Name: Sai Koumudi Student ID: 800959723

Homework - Assignment -1

ITCS6157 – Visual Databases

1.K-means Clustering:

(A) K-means Breast Cancer dataset:

Load the dataset into R.

```
bcw <- read.csv('D:/hw/dataset3/bcw.csv')
names(bcw)
```

After loading dataset, retrieving the column names from the dataset.

```
> bcw <- read.csv('D:/hw/dataset3/bcw.csv')
> names(bcw)
 [1] "v2"  "v3"  "v4"  "v5"  "v6"  "v7"  "v8"  "v9"  "v10" "v11" "v12"
[12] "v13" "v14" "v15" "v16" "v17" "v18" "v19" "v20" "v21" "v22" "v23"
[23] "v24" "v25" "v26" "v27" "v28" "v29" "v30" "v31" "v32"
>
```

Now we will split attributes into two variables,
Variable1 = x
Variable2= y
Here, y is the decision attribute, which have values as M or B. x contains remaining all columns.

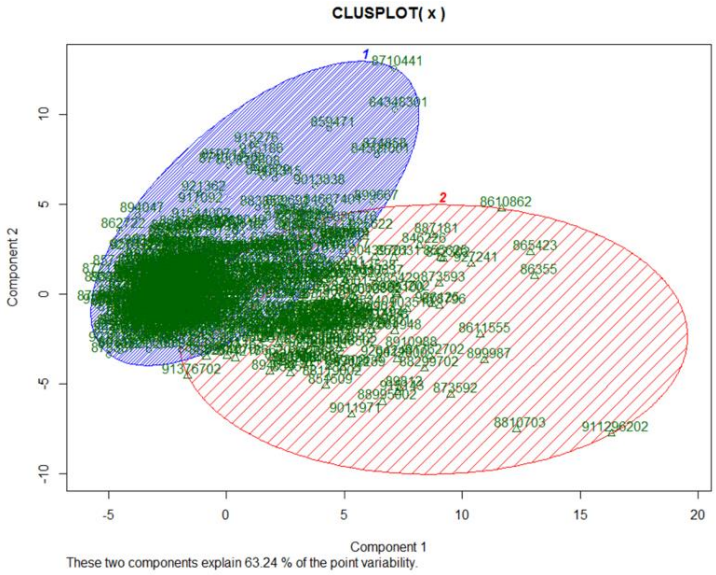
Step1: K-means clustering:

```
> bcw_kmeans
K-means clustering with 2 clusters of sizes 131, 438

Cluster means:
        v3       v4        v5        v6        v7         v8         v9
1 19.37992 21.69458 128.23130 1185.9298 0.1012946 0.14861298 0.17693947
2 12.55630 18.57037  81.12347  496.0619 0.0948845 0.09109982 0.06243776
         v10       v11        v12       v13       v14       v15       v16
1 0.10069878 0.1915397 0.06060290 0.7428038 1.222538 5.250580 95.67817
2 0.03343254 0.1780580 0.06345402 0.3041909 1.215153 2.152881 23.78529
          v17        v18        v19        v20        v21         v22
1 0.006598687 0.03217669 0.04241977 0.01567398 0.02030397 0.003953389
2 0.007173263 0.02347469 0.02874551 0.01063632 0.02061358 0.003747503
       v23       v24       v25       v26       v27       v28       v29
```

```
          v17         v18         v19         v20         v21          v22
1 0.006598687 0.03217669 0.04241977 0.01567398 0.02030397 0.003953389
2 0.007173263 0.02347469 0.02874551 0.01063632 0.02061358 0.003747503
       v23      v24       v25       v26       v27       v28       v29
1 23.70947 28.91267 158.49618 1753.0229 0.1404247 0.3577577 0.4493061
2 14.04390 24.70954  91.93751  619.6479 0.1299591 0.2233118 0.2192149
         v30       v31        v32
1 0.19243107 0.3118817 0.08616550
2 0.09132984 0.2835537 0.08328194
```

```
Clustering vector:
  [1] 1 1 1 2 1 2 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 1 2 1 1 1 1 2 1 1 1
 [36] 1 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2
 [71] 1 2 1 2 2 1 2 1 1 2 2 2 1 1 2 1 2 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2
[106] 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 2 2 1 2 1 2 2 2 2 1 2 2 2 2 2
[141] 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 1 1 2 1 2 2 1 1 2 2 2 2 2 2
[176] 2 2 2 2 2 1 1 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 1 2 2 1 1 2 2 2 2 1 2 2
[211] 1 2 1 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 1 1 2 1 2 2 2 2 1
[246] 2 2 2 2 2 1 2 1 1 1 2 1 2 1 2 1 1 1 2 1 1 2 2 2 2 2 2 1 2 1 2 2 1 2 2
[281] 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2
[316] 2 2 1 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 2 1 2 2 2 1 2 2 2 2 2
[351] 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 1 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2
[386] 2 2 2 2 1 2 2 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2
[421] 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 1 2 2 1 2 1 2 2 1 2 1 2 2 2
[456] 2 2 2 2 2 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2
[491] 2 1 1 2 2 2 2 2 2 1 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 1 2 2 2
[526] 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[561] 2 2 2 1 1 1 2 1 2

Within cluster sum of squares by cluster:
[1] 49383423 28559677
 (between_SS / total_SS =  69.6 %)
```

