Introduction to Business Analytics

Lecture 13: Intro to Machine Learning in R (2)

Iegor Vyshnevskyi

Woosong University

May 22/23, 2023

Agenda

1. Unsupervised Machine Learning
2. Unsupervised Machine Learning in Practice
3. In-class Assignment
4. Final Exam instructions

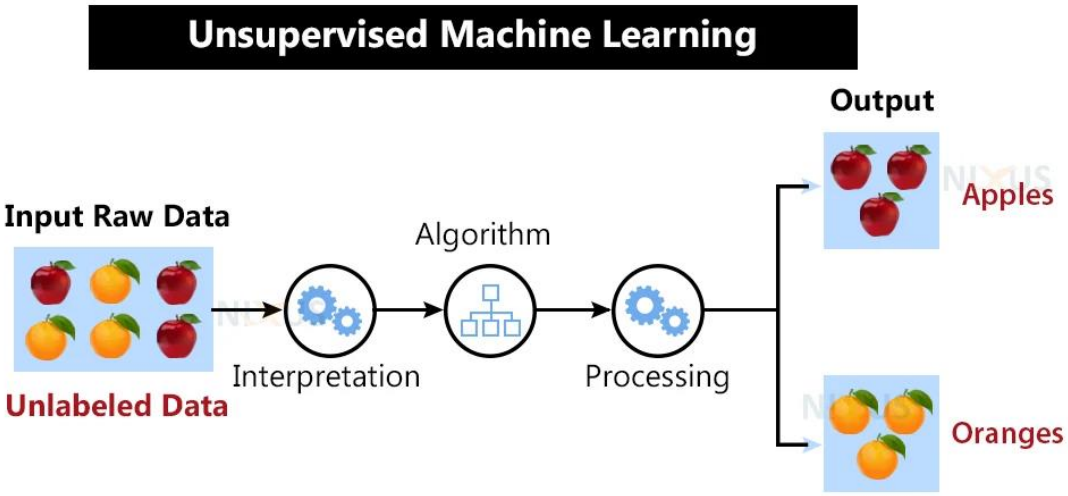
Acknowledgment: Used a number of open sources and materials from the web.

1. Unsupervised Machine Learning

*What is **Unsupervised ML?***

Unsupervised learning, also known as *unsupervised machine learning*, uses machine learning algorithms to analyze and cluster unlabeled datasets.

These algorithms discover hidden patterns or data groupings without the need for human intervention.



Why Unsupervised ML?

Here, are prime reasons for using Unsupervised Learning in Machine Learning:

- Unsupervised machine learning finds all kind of unknown patterns in data and it helps identify anomalies and outliers.
- Unsupervised machine learning helps identify the data structure and reduce data daimonions for better visualization.
- Unsupervised methods help you to find features which can be useful for categorization.
- It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.
- It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention.

*What are **Unsupervised ML techniques?** (cont.)*

Unsupervised ML generally differentiates between

- **Clustering**, where the goal is to find homogeneous subgroups within the data; the grouping is based on distance between observations.

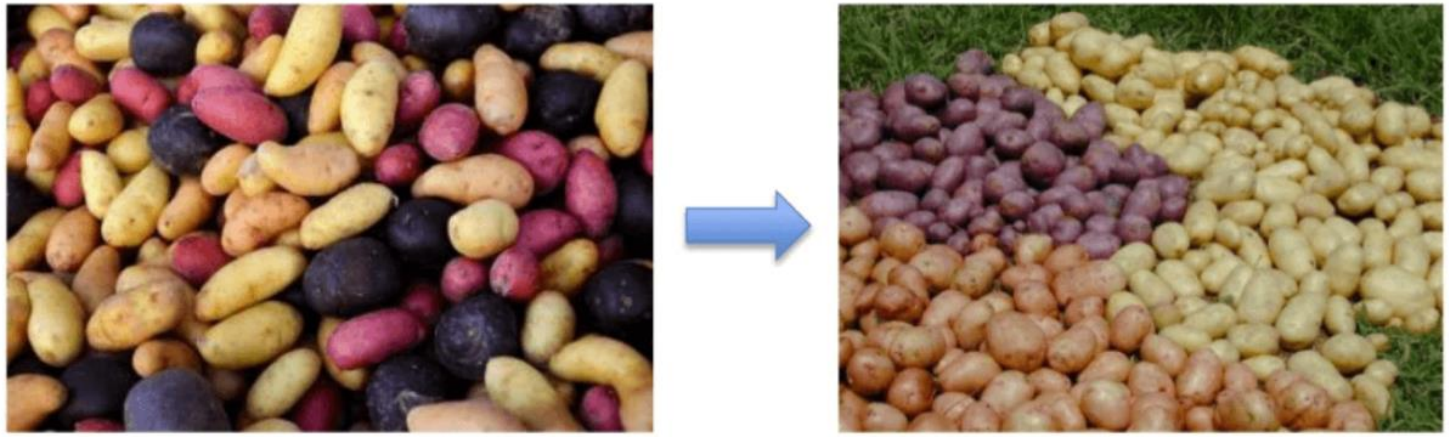
- **Dimensionality reduction**, where the goal is to identify patterns in the features of the data. Dimensionality reduction is often used to facilitate visualisation of the data, as well as a pre-processing method before supervised learning (*can be under both supervised and unsupervised ML*).

