

Class 7: Machine Learning 1

Tin Nguyen

Table of contents

Clustering	1
K-means	3
Hierarchical clustering	6
Principal Component Analysis (PCA)	8
Data import	8
PCA to the rescue	10
interpreting PCA Results.	10

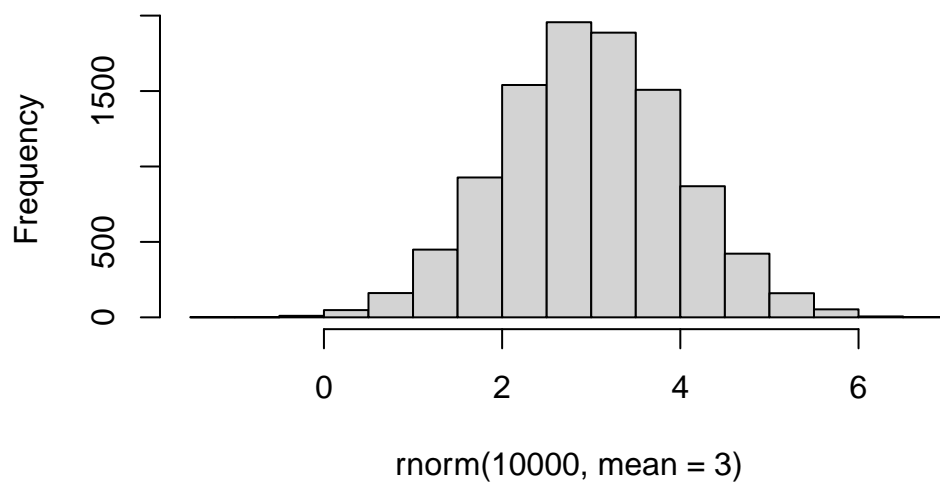
Today we will explore unsupervised machine learning methods starting with clustering and dimensionality reduction.

Clustering

To start let's make up some data to cluster where we know what the answer should be. The 'rnorm()' function will help us here.

```
hist( rnorm(10000, mean = 3) )
```

Histogram of rnorm(10000, mean = 3)

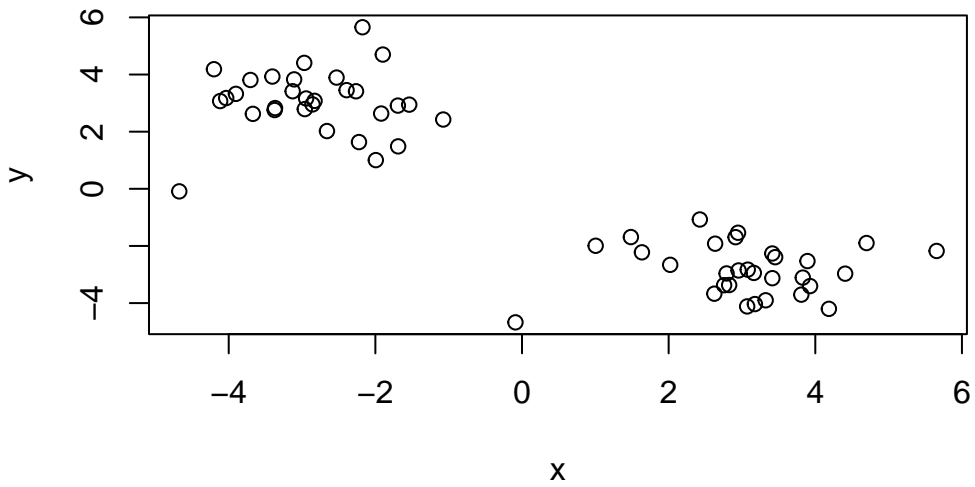


Return 30 numbers centered on -3

```
tmp <- c( rnorm(30, mean = -3), rnorm(30, mean = +3) )  
x <- cbind(x = tmp, y = rev(tmp))
```