Step2: Plotting cluster plot for clustering variables.



Results:

- As we already know the class to which each record belongs to, now we compare the clustering results to the original classes. This gives a matrix.
- Now, we can compare the clustering results to that of the original classes. This gives us a matrix as to how the clustering has fared.

```
> table(y,bcw_kmeans$cluster)

y     1   2
  B 356   1
  M  82 130
```

- Now, map cluster 1 to “B” and cluster 2 to “M”. Generating confusion matrix and accuracy results. Y is the original class label and bcw$x is mapped class labels obtained from k-means clustering.

```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 356   1
         M  82 130

               Accuracy : 0.8541
                 95% CI : (0.8224, 0.8821)
    No Information Rate : 0.7698
    P-Value [Acc > NIR] : 3.412e-07

                  Kappa : 0.6618
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8128
            Specificity : 0.9924
         Pos Pred Value : 0.9972
         Neg Pred Value : 0.6132
             Prevalence : 0.7698
         Detection Rate : 0.6257
   Detection Prevalence : 0.6274
      Balanced Accuracy : 0.9026

       'Positive' Class : B
```

Therefore, the accuracy of k-means clustering algorithm on breast cancer data = 85.41%

One instance of B wrongly clustered into class M

82 instances of M wrongly clustered into B class.

Name: Sai Koumudi Student ID: 800959723

(B) k-means-Iris dataset:

This data set contains 50 instances with three different classes.

The classes: Setosa, Versicolour, Virginica

Attributes:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

Step1: Loading dataset

```
iris <- read.csv('D:/hw/dataset2/iris.csv')
names(iris)
```

Step2: For clustering, only attributes are taken.

For analysis species(classes) data is used. Splitting dataset into two attributes x and y

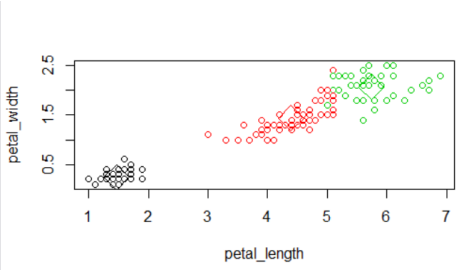
Clustering using k-means function:

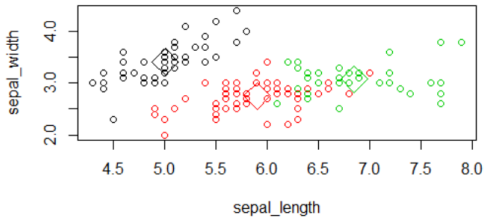
```
kmeans_iris <- kmeans(x,3)
kmeans_iris
```

Number of centers =3

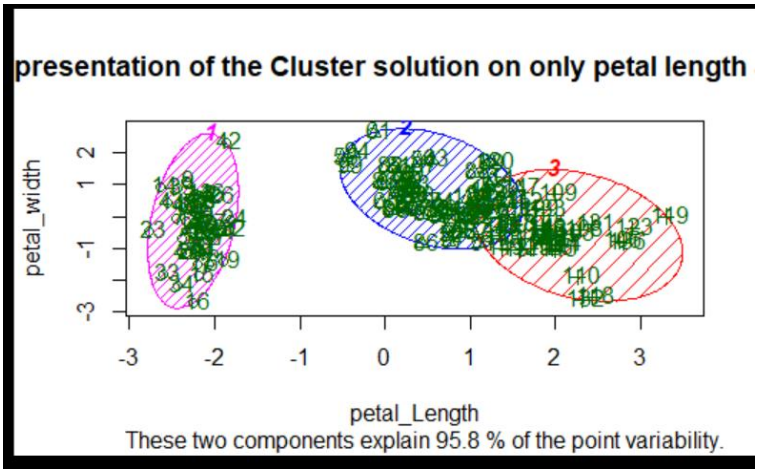
Step3:

Plotting graphs for three clusters:





Cluster plots:



Results:

- As we already know the class to which each record belongs to, now we compare the clustering results to the original classes. This gives a matrix.
- Now, we can compare the clustering results to that of the original classes. This gives us a matrix how the clustering has done.

```
table(y,kmeans_iris$cluster)
```

- As we know that y is the original class labels, we are comparing them to the clustered classes 1,2,3.

```
y                1  2  3
  setosa         0 50  0
  versicolor     2  0 48
  virginica     36  0 14
```

- From the above table, we can observe that, two records of versicolor fall into the cluster of virginica and 14 records of virginica fall into the cluster of versicolor.
- From above table, mapping class values to that particular levels. Mapping Cluster 2 to setosa, 3 to versicolor and 1 to virginica
- y is the original class label and iris$species is mapped class labels obtained from k-means clustering.

```
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0         14        36

Overall Statistics

               Accuracy : 0.8933
                 95% CI : (0.8326, 0.9378)
    No Information Rate : 0.4133
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.84
 Mcnemar's Test P-Value : NA
```

```
Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.7742           0.9474
Specificity                 1.0000            0.9773           0.8750
Pos Pred Value              1.0000            0.9600           0.7200
Neg Pred Value              1.0000            0.8600           0.9800
Prevalence                  0.3333            0.4133           0.2533
Detection Rate              0.3333            0.3200           0.2400
Detection Prevalence        0.3333            0.3333           0.3333
Balanced Accuracy           1.0000            0.8757           0.9112
```

- From the above results, it is observed that accuracy of k-means clustering algorithm = 89.33%
- Two instances of versicolor are wrongly clustered.
- 14 instances of virginica are wrongly clustered.

(C) K-means YearPredicitonMSD Dataset:

Loading data into R studio.

YP_MSD <- read.csv('D:/hw/dataset1/YearPredictionMSD/YearPredictionMSD.csv')

Retrieving column names:

names(YP_MSD)

```
 [1] "V1"  "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11" "V12" "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20"
[21] "V21" "V22" "V23" "V24" "V25" "V26" "V27" "V28" "V29" "V30" "V31" "V32" "V33" "V34" "V35" "V36" "V37" "V38" "V39" "V40"
[41] "V41" "V42" "V43" "V44" "V45" "V46" "V47" "V48" "V49" "V50" "V51" "V52" "V53" "V54" "V55" "V56" "V57" "V58" "V59" "V60"
[61] "V61" "V62" "V63" "V64" "V65" "V66" "V67" "V68" "V69" "V70" "V71" "V72" "V73" "V74" "V75" "V76" "V77" "V78" "V79" "V80"
[81] "V81" "V82" "V83" "V84" "V85" "V86" "V87" "V88" "V89" "V90" "V91"
```

From the above column names, column v1 is the decision attribute. It is the class label which contains 90 classes: year 1991 to year 2011

The attributes are divided into two variables x and y.y is the decision attribute which is v1 column.

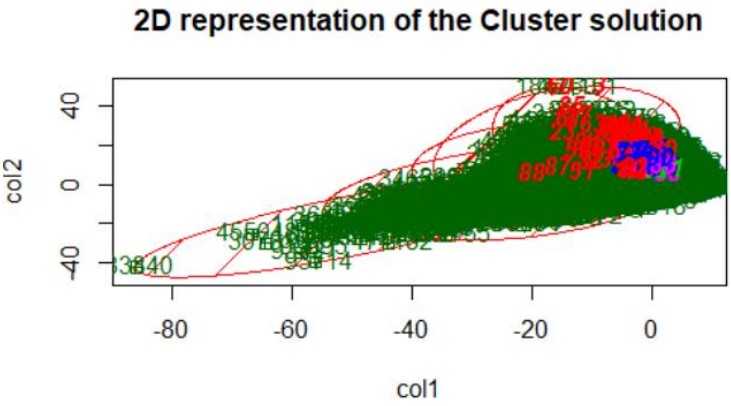
x contains remaining all columns.

k-means clustering:

```
kmeans_YP_MSD <- kmeans(x,90)
kmeans_YP_MSD
```

Since we know that there are 90 classes, we have taken 90 centers.