Unsupervised ML presents specific challenges and benefits:

- there is no single goal in UML
- there is generally much more unlabelled data available than labelled data.

Clustering

- Clustering is an important concept when it comes to unsupervised learning.
- It mainly deals with finding a structure or pattern in a collection of uncategorized data.
- Unsupervised Learning Clustering algorithms will process your data and find natural clusters(groups) if they exist in the data.
- You can also modify how many clusters your algorithms should identify.
- It allows you to adjust the granularity of these groups.



sample

Cluster/group

Clustering Types

Following are the clustering types of Unsupervised ML:

- K-mean Clustering: The k-means clustering algorithms aims at partitioning n observations into a fixed number of k clusters. The algorithm will find homogeneous clusters.

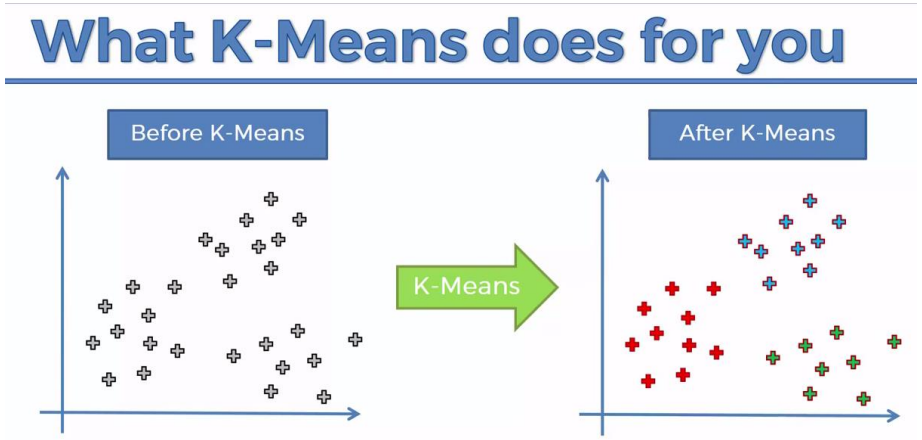
- Hierarchical Clustering: Hierarchical clustering is an algorithm which builds a hierarchy of clusters. It begins with all the data which is assigned to a cluster of their own. Here, two close cluster are going to be in the same cluster. This algorithm ends when there is only one cluster left.

2. Unsupervised Machine Learning in Practice

What we do today

K-means clustering

1. It starts with K as the input which is how many clusters you want to find. Place K centroids in random locations in your space.
2. Now, using the euclidean distance between data points and centroids, assign each data point to the cluster which is close to it.
3. Recalculate the cluster centers as a mean of data points assigned to it.
4. Repeat 2 and 3 until no further changes occur.



Example from: https://towardsdatascience.com/clustering-unsupervised-learning-788b215b074b

Example Dataset

Edgar Anderson's Iris Data

- From the *iris* manual page:
  - This famous (Fisher's or Anderson's) *iris* data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.
  - K-means Clustering is used with *unlabeled data*, but in this case, we have a labeled dataset so we have to use the iris data without the Species column. In this way, algorithm will cluster the data and we will be able to compare the predicted results with the original results, getting the accuracy of the model.