Make a plot of 'x'

```
plot(x)
```



K-means

The main function in “base” R for K-means clustering is called ‘kmeans()’:

```
km <- kmeans(x, centers = 2)
km
```

K-means clustering with 2 clusters of sizes 30, 30

Cluster means:

	x	y
1	-2.843567	3.047474
2	3.047474	-2.843567

Clustering vector:

[illegible]

Within cluster sum of squares by cluster:

```
[1] 59.67693 59.67693
(between_SS / total_SS = 89.7 %)
```

Available components:

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

The `'kmeans()'` function return a “list” with 9 components. You can see the named components of any list with the `'attributes()'` function.

```
attributes(km)
```

```
$names
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"

$class
[1] "kmeans"
```

Q. How many points are in each cluster ?

km\$size

[1] 30 30

Q. Cluster assignment/membership vector?

```
km$cluster
```

[illegible]

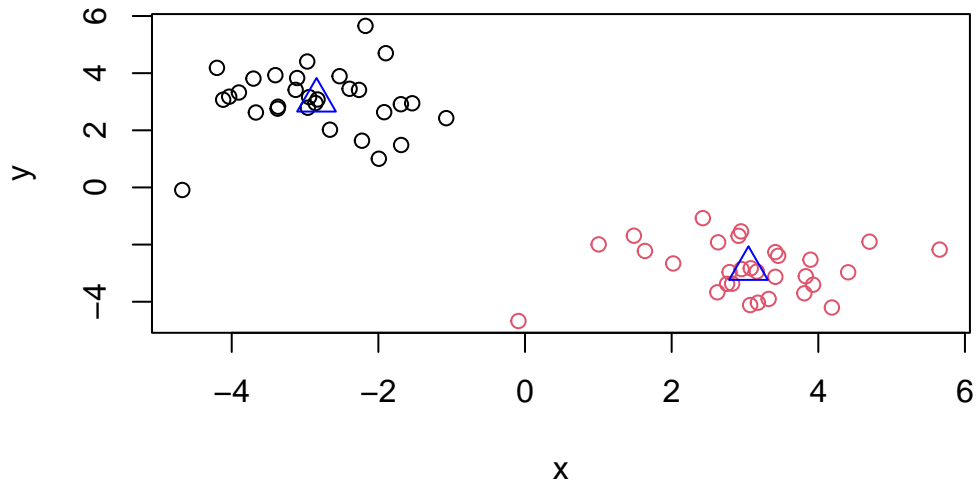
Q. Cluster centers?

km\$centers

	x	y
1	-2.843567	3.047474
2	3.047474	-2.843567

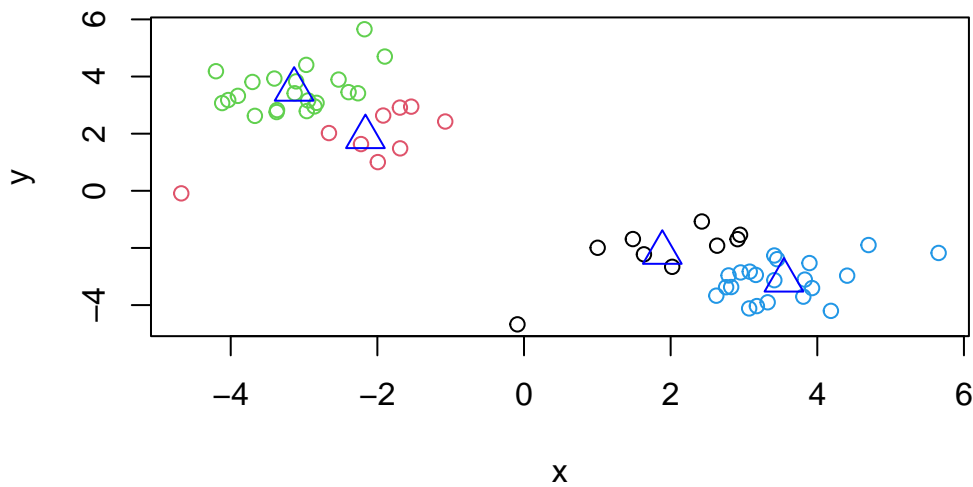
Q. Make a plot of our 'kmeans()' results showing cluster assignment using different colors for each cluster/group of points and cluster centers in blue.

```
plot(x, col = km$cluster )
points(km$centers, col = "blue", pch = 2, cex = 2)
```



