Djillali Liabes University of Sidi Bel Abbes
Faculty of Exact Sciences
Department of Computer Sciences

Module : Apprentissage Automatique
1st year of Master's degeree in
Networks,Systems & Information Security(RSSI)
2021/2022

Solution To TD-01 + TD-02

Author:
Hadjazi Mohammed Hisham

Supervisor:
Pr.ELBERRICHI Zakaria

October 29, 2021

Contents

List of Figures

Chapter 1

Fiche TD-01 Solutions

1.1 Quelle est la relation entre l'IA et le machine learning ?

Machine Learning is considered to be a subset of Artificial Intelligence, To understand the relation between AI and ML we need to define each of them first.

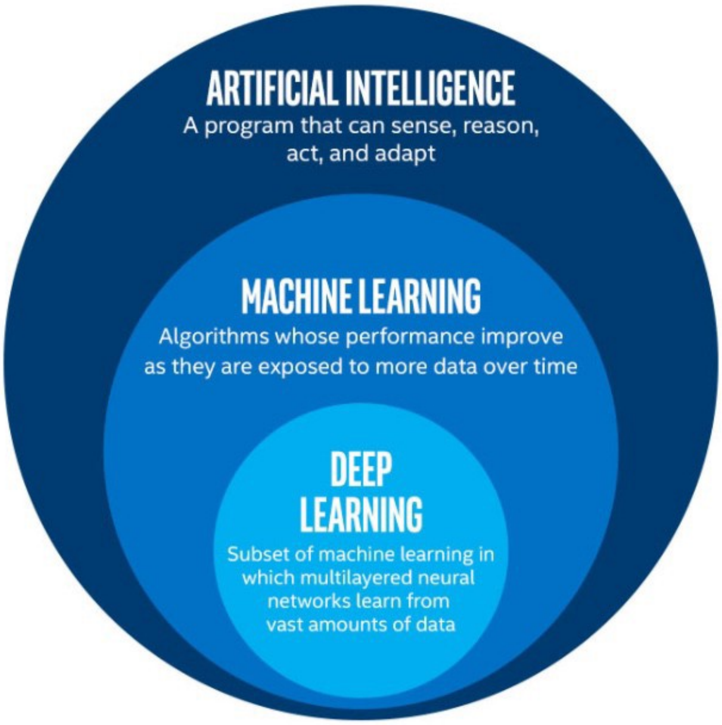
ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

FIGURE 1.1: Relation of ML and AI[3]

Artificial Intelligence is :

"Artificial intelligence is the science and engineering of making computers behave in ways that, until recently, we thought required human intelligence. "[3]

Pr.Andrew Moore, Carnegie Mellon University

So in order to make our computers or machines behave intelligently it is required to learn from something and now the field of Machine learning comes to play.

Machine learning is :

"Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience. "[2]

Pr. Tom M. Mitchell, Carnegie Mellon University

So with Machine learning we can use existing algorithms and statistical models or build our own algorithms and models that can learn from Data to teach a computer program how to perform a specific task, for example a program can read my previous years exam scores and from that data can answer the question of will student M.H.Hadjazi pass the exam of ML or not ?

1.2 Quelle est la différence fondamentale entre l'apprentissage supervisé et non supervisé ?

If we just try to understand the difference from the meaning of each word,

Supervised is to observe and direct the execution of the task done by the Machine learning model. but how to direct or what to direct ? it all comes to the dataset as in supervised machine learning models we provide a **labelled** dataset that we already have correct information about and the job of the model or algorithm is to generate rules from these models to classify other unlabelled input data according to the rules made with the labelled dataset that was given to it.

There are two main categorise of supervised machine learning models:

- Classification Models
- Regression Models

On the other hand Unsupervised Machine learning models are the opposite as we give it **unlabelled** data, and the model or algorithm job is to find patterns in the data to exploit and generate classes and groups by itself, its able to find relations between data that the human eye can't detect or recognise. therefore the main goal is to group that unlabelled dataset into groups and present the data in a more compressed manner that could be easier to understand.