Loading packages and the data

Install and load needed packages first.

```
# Install libraries
install.packages('tidyverse')
install.packages('DT')
install.packages('ggvis')
install.packages('gridExtra')
install.packages("class")
install.packages('gmodels')

# Load libraries
library(tidyverse)
library(DT)
library(ggvis)
library(gridExtra)
library(class)
library(gmodels)
```

Load data.

```
# Load data
data(iris)

iris
```

kmeans is installed in the **base** package from R, so we don't have to install any package.

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
```

Initial Overview of the Data Set

```
# Dataset information
datatable(iris)

glimpse(iris)

summary(iris)
```

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Show 10 entries

Search:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0
2	4.9	3	1.4	0
3	4.7	3.2	1.3	0
4	4.6	3.1	1.5	0
5	5	3.6	1.4	0
6	5.4	3.9	1.7	0
7	4.6	3.4	1.4	0

Showing 1 to 10 of 150 entries

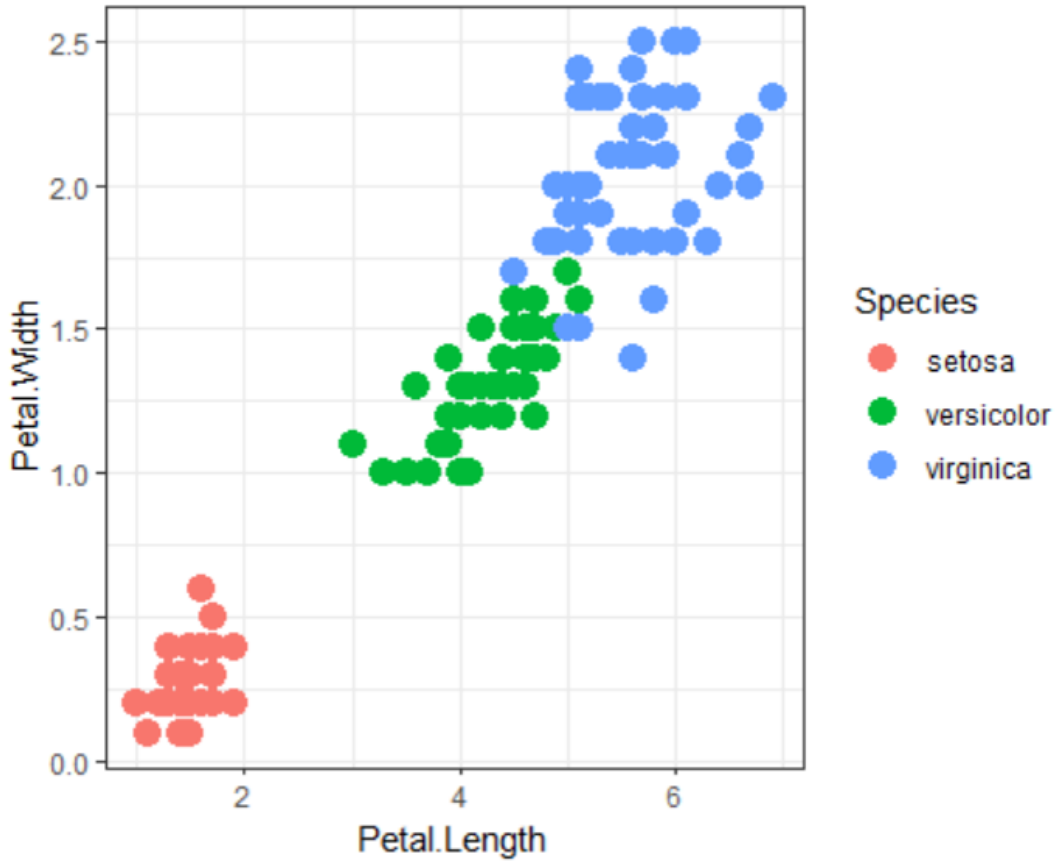
Previous 1 2 3 4 5 … 15 Next

```
> glimpse(iris)
Rows: 150
Columns: 5
$ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.8, 4.8, 4.3, 5.8, 5.7, 5.4, 5.1, 5.7, 5.1, 5.4, 5.1, …
$ Sepal.Width  <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.4, 3.0, 3.0, 4.0, 4.4, 3.9, 3.5, 3.8, 3.8, 3.4, 3.7, …
$ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.6, 1.4, 1.1, 1.2, 1.5, 1.3, 1.4, 1.7, 1.5, 1.7, 1.5, …
$ Petal.Width  <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.2, 0.1, 0.1, 0.2, 0.4, 0.4, 0.3, 0.3, 0.3, 0.2, 0.4, …
$ Species      <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa, setosa…
```

```
# For more details, see
?iris
```

Initial Overview of the Data Set (cont.)