Cluster plot:



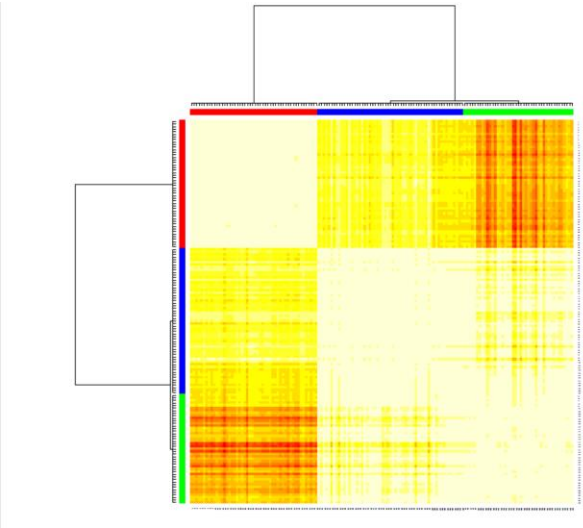
2) AP-Clustering

(A) AP-clustering -Iris dataset:

- Import dataset into R studio. Split the attributes into two different variables
- Now, we know that Iris data has three classes. Using the minimum off-diagonal similarities, we get.

```
APResult object

Number of samples     =  150
Number of iterations  =  126
Input preference      =  -50.2
Sum of similarities   =  -84.49
Sum of preferences    =  -150.6
Net similarity        =  -235.09
Number of clusters    =  3

Exemplars:
   8 56 113
Clusters:
   Cluster 1, exemplar 8:
      1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
      27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
      50

   Cluster 2, exemplar 56:
      52 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
      76 79 80 81 82 83 84 85 86 88 89 90 91 92 93 94 95 96 97 98 99 100
      102 107 114 120 122 124 127 128 134 139 143 150
   Cluster 3, exemplar 113:
      51 53 77 78 87 101 103 104 105 106 108 109 110 111 112 113 115 116
      117 118 119 121 123 125 126 129 130 131 132 133 135 136 137 138 140
      141 142 144 145 146 147 148 149
```

Heat map can be obtained as follows



- From the above heatmap we can say that, three clusters are reasonable. As the dataset has three classes, this can be concluded as clustering is done well.
- From the cluster data, observed that 13 instances of versicolor wrongly grouped into virginica. And 5 instances of versicolor have been wrongly grouped into virginica. 18 instances have been wrongly clustered.
- Therefore, error rate is 18/150 * 100 = 12%
- Accuracy = 88%

(B) AP Clustering - Breast Cancer dataset:

Load data into R studio and split the columns into two variables.

Run Ap-clustering algorithm. using apcluster library in R.

```
AP_bcw <- apcluster(negDistMat(r=2),x, q= 0)
AP_bcw
> AP_bcw

APResult object

Number of samples     =  569
Number of iterations  =  143
Input preference      =  -22458963
Sum of similarities   =  -47511870
Sum of preferences    =  -67376888
Net similarity        =  -114888758
Number of clusters    =  3

Exemplars:
   122 498 504
```