Most Unsupervised Machine learning models can be categorised into these :

- Clustering
- Association

1.3 Quelle est la différence fondamentale entre la classification et la régression ?

The main difference is in the output as regression outputs in continues form **real numbers** for example estimating the price of a stock from previous data or the salary of an employee by using data of other employees. However classification outputs are **classes or categories**, note that these classes also can be numbers but predefined as a class or category, an example is trying to find if someone is going to pay back his loan or not.

There are other differences like what are we looking for in the data and how each is evaluated for example in classification we use the accuracy measurement, while in regression we use the sum of squared error method.

1.4 Pourquoi appelle-t-on la régression, régression ?

In the English Dictionary Regression is the process of going back to an earlier form. but the origin of the word came from Sir Francis Galton from thee 19th century who was many things but most importantly a statistician. he used the word Regression when he made his experiment to calculate heights of 205 sets of parents with adult children.

To make male and female heights directly comparable, he rescaled the female heights, multiplying them by a factor 1.08. Then he calculated the average of the two parents' heights (which he called the "mid-parent height") and divided them into groups based on the range of their heights.[1]

1.5 Pour un algorithme machine learning, est-ce un avantage ou un inconvénient d'avoir des données en quantité ?

This question is a bit tricky, but if we consider that the dataset is of a high quality with minimum amount of errors then yes in general this can lead to more accurate results and lead to a better understanding of the problems. In the other hand to make this dataset is probably one of the hardest jobs and most time consuming tasks during any machine learning process. not to mention the time to process the large amount of data, for example a ML engineer or data scientist needs to first fill data gaps and any missing data , then he needs to assess the relevance of the data, after that comes the detect of any anomalies, to finish with identifying and removing duplicates ,but I still think in a broader way it is better to have a larger dataset than a small one.

1.6 Pourquoi la qualité des données est très importante en machine learning ?

It is just a natural thing that a good input leads to a good output therefore the quality of a machine learning model is directly proportional to the quality of data. hence the field of **Data Quality Analysis**.

"Data preparation accounts for about 80% of the work of data scientists. "[4]

forbes.com

"Data collection and preparation are typically the most time-consuming activities in developing an AI-based application, much more so than selecting and tuning a model"[5]

MIT sloan survay

1.7 Qu'est-ce que l'overfitting ?

Overfitting is the problem of when a machine learning model tries to adapt itself too much to the dataset. it all comes to reducing the amount of error a model can generate when presented with new dataset, underfitting makes the model more buyest while overfitting makes the model with a higher variance and more related to the dataset it was learning from and both are undesired phenomena's in machine learning.