```
# Iris scatter plot
ggplot(iris, aes(Petal.Length, Petal.Width)) +
  geom_point(aes(col=Species), size=4) +
  theme_bw()
```



As we can see, setosa is going to be clustered easier.

Meanwhile, there is noise between versicolor and virginica even when they look like perfectly clustered.

The Actual K-means Model

```
# The Actual K-means Model
set.seed(101)
irisCluster <- kmeans(iris[,1:4], center=3, nstart=20)
irisCluster
```

- In the kmeans function, it is necessary to set center, which is the number of groups we want to cluster to. In this case, we know this value will be 3.

```
> irisCluster
K-means clustering with 3 clusters of sizes 38, 62, 50

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.850000    3.073684     5.742105    2.071053
2     5.901613    2.748387     4.393548    1.433871
3     5.006000    3.428000     1.462000    0.246000

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [43] 3 3 3 3 3 3 3 3 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2
 [85] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 2 1 1 1 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1
[127] 2 2 1 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 2 1 1 2

Within cluster sum of squares by cluster:
[1] 23.87947 39.82097 15.15100
 (between_SS / total_SS =  88.4 %)

Available components:

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

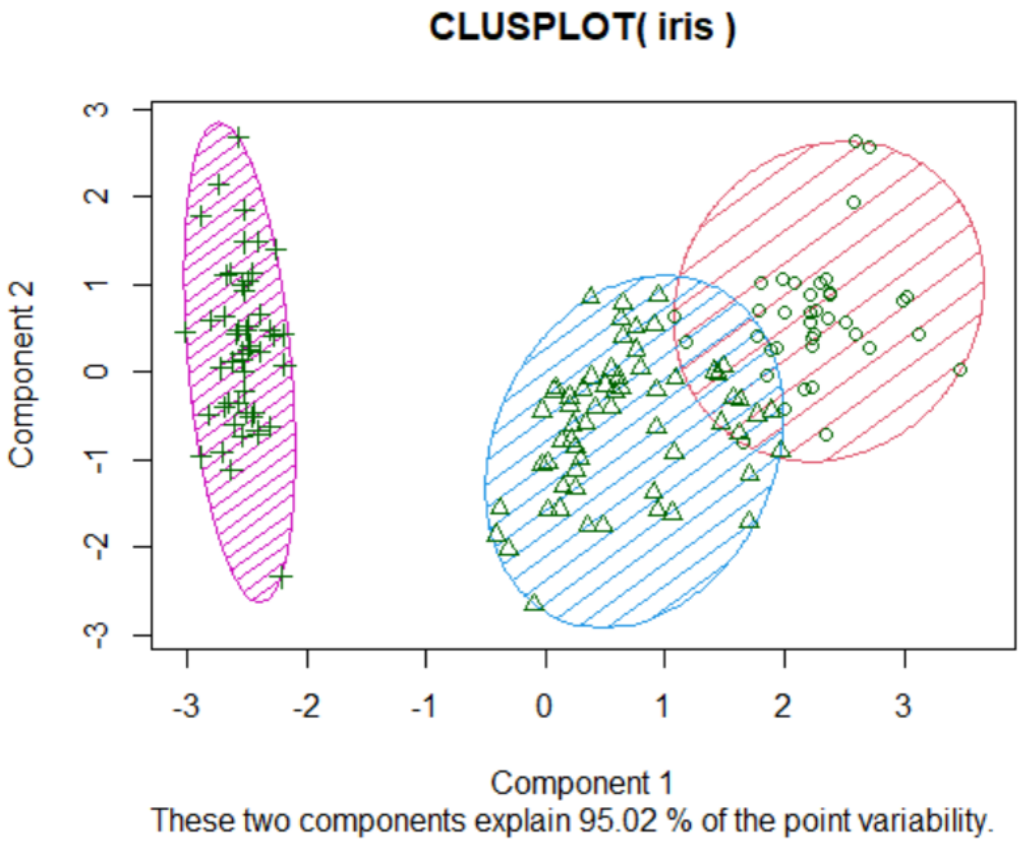
The Model Evaluation

```
# Comparison with original data
table(irisCluster$cluster, iris$Species)
# Plotting the clusters
clusplot(iris, irisCluster$cluster, color=T, shade=T, labels=0, lines=0)
```