Q, Run 'kmmeans()' again on 'x' and this cluster into 4 groups/ clusters and plot the same results figure as above.

```
km2 <- kmeans(x, centers = 4)
plot(x, col = km2$cluster )
points(km2$centers, col = "blue", pch = 2, cex = 2)
```



key-point: K-means clustering is super popular but can be miss-used. One big limitation is that it can impose a clustering pattern on your data even if clear natural grouping don't exist - i.e. it does what you tell it to do in terms of 'centers'.

Hierarchical clustering

The main function in “base” R for herarchical clustering is called ‘hclust()’.

You can't just pass our dataset as is into ‘hclust()’ you must give “distacne matrix” as input. WE can get this from ‘dist()’ function in R.

```
d <- dist(x)
hc <- hclust(d)
hc
```

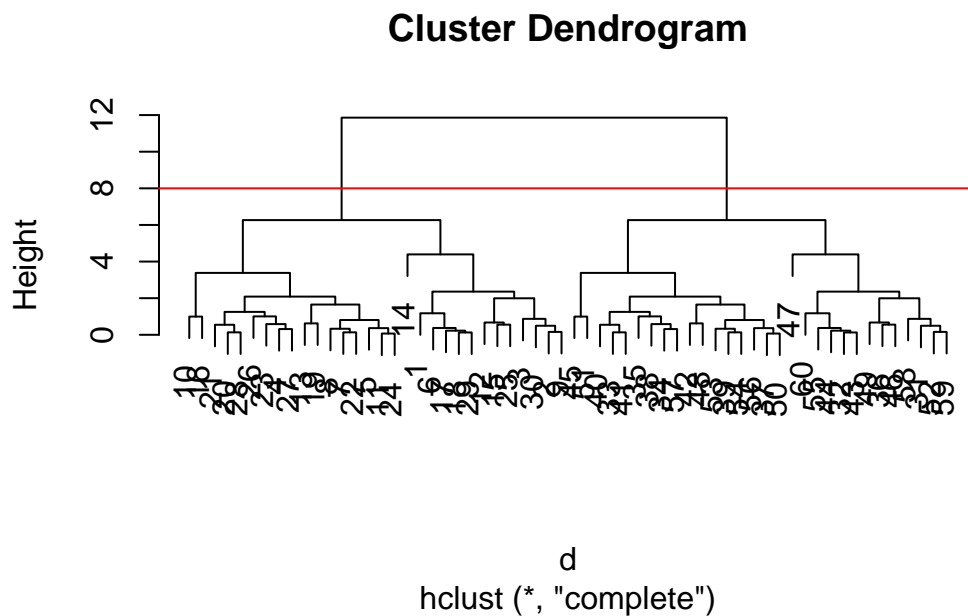
Call:

```
hclust(d = d)
```

```
Cluster method   : complete
Distance         : euclidean
Number of objects: 60
```

The results of 'hclust()' don't have a useful 'print()' method but do have a special 'plot()' method.

```
plot(hc)
abline(h=8, col="red")
```