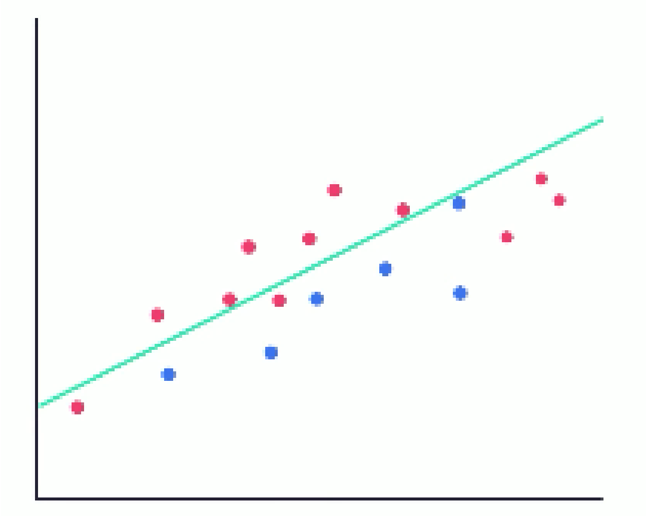

FIGURE 1.2: example of underfitting

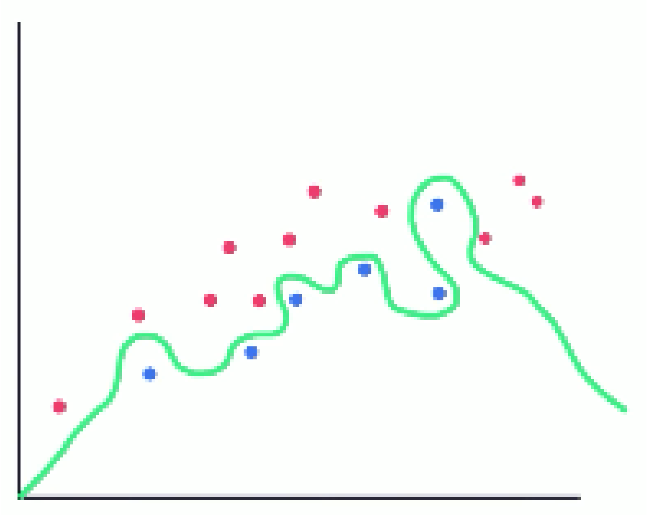

FIGURE 1.3: example of overfitting

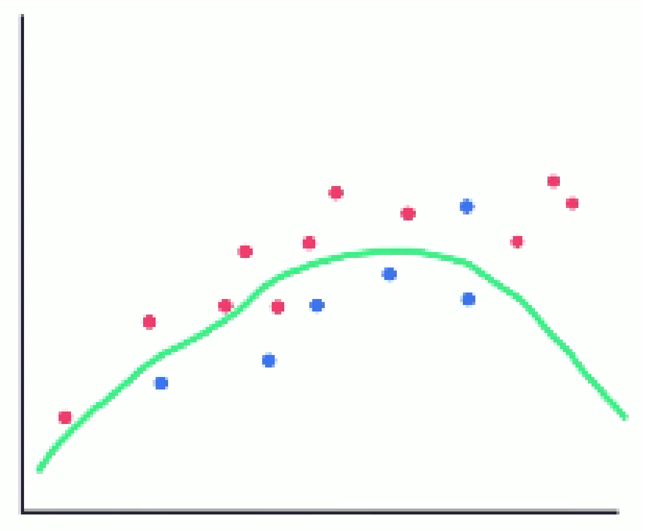

FIGURE 1.4: the desired output of the model

1.8 Pourquoi l'approche basée instances est différentes des autres approches ?

The Instance approach uses memorization instead of generalization, which in practice it uses databases to store all the data, of course this has its advantages and disadvantages and depends on the application in use.

1.9 Classifier selon l'arbre donné au cours l'instance Outlook= rainy ;Temperature= hot ;Humidity= normal ; Windy= true ?

Its true we don't have all the information in this instance but we have enough to reach a decision, therefore we don't play as we simply follow the decision tree with the values given. **Outlook—>Rainy—>Windy—>True—>NO**

1.10 Générer les règles produites par ce même arbre, essayer de classifier l'instance précédente à l'aide des règles cette fois. Expliquer pourquoi on dit que ces règles sont non ambiguës ?

The rules are as follow :

- If (Outlook == Sunny) and (Humidity == Normal) then Play
- If (Outlook == Sunny) and (Humidity == High) then No
- If (Outlook == Overcast) then Play
- If (Outlook == Rainy) and (Windy == True) then No
- If (Outlook == Rainy) and (Windy == False) then Play

The instance will work with the rules and generate the same result, the rules are unambiguous because we are creating with these rules **Boundaries** in our data space where they are separated from each other.

1.11 Pour l'exemple de régression, quelle sera le prix d'un appartement de 110 M 2 ?

Y = 4.7 + 1.3 x (110) = 147.7 , so the price is 147.7 according to the formula.

1.12 Pour l'exemple clustering hierarchique, donner les groupes pour chaque cas (2 groupes)(3 groupes)(4 groupes)(5 groupes) ?