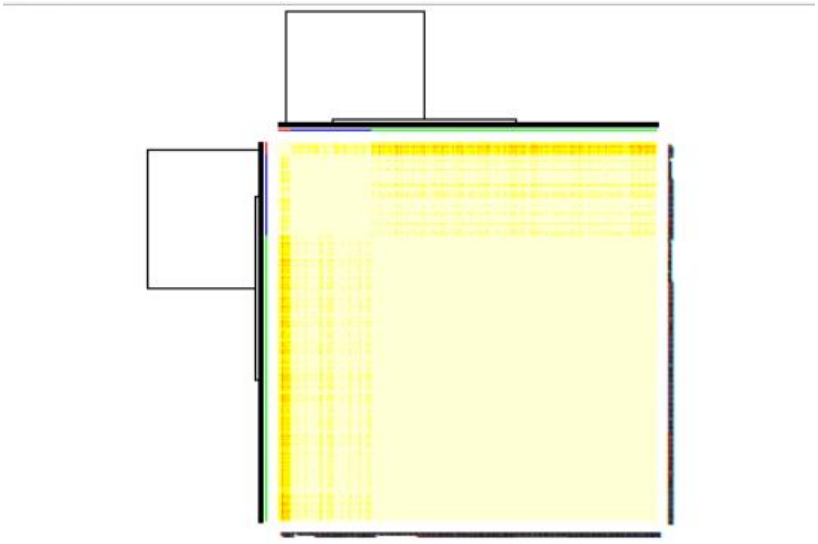
```
Clusters:
   Cluster 1, exemplar 122:
      1 2 3 5 7 11 12 13 18 25 26 28 29 30 31 33 34 35 36 43 46 54 57 71
      73 76 78 79 84 86 88 96 119 120 122 128 130 132 133 135 142 157 158
      162 163 168 169 182 183 187 198 199 202 203 208 211 214 219 231 234
      238 240 245 251 253 254 255 257 259 261 262 263 265 275 278 281 283
      301 303 318 322 324 329 331 336 338 344 366 367 371 373 374 390 393
      394 401 409 418 433 434 442 445 447 450 452 461 469 488 490 492 493
      499 500 517 518 534 536 564 565 566 567 568
```

```
  Cluster 2, exemplar 498:
     4 6 8 9 10 14 15 16 17 20 21 22 23 27 32 37 38 39 40 41 42 44 45 47
     48 49 50 51 52 53 55 56 58 59 60 61 62 63 64 65 66 67 68 69 70 72 74
     75 77 80 81 82 85 87 89 90 91 92 93 94 95 97 98 99 100 101 102 103
     104 105 106 107 108 110 111 112 113 114 115 116 117 118 121 124 125
     126 127 129 131 134 136 137 138 139 140 141 143 144 145 146 147 148
     149 150 151 152 153 154 155 156 159 160 161 164 166 167 170 171 172
     173 174 175 176 177 178 179 180 184 185 186 188 189 190 191 192 193
     194 195 196 197 200 201 204 205 206 207 209 210 212 215 216 217 218
     221 222 223 224 225 226 227 228 229 230 232 233 235 236 239 241 242
     243 244 246 247 248 249 250 252 256 258 260 264 267 268 269 270 271
     272 274 276 277 279 280 282 284 285 286 287 288 289 290 291 292 293
     294 295 296 297 298 299 300 302 304 305 306 307 308 309 310 311 312
     313 314 315 316 317 319 320 321 323 325 326 327 328 330 332 333 334
     335 337 339 341 342 343 345 346 347 348 349 350 351 352 354 355 356
     357 358 359 360 361 362 363 364 365 368 372 375 376 377 378 379 380
     381 382 383 384 385 386 387 388 389 391 392 395 396 397 398 399 400
     402 403 404 405 406 407 408 410 411 412 413 414 415 416 417 419 420
     421 422 423 424 425 426 427 428 429 430 431 432 435 436 437 438 439
     440 441 443 444 446 448 449 451 453 454 455 456 457 458 459 460 463
     464 465 466 467 468 470 471 472 473 474 475 476 477 478 479 480 481
     482 483 484 485 486 487 489 491 494 495 496 497 498 501 502 503 505
     506 507 508 509 510 511 512 513 514 515 516 519 520 521 523 524 525
     526 527 528 529 530 531 532 533 535 537 538 539 540 541 542 543 544
     545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561
     562 563 569
```

```
Cluster 3, exemplar 504:
   19 24 83 109 123 165 181 213 220 237 266 273 340 353 369 370 462 504
   522
```

- From the above result, there are three clusters in Breast cancer dataset. But originally there are only two classes. May be the remaining 19 instances in cluster 3 belongs to the above two clusters only.

Heatmap of Breast cancer data is.



- As 19 instances placed in wrong cluster,
- Error rate= 19/569 * 100 = 3.3%

3) **Spectral Clustering:**

(A) Spectral Clustering Iris Data:

After loading the Iris dataset,

Retrieving the column names:

```
> names(iris);
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
>
```

- Since “Species” is the class column: Setosa or Versicolour or Virginica we do not need it for clustering. It can be used later to analyze the clustering results.
- Splitting the attributes into variables. The first 4 that are used for clustering as one variable and the last column into another.

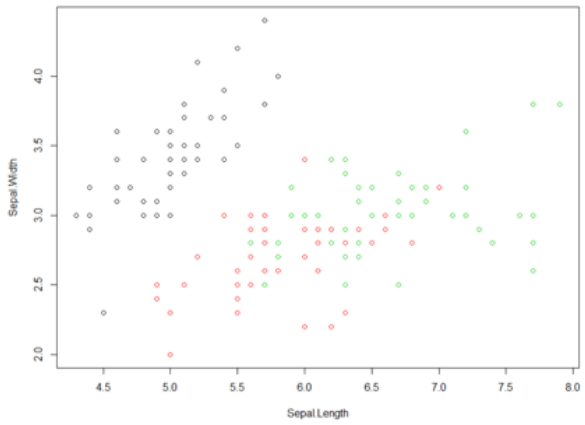
```
x = iris[,-5];
y = iris$Species;
```

Clustering using the specClust function:

```
Speciris <- specClust(iris[1:4],3,  nn=5)
Speciris
```

