Practical 6: Data Stream Mining

In this practical, we use the *stream* R package for analysing stream data. Please install the stream package to complete the practical.

I. Creating a data stream

1. We firstly create a generator to generate stream data points that will belong to one of three clusters (k=3). Each data point will have 2 dimensions (d=2). The data points will follow Gaussian distribution with 5% noise. When a new data point is requested from this data generator, a cluster will be chosen randomly using the probability weights in p.

```
library("stream")
stream <- DSD_Gaussians(k = 3, d = 2, noise = .05, p = c(.5, .3, .1))
stream
```

2. Generate 5 data points using the generator.

```
p <- get_points(stream, n = 5)
p
```

3. Use option class=TRUE to see which cluster a data point belongs to. Please note that noise data points (5%) do have the class labels (NA).

```
p <- get_points(stream, n = 10, class = TRUE)
p
```

4. Plot the 500 points from the data stream

```
plot(stream, n=500)
```

II. Reading and writing data streams

1. Write the created stream with 100 data points to a file called data.csv.

```
write_stream(stream, "data.csv", n = 100, sep = ",")
```

2. Read back the data.csv file to R.

```
stream_data = DSD_ReadStream("data.csv")
```

3. Note that the data has not been read to the stream_data until we use get_points

```
get_points(stream_data, n=5)
```

III. Reservoir Sampling

1. Create a stream with 3 clusters and 5% noise

```
stream <- DSD_Gaussians(k = 3, d = 2, noise = .05, p = c(.5, .3, .1))
```

2. Create a Reservoir sampling mechanism with 20 points will be sampled from the stream

```
sample <- DSAggregate_Sample(k = 20)
```

3. Update the data for sample using 500 data points from stream

```
update(sample, stream, 500)
sample
```

4. Get the data from sample

```
get_points(sample)
```

5. Plot the data points in sample

```
plot(get_points(sample))
```

IV. Data Stream Clustering

1. We firstly prepare the clustering algorithm. We use DSC_DStream which implements the D-Stream algorithm (Tu and Chen 2009). D-Stream assigns points to cells in a grid. For the example we use a gridsize of 0.1.

```
dstream <- DSC_DStream(gridsize = .1, Cm = 1.2)
dstream
```

2. The clusters are currently empty, but they are ready to get data points from the stream.

```
update(dstream, stream, n = 500)
dstream
plot(dstream, stream)
```

3. There are a number of micro-clusters. We can get the centers of the micro-clusters using:

```
head(get_centers(dstream))
```