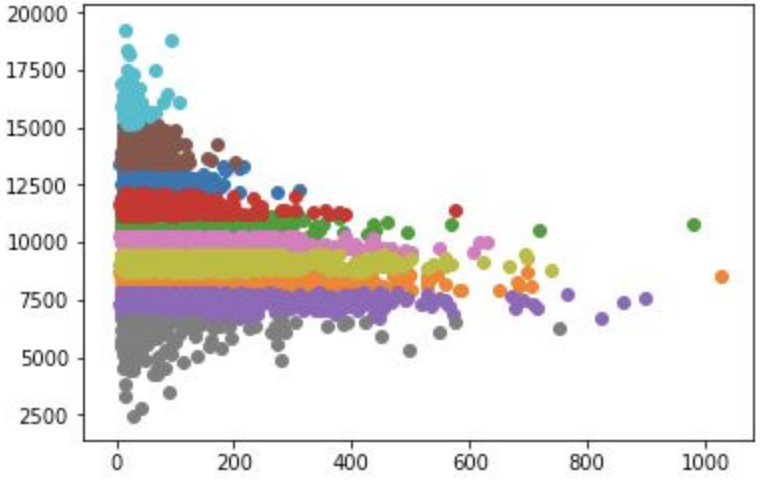
- Most attributes in the dataset are discrete,all methods mentioned in class not working
- Searched for papers building categorical clustering algorithm
- Picked ROCK algorithm,tried to implement from library,but library doing totally different work and only works for continuous features
- BUILD THE WHOLE ALGORITHM BY MYSELF

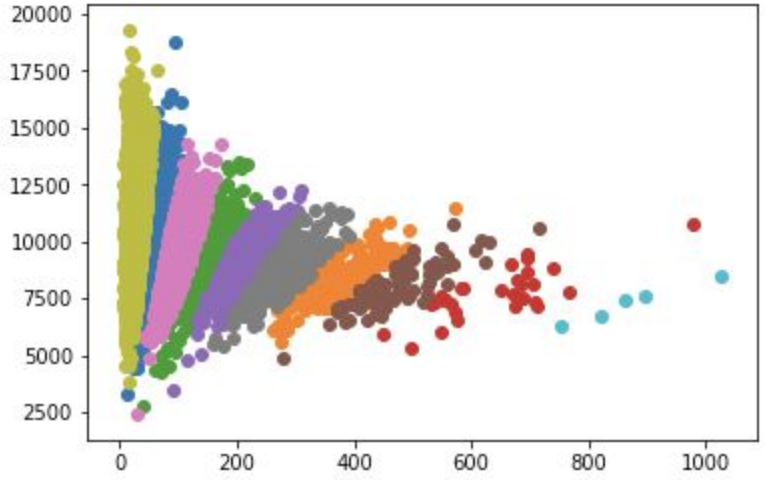
ROCK algorithm

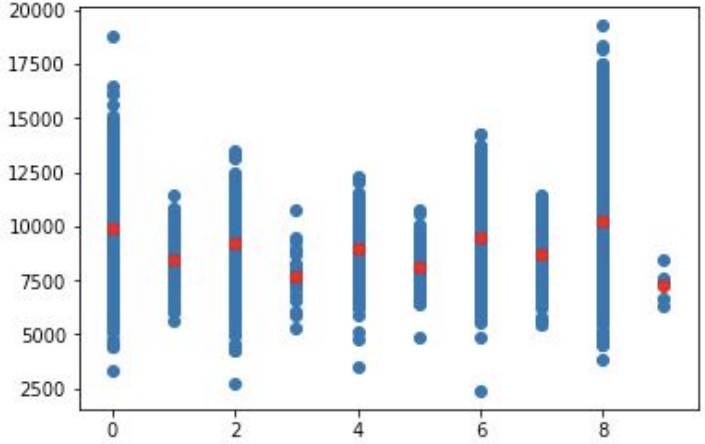
- Neighbors:Jaccard score and threshold
- Link(Ci,Cj) and Goodness(Ci,Cj)
- Clusters merging
- Disadvantage:computationally expensive O(n^3)
- Result for first 300 samples:10 clusters (ni=25,124,115,16,5,2,6,2,2,3)