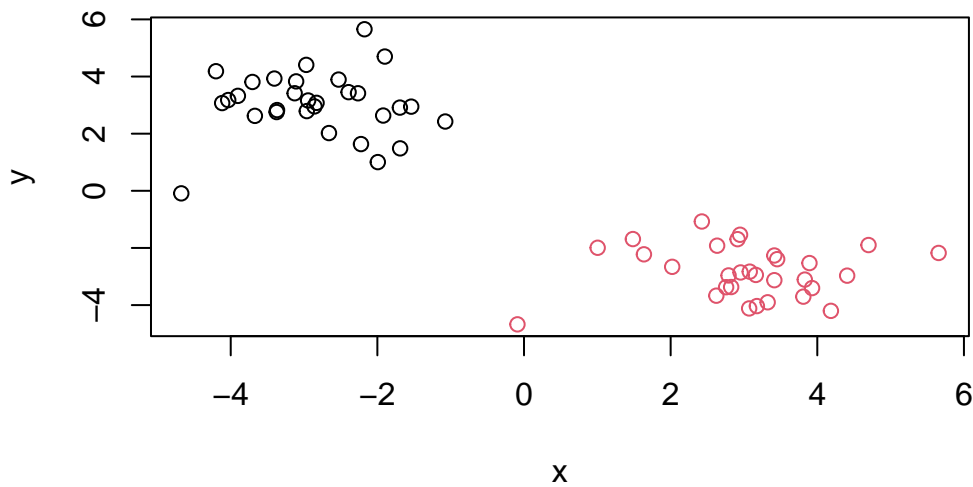
To get our main cluster assignment (membership vector) we need to “cut” the tree at the big goalposts...

```
grps <- cutree(hc, h = 8)
```

```
table(grps)
```

```
grps
 1  2
30 30
```

```
plot(x, col = grps)
```



Hierarchical Clustering is distinct in that the dendrogram (tree figure) can reveal the potential grouping in your data (unlike K-means)

Principal Component Analysis (PCA)

PCA is a common and highly useful dimensionality reduction technique used in many fields - particularly bioinformatics.

Here we will analyze some data from the UK on food consumption.

Data import

```
url <- "https://tinyurl.com/UK-foods"
x <- read.csv(url)

head(x)
```

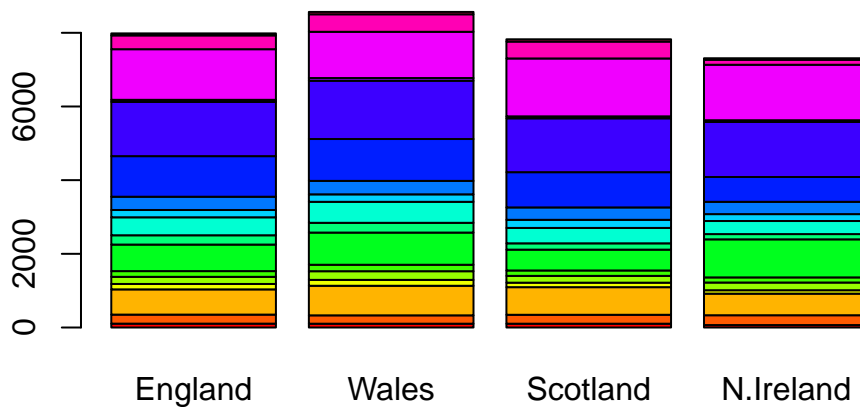
	X	England	Wales	Scotland	N.Ireland
1	Cheese	105	103	103	66
2	Carcass_meat	245	227	242	267
3	Other_meat	685	803	750	586

4	Fish	147	160	122	93
5	Fats_and_oils	193	235	184	209
6	Sugars	156	175	147	139

```
x <- read.csv(url, row.names = 1)
head(x)
```

