Assignment: Appendix

(Dataset Format)

Datasets for Clustering & Classification

Dataset	N	D	C
toydata	400	2	NA
iris	150	4	3
glass	214	9	6
vowel	528	10	11
vehicle	846	18	4
Letter	16000	16	26
DNA	2000	180	3

D: # of attributes; N: # of samples; C: # of classes; NA: no class label available

Input File Format of Dataset

N D C
sample1_attribute#1 sample1_attribute#2 … sample1_attribute#D sample1_class_label
sample2_attribute#2 sample2_attribute#2 … sample2_attribute#D sample2_class_label
…
sampleN_attribute#1 sampleN_attribute#2 … sampleN_attribute#D sampleN_class_label

Comments:
If C=0, it means that the dataset does not contain the class label information.
If C=1, it means that the dataset contains the class label information.

Note for Dataset used for Clustering

For data clustering, we usually assume the class label is unknown. Therefore, the class label will not be used. DO remember to discard the class label when you read a file in your program for clustering.

Note for Dataset used for Classification

For data classification, we provide two files for each dataset: one is the training set and the other is the test set. For the test set, the class label information is also provided, which should only used for the performance evaluation purpose.

Sample Output Format f Classification Task

Accuracy = 95.15%
Time cost = 15.31 seconds

(Also you should output a file that includes the predicted class label results).

Output File Format of Clustering results

Below shows an example output file for your program for a clustering task with 4 clusters of 10 data examples in 2-D.

Within-cluster SSE = 0.2655
(Final Center Points:

0	[-0.975313 0.857799]	SSE = 0.0274
1	[0.85796 0.0531374]	SSE = 0.1171
2	[0.872382 0.564819]	SSE = 0.1084
3	[-0.961812 0.479666]	SSE = 0.0126

)

Point	Center	Squared Dist
0	2	0.0464422
1	2	0.0619897
2	3	0.0126129
3	0	0.00525171
4	1	0.00585648
5	0	0.0221774
6	1	0.0500088
7	1	0.0200706
8	1	0.00378334
9	1	0.0373406