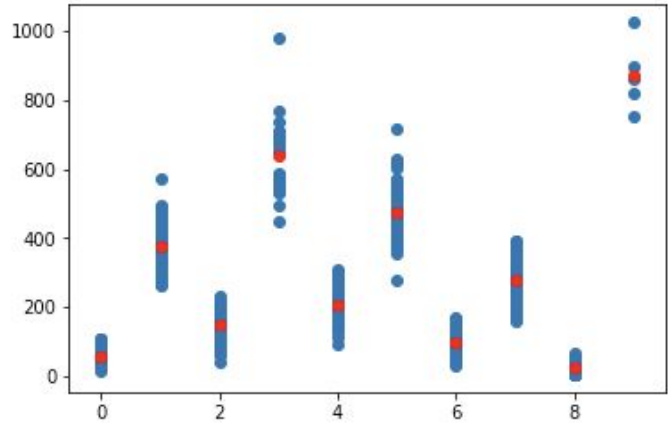
```
Number of clusters<=10
[[0, 18, 83, 190, 206, 48, 72, 96, 115, 149, 196, 5, 10, 36, 66, 197, 215, 103, 163, 192, 244, 272, 111, 135], [1, 13, 16, 24,
28, 49, 58, 63, 81, 82, 84, 97, 110, 114, 116, 117, 125, 139, 170, 177, 181, 184, 202, 221, 247, 251, 255, 274, 289, 175, 238,
278, 283, 91, 128, 140, 14, 218, 226, 257, 258, 80, 166, 223, 161, 267, 19, 252, 256, 159, 213, 70, 122, 185, 188, 195, 107, 2,
231, 55, 143, 118, 209, 266, 60, 210, 224, 74, 178, 38, 158, 89, 98, 127, 141, 241, 54, 8, 245, 240, 131, 187, 172, 263, 225, 2
95, 296, 37, 67, 65, 148, 198, 186, 234, 269, 292, 11, 237, 30, 87, 108, 271, 182, 168, 298, 20, 35, 132, 222, 230, 165, 205, 2
12, 219, 254, 40, 105, 208, 216, 138, 21, 33, 59, 153, 201, 193, 220, 43, 44, 104, 176, 273, 157, 293, 249, 53, 99, 156, 229, 2
61, 285], [3, 6, 12, 22, 41, 45, 100, 121, 126, 144, 145, 173, 279, 73, 90, 200, 243, 259, 265, 268, 294, 79, 112, 113, 130, 6
2, 207, 204, 277, 77, 102, 211, 253, 291, 57, 106, 286, 236, 250, 39, 92, 46, 235, 61, 194, 246, 297, 4, 7, 26, 120, 129, 137,
169, 232, 264, 287, 32, 124, 262, 203, 227, 78, 47, 147, 68, 86, 270, 134, 42, 69, 76, 85, 155, 123, 179, 189, 9, 15, 27, 75, 1
52, 34, 51, 151, 275, 88, 101, 214, 288, 142, 260, 23, 25, 239, 119, 146, 162, 191, 282], [17, 94, 136, 29, 31, 248, 281, 284,
50, 52, 171, 183, 242, 280, 160, 233], [56, 64, 167, 199], [71, 174], [93, 95, 133, 164, 109, 180], [150, 290], [154, 228], [21
7, 276, 299]]
```

K-means approach


20000
17500
15000
12500
10000
7500
5000
2500
0
200
400
600
800
1000


20000
17500
15000
12500
10000
7500
5000
2500
0
200
400
600
800
1000


20000
17500
15000
12500
10000
7500
5000
2500
0
2
4
6
8


1000
800
600
400
200
0
0
2
4
6
8

Discussion about evaluation metrics for categorical clustering

- In K-modes, is the euclidean distance on one-hot transformed data really meaningful?
- In ROCK,is there anything we can do to evaluate it?

  - Found from paper:ANOVA distance

  - No libraries available,complex to go deeper and implement,computationally expensive