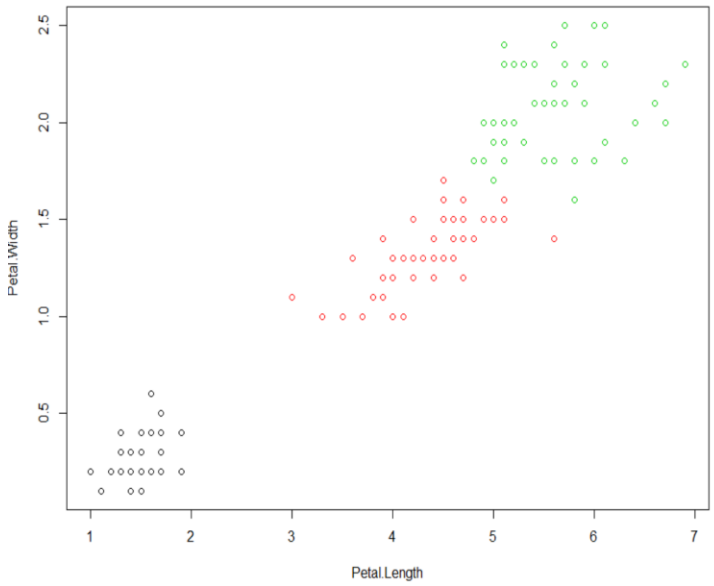
```
Cluster means:
       [,1]       [,2]          [,3]
1 0.5988062 -0.4671027 -6.182545e-01
2 0.6150727  0.7884704  1.791307e-05
3 0.5533649 -0.4316863  6.824605e-01

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [42] 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1
 [83] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 1
[124] 3 3 3 3 3 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3

Within cluster sum of squares by cluster:
[1] 2.132398e+00 3.000920e-08 2.000730e+00
 (between_SS / total_SS =  95.8 %)

Available components:

 [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
 [6] "betweenss"    "size"         "iter"         "ifault"       "eigenvalue"
[11] "eigenvector"  "data"         "indAll"       "indUnique"    "L"
[16] "archetype"    "call"
>
```

Plotting Sepal width vs Sepal Length:

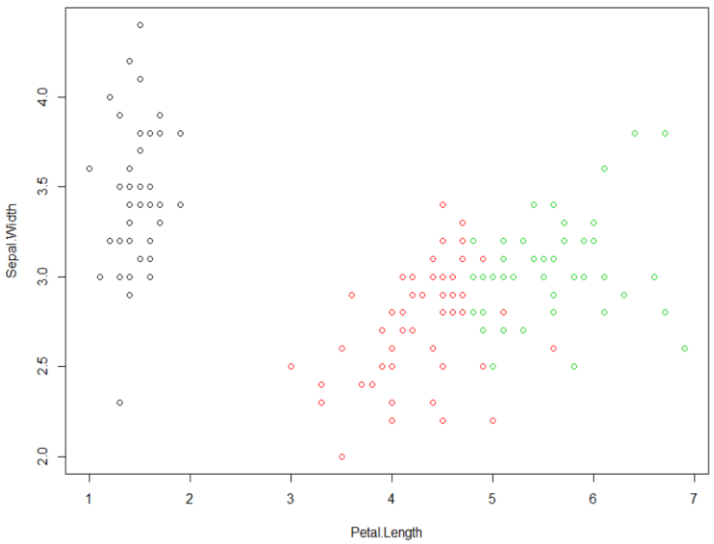


Name: Sai Koumudi Student ID: 800959723

Plotting Petal Width vs Petal Length:

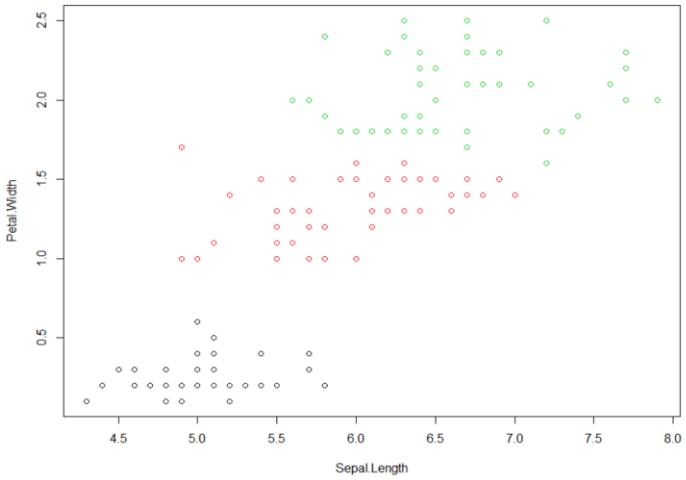


Plotting Sepal width vs Petal length:

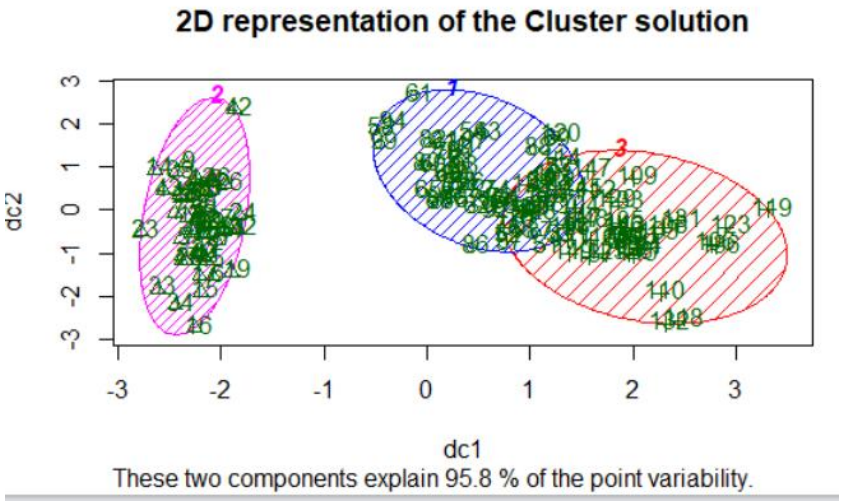


Name: Sai Koumudi Student ID: 800959723

Plotting Petal width vs Sepal Length:



Cluster plot:



Results:

- As we already know the class to which each record belongs to, now we compare the clustering results to the original classes. This gives a matrix.
- Now, we can compare the clustering results to that of the original classes. This gives us a matrix how the clustering has done.

```
table(y,Speciris$cluster)
```

```
            1  2  3
setosa     50  0  0
versicolor  0 48  2
virginica   0  4 46
```

- From the above table, we can observe that, two records of versicolor fall into the cluster of virginica and 14 records of virginica fall into the cluster of versicolor.
- From above table, mapping class values to that particular levels. Mapping Cluster 2 to setosa, 3 to versicolor and 1 to virginica.

```
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          4        46

Overall Statistics

               Accuracy : 0.96
                 95% CI : (0.915, 0.9852)
    No Information Rate : 0.3467
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.94
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.9231           0.9583
Specificity                 1.0000            0.9796           0.9608
Pos Pred Value              1.0000            0.9600           0.9200
Neg Pred Value              1.0000            0.9600           0.9800
Prevalence                  0.3333            0.3467           0.3200
Detection Rate              0.3333            0.3200           0.3067
Detection Prevalence        0.3333            0.3333           0.3333
Balanced Accuracy           1.0000            0.9513           0.9596
```

- The accuracy of the spectral algorithm has turned out to be 0.96 = 96% accuracy.

- 2 instances of versicolor are wrongly clustered.
- 4 instances of virginica are wrongly clustered.
- Compared to K-means and Ap clustering, spectral clustering has higher accuracy.

(B) Spectral - Breast Cancer Dataset:

After loading the Breast cancer dataset,

Retrieving the column names:

```
> library(kknn)
> library(cluster)
> bcw <- read.csv('D:/hw/dataset3/bcw.csv')
> names(bcw)
 [1] "v2"  "v3"  "v4"  "v5"  "v6"  "v7"  "v8"  "v9"  "v10" "v11" "v12" "v13" "v14"
[14] "v15" "v16" "v17" "v18" "v19" "v20" "v21" "v22" "v23" "v24" "v25" "v26" "v27"
[27] "v28" "v29" "v30" "v31" "v32"
```

- The column V2 is the class label. It has 2 classes: M or B
- Splitting the attributes into variables. The first column as one variable and the rest as another variable.

```
x = bcw[, -1]
y = bcw$v2
```

