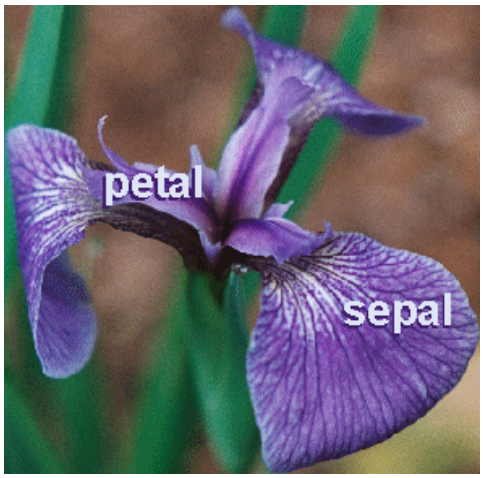
BL5229: Data Analysis with Matlab
Lab: Learning: Clustering

The following hands-on exercises were designed to teach you step by step how to perform and understand various clustering algorithm. We will look at two standard data sets: the Fisher data set on irises, and a gene expression dataset related to lung cancers.

Exercise 1: clustering the Fisher iris data set.

In the 1920's, botanists collected measurements on the sepal length and width, and on the petal length and width of 150 irises, 50 from each of three species (*setosa, versicolor, virginica*) The measurements became known as Fisher's iris data



The question is: is it possible to define the specie of an iris based on these four measurements? We attempt to analyse this question by clustering the Fisher's iris dataset.

1) Statistics of the dataset

Load the Fisher iris data set:

```
>> load fisheriris
```

Note that this will create two arrays:

- **species**, that gives the species of each iris considered (of size 150):

```
>> species(1:5)

ans =

    'setosa'
    'setosa'
```

```
    'setosa'
    'setosa'
    'setosa'

>> species(145:150)

ans =

    'virginica'
    'virginica'
    'virginica'
    'virginica'
    'virginica'
    'virginica'
```

- **meas**, matrix of size 150x4, that gives the four measurements for each iris (in the order sepal length, sepal width, petal length, petal width).

```
>> meas(1:4,1:4)
```

	Sepal length	Sepal width	Petal length	Petal width
iris1 ->	**5.1000**	**3.5000**	**1.4000**	**0.2000**
iris2 ->	**4.9000**	**3.0000**	**1.4000**	**0.2000**
iris3 ->	**4.7000**	**3.2000**	**1.3000**	**0.2000**
iris4 ->	**4.6000**	**3.1000**	**1.5000**	**0.2000**

You will:
a) create three data arrays, **`setosa, versicolor`**, and **`virginica`**, that are specific to each iris species

b) Compute the statistics (mean, standard deviation) for each measurement, for each iris species. Which two measures have the highest correlation coefficients (for each specie)? Fill up the table:

	Setosa			**Versicolor**			**Virginica**		
Measure	Mean	Std deviation	Best correlation	Mean	Std deviation	Best correlation	Mean	Std deviation	Best correlation
Sepal length									
Sepal width									
Petal length									
Petal width									

c) Generate on a single page 6 subplots, each showing the values of two characteristics of the irises, for all three types of irises. For example, the first subplot will have “Sepal length” as X axis and “Sepal Width” as Y axis, with Setosa in red, Versicolor in green, and Virginica in blue.

2) Clustering

We go back to the full measurement array, **meas**. We will cluster the 150 irises into 3 clusters, and compare the results with the actual species of these 150 irises.

There are (at least) two cluster methods implemented in Matlab:

- ***hierarchical clustering***: use the function **`clusterdata`**. The basic syntax for this function is:

```
idx = clusterdata(meas,'Distance',dist_method,'Linkage',
link_method,'maxclust',Nclust)
```

where:

- **`'Distance'`** specifies the method used to measure the similarity of two measurements. Standard values for **`'Distance'`** are **`'euclidean','correlation', 'cosine', …`**
- **`'Linkage'`** specifies the linkage method used to compute the distance between two clusters. Standard methods include **`'single','average'`**, and **`'complete'`**.
- **`'maxclust'`** is the number of clusters expected.
- **`idx`** is the array of cluster assignment for each data point (in our case, irises), after clustering.

For example:

```
Idx = clusterdata(meas,'Distance','euclidean','Linkage',
'single','maxclust',4)
```

will cluster the data using the Euclidian distance to compare data points, single linkage to compare clusters, and output 4 clusters.

- ***kmeans clustering***: use the function **`kmeans`**. The basic syntax for the function is:

```
idx = kmeans(meas,K,'Distance',dist_method)
```

where:

- K is the number of clusters.
- 'Distance' specifies the method used to measure the similarity of two measurements. Standard values for 'Distance' are 'sqEuclidean','correlation', 'cosine', …
- idx is the array of cluster assignment for each data point (in our case, irises), after clustering.

For example,

```
Idx = kmeans(meas,4,'Distance','correlation')
```

will cluster the data using k-means, with k set to 4 clusters, and using correlation to compare data points.
Once the irises have been clustered, we need to compare the cluster assignments of the irises with their actual species. Useful tools for this include:

- crosstab:

`crosstab(idx, species)` will build a matrix C of size Ncluster x 3, with C(i,j) gives the number of elements of cluster I that belongs to specie j.

-randindex:

`R=randindex(idx,species)` gives an Index of quality for the two partitioning of the data, idx (from the clustering technique), and species. Note that randindex is not a Matlab built in function. You can get it from here.

Experiment with the different clustering techniques. Fill in this table:

Cluster technique			
Technique	Distance	Linkage	RandIndex
Hierarchical	Euclidean	Single	
Hierarchical	Euclidian	Average	
Hierarchical	Euclidian	Complete	
Hierarchical	Correlation	Single	
Hierarchical	Correlation	Average	
Hierarchical	Correlation	Complete	
K-means	sqEuclidean		
K-means	Correlation		
K-means	Cosine		

Which method works “best”?

For this method, plot the silhouette for each flower, organized in clusters.

Exercise 2) Clustering a gene expression dataset related to lung cancers.

The pathological distinction between malignant pleural mesothelioma (MPM) and adenocarcinoma (AD) of the lung can be cumbersome. In 2002, Gordon et al studied the expression of a large number of genes for 181 patients suffering from either MPM or AD to see if the gene expression pattern of a patient was enough to identify which disease she suffers from.

I have prepared a Matlab mat file that contains these data, gordon_lung.mat. To load these data, just type:

```
load gordon_lung
```

This will generate two arrays:

- **`gene`**: a matrix of size 181x1626 that contains the expression values of 1626 genes for all 181 patients.

- **`type`**: an array of size 181 that gives the name of the disease for each of the 181 patients.

Repeat the clustering analysis you have performed on the iris dataset with these data. Fill in the table:

Cluster technique			
Technique	Distance	Linkage	RandIndex
Hierarchical	Euclidean	Single	
Hierarchical	Euclidian	Average	
Hierarchical	Euclidian	Complete	
Hierarchical	Correlation	Single	
Hierarchical	Correlation	Average	
Hierarchical	Correlation	Complete	
K-means	sqEuclidean		
K-means	Correlation		
K-means	Cosine		

Which method works “best”?