- We can compare the predicted clusters with the original data and plot these clusters.

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

We can see the *setosa* cluster perfectly explained, meanwhile *virginica* and *versicolor* have a little noise between their clusters.

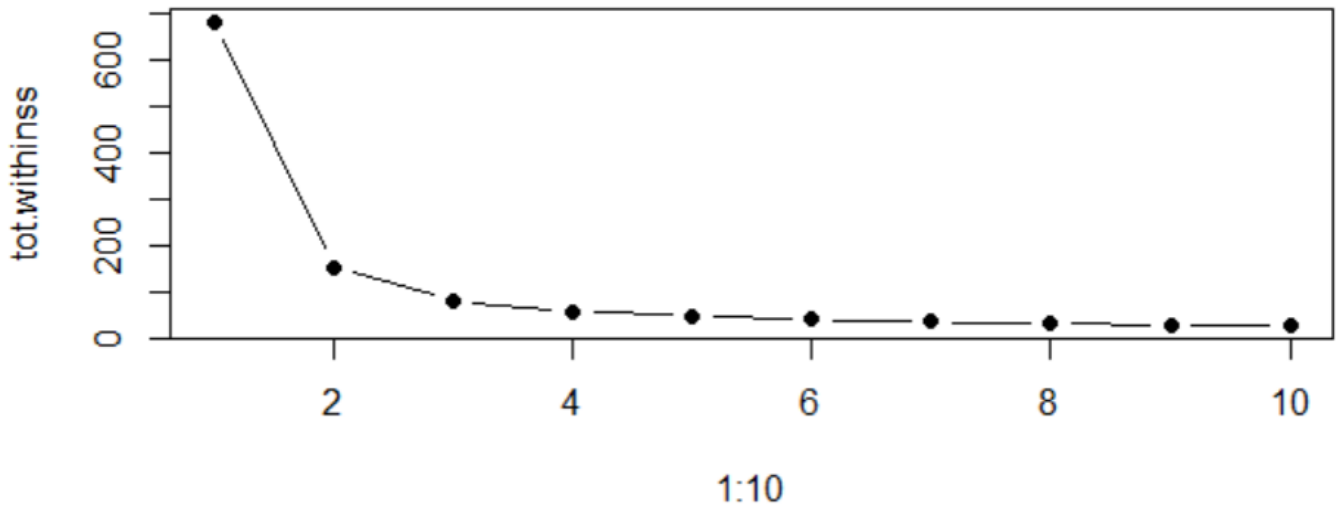


The Elbow Method

- we will not always have the labeled data. If we would want to know the exactly number of centers, we should have built the elbow method.

```r
# The Elbow Method
tot.withinss <- vector(mode="character", length=10)
for (i in 1:10){
  irisCluster <- kmeans(iris[,1:4], center=i, nstart=20)
  tot.withinss[i] <- irisCluster$tot.withinss
}
# Visualizing it
plot(1:10, tot.withinss, type="b", pch=19)
```

- As we saw, the optimal number of clusters is 3.



3. In-class Assignment

4. Final Exam instructions

Final Exam

- Group project report/presentation (Week 15): **30%**;
  - A group project report including the application of analytics principles and methods to a business issue will be required of the students.
  - A 20-minute PPT presentation and also an active participation in asking some questions to other presenters for a 15- minute discussion on the topic.
  - Students will be allocated to teams (up to 5 people in a group).
  - The assignment will be given today.
    - You have the remaining time and the time of our next class to work on this.
  - Please remember to upload your files in advance (script and ppt slides).
  - Only two groups will get the highest score.