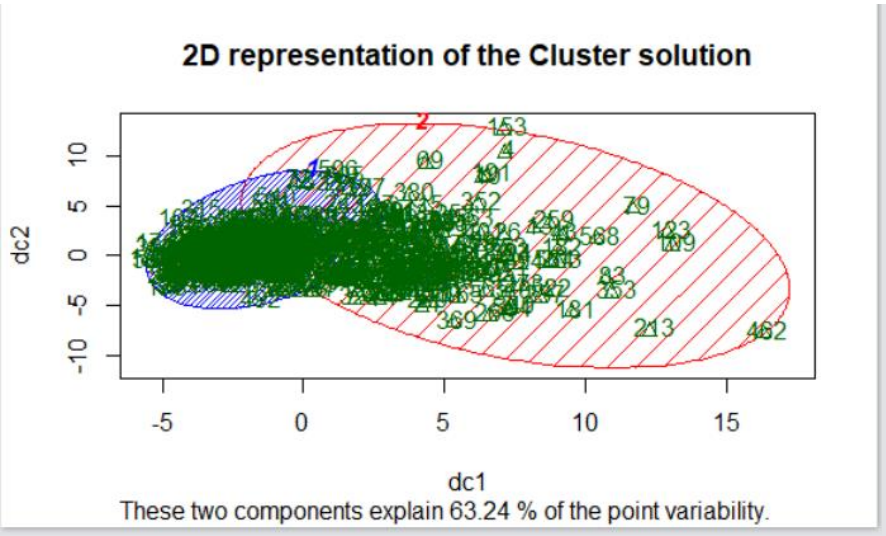
- Spectral clustering using specClust function in R

```
Specbcw <- specClust(x,2,nn=5)
Specbcw
```

- Results of the spectral clustering with 2 clusters

```
Cluster means:
       [,1]       [,2]
1 0.8206427 -0.5232670
2 0.6479200  0.7197535

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 1 2 2 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1
 [42] 1 2 2 1 2 1 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 2 1 2 2 1 1 2 1 2 2 2 1 1 2 1 2 2 1 1 2
 [83] 2 2 1 2 1 2 1 2 1 1 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 1 2 2 2 2 1 2 2
[124] 1 1 1 1 2 2 2 1 2 2 1 2 1 1 1 2 1 1 2 1 1 1 1 2 1 1 1 1 2 2 1 1 1 2 1 1 1 1 2 2 1
[165] 2 1 1 2 2 1 1 1 2 1 1 1 2 2 1 1 2 2 2 1 1 1 2 1 1 1 2 1 1 2 2 1 2 2 2 2 1 2 2 2 1
[206] 1 1 2 2 1 2 1 2 2 2 2 1 1 2 2 1 1 1 2 1 1 1 1 1 2 2 1 1 2 1 1 2 2 1 2 1 1 2 1 2 1
[247] 1 2 1 1 2 1 2 2 2 1 2 2 2 2 2 1 2 1 2 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 2 2 1 1 1
[288] 1 2 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 2 1 2 1 1 1 1 1
[329] 2 2 2 1 1 1 1 2 1 2 1 2 1 1 1 2 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 2
[370] 2 2 1 2 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1 2 2 1 1 2 2 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1
[411] 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 2 1 2 2 1 2 1 1 1 1 1 2 1 1 2 1 2 1 1 2 1
[452] 2 1 1 1 1 1 1 1 1 2 2 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 2 1 1 1 1
[493] 2 1 1 1 1 1 2 2 1 2 1 2 1 1 1 1 1 2 1 1 2 1 1 1 2 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1
[534] 2 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 1

Within cluster sum of squares by cluster:
[1] 19.14359 12.80384
 (between_SS / total_SS =  86.6 %)

Available components:

 [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
 [6] "betweenss"    "size"         "iter"         "ifault"       "eigenvalue"
[11] "eigenvector"  "data"         "indAll"       "indUnique"    "L"
[16] "archetype"    "call"
>
```

Cluster Plot:



Results:

- As we already know the class to which each record belongs to, now we compare the clustering results to the original classes. This gives a matrix.
- Now, we can compare the clustering results to that of the original classes. This gives us a matrix as to how the clustering has fared.

```
table(y, Specbcw$cluster)
```

```
> table(y, Specbcw$cluster)

y     1   2
  B 334  23
  M  29 183
>
```

Now, map cluster 1 to “B” and cluster 2 to “M”. Generating confusion matrix and accuracy results.

y is the original class label and bcw$x is mapped class labels obtained from k-means clustering.

```
confusionMatrix(y,bcw$x)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   B   M
         B 334  23
         M  29 183

               Accuracy : 0.9086
                 95% CI : (0.8819, 0.931)
    No Information Rate : 0.638
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8034
 Mcnemar's Test P-Value : 0.4881

            Sensitivity : 0.9201
            Specificity : 0.8883
         Pos Pred Value : 0.9356
         Neg Pred Value : 0.8632
             Prevalence : 0.6380
         Detection Rate : 0.5870
   Detection Prevalence : 0.6274
      Balanced Accuracy : 0.9042

       'Positive' Class : B
```

- Therefore, the accuracy of k-means clustering algorithm on breast cancer data = 90.86%
- 23 instances of B wrongly clustered into class M
- 29 instances of M wrongly clustered into B class.

Performance Analysis: The accuracy of the Spectral clustering is more when compared to the accuracies of the K-means clustering and Ap clustering.