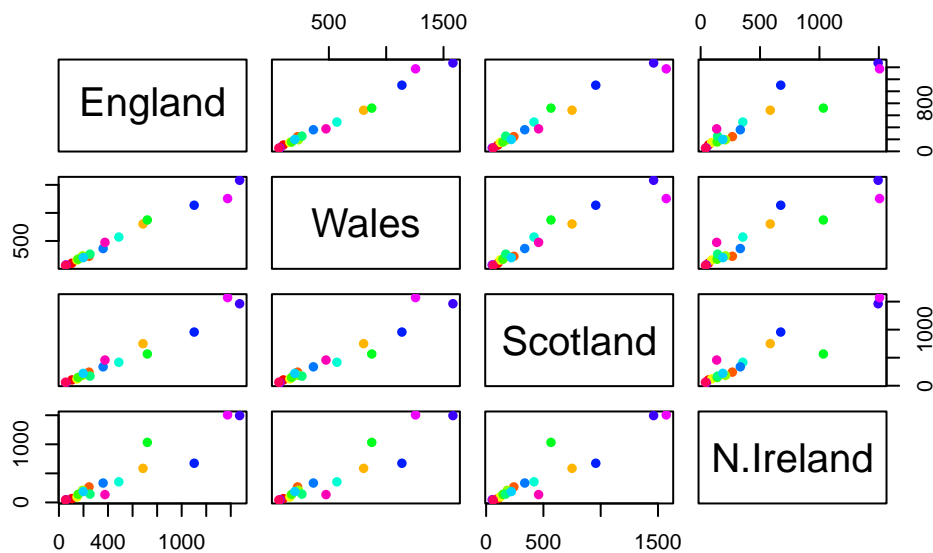
	England	Wales	Scotland	N.Ireland
Cheese	105	103	103	66
Carcass_meat	245	227	242	267
Other_meat	685	803	750	586
Fish	147	160	122	93
Fats_and_oils	193	235	184	209
Sugars	156	175	147	139

```
barplot(as.matrix(x), beside=F, col=rainbow(nrow(x)))
```



One conventional plot that can be useful is called a “paris” plot.

```
pairs(x, col=rainbow(nrow(x)), pch=16)
```



PCA to the rescue

The main function in base R for PCA 'prcomp()'.

```
pca <- prcomp( t(x) )
summary(pca)
```

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	324.1502	212.7478	73.87622	2.921e-14
Proportion of Variance	0.6744	0.2905	0.03503	0.000e+00
Cumulative Proportion	0.6744	0.9650	1.00000	1.000e+00

interpreting PCA Results.

The 'prcomp()' function returns a list object of our results with five attributes/components

```
attributes(pca)
```

```
$names
[1] "sdev"      "rotation" "center"    "scale"     "x"

$class
[1] "prcomp"
```

The two main “results” in here are ‘*pca\$x*’ and ‘*pca\$rotation*’. The first of these (‘*pca\$x*’) contains the scores of the data on the new PC axis - we use these to make our “PCA plot”.