2 groups

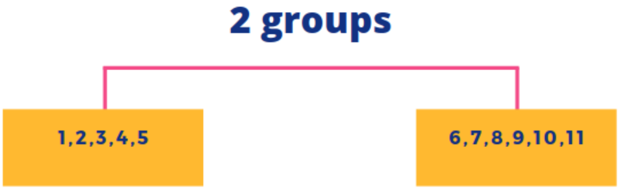

FIGURE 1.5: 2 groups

3 groups

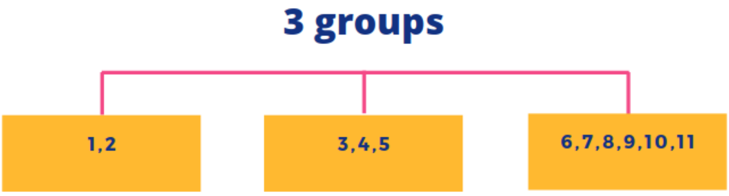

FIGURE 1.6: 3 groups

4 groups

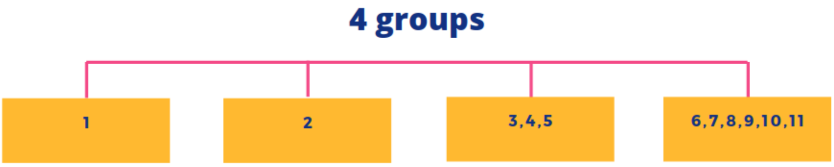

FIGURE 1.7: 4 groups

5 groups

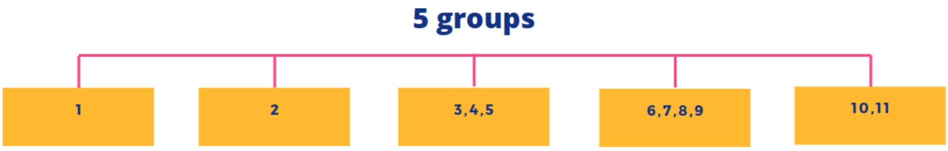

FIGURE 1.8: 5 groups

1.13 Quelle(s) application(s) machine learning vous utilisez quotidiennement ?

There many applications to count but in a daily basis, the following comes to my mind are :

- The Spam filter in my gmail app.
- When I take photos with my phone the camera detects faces automaticity.
- Voice commands to my phone (Google now).
- I'm sure Facebook is applying many ML algorithms on me, however not sure what they are other than the recommendations and automatic friend tagging.

There are many others but not on a daily basis like google maps and google Translation.

1.14 Quelle est le meilleur cours/livre de machine learning sur le Net ?

As this topic is new to me I don't really have any recommendation, however as i have searched I found the following resources to be useful.

For a textbook I'm looking now at

Fundamentals of Machine Learning for Predictive Data Anayltics: Algorithms, Worked Examples, and Case Studies by *John D. Kelleher, Brian Mac Namee, Aoife D'Arcy*. ISBN:0262044692.

Introduction to Machine Learning by *Ethem Alpaydin* ISBN:0262012111.

Machine Learning by *Thomas Mitchell* ISBN:0070428077.

The website of Pr.Saed Sayad, it is quite old and requires Adobe Flash and ActionScript.v3 support to display the interactive materials, but very easy to understand and straightforward to the point which I like.

Another good resource is the stanford university openclassroom website, this one is also old and requires Adobe flash to view the content but a good thing it is also available on youtube to watch.

It was recommended to me to join an introduction class of Machine learning provided by University of Washington on coursera and Instructed by Pr.Emily Fox and Pr.Carlos Guestrin, However the course is 7 months long and very practical as it has may python work on labs and exercises and I don't believe that in the time frame we have I can finish it. so maybe a course for the summer vacation.

1.15 Quel est le meilleur langage de programmation pour faire le machine learning ? Qu'attendez- vous pour le maitriser ?

Again I really don't know as I'm new to Machine Learning but I have observed that most of the programming examples are done in **Python**, which leads me to think it is probably the best language available for Machine Learning specially if there was a need to use one of its ML libraries like **TensorFlow**, **PyTorch**, **OpenCV**, **Numpy**, **SciPy** and many more, I also noticed the usage of **R** and **Matlab** in some materials, but again it seems python is dominating the field of ML.

The thing about Python is i always had the thinking of it as a slow interpreted language, that is the main reason why I didn't learn it, I tend to like C and C++ and maybe JavaScript and now looking at Rust, but I never liked Python and Java although I'm sure it wouldn't take me much to master Python if i want to.

Chapter 2

Fiche TD-02 Solutions

2.1 Exercise 2

First we need to generate our frequency table for each category