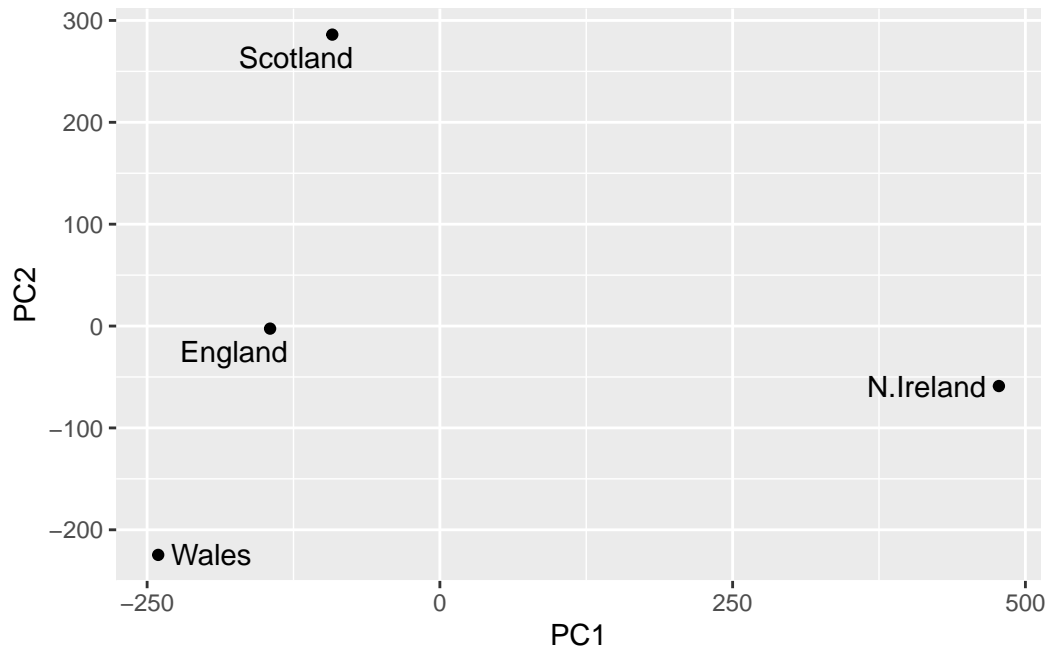
```
pca$x
```

	PC1	PC2	PC3	PC4
England	-144.99315	-2.532999	105.768945	-9.152022e-15
Wales	-240.52915	-224.646925	-56.475555	5.560040e-13
Scotland	-91.86934	286.081786	-44.415495	-6.638419e-13
N.Ireland	477.39164	-58.901862	-4.877895	1.329771e-13

```
library(ggplot2)
library(ggrepel)

# Make a plot of pca$x with PC1 vs PC2

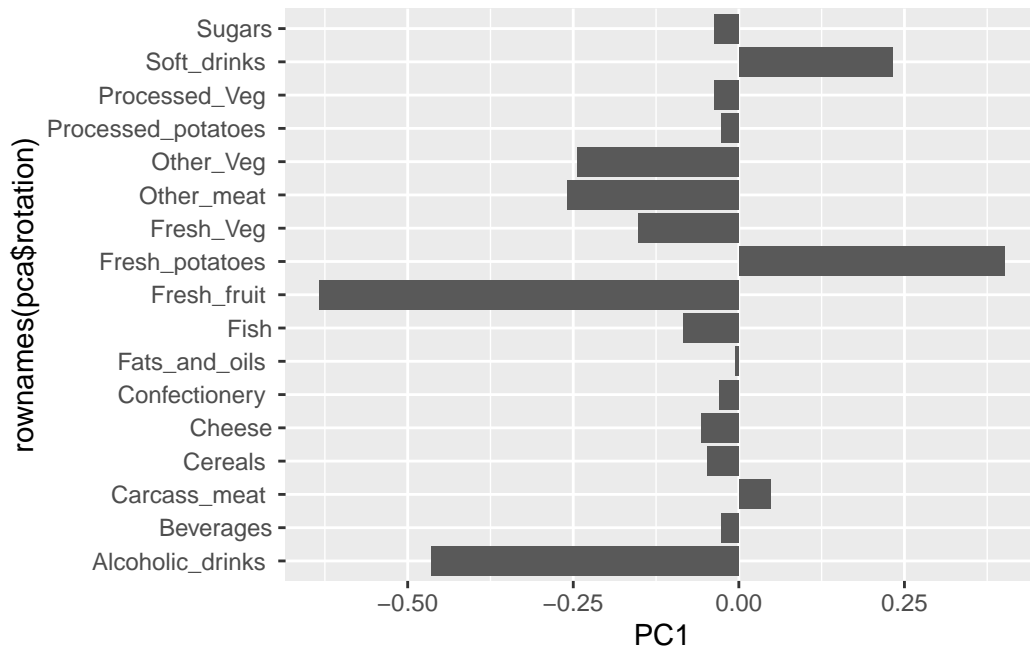
ggplot(pca$x) +
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point() +
  #geom_label()+
  geom_text_repel()
```



PCA picks find the principle directions that describe the most different between the countries. It found that in the most principle direction (PC1) Ireland is the most different from other country.

The second major result is contained in the 'pca\$rotation' object or component. :et's plot this to see what PCA is picking up...

```
ggplot(pca$rotation) +  
  aes(PC1, rownames(pca$rotation)) +  
  geom_col()
```



In the most principle direction PC1, this graph of rotation shows factors that contribute to the different in the countries. This shows that Soft drinks and Freshfruit consumption is significantly higher in Irelands compares to other countries. And other countries consume more frsh fruit and alcoholic dirnks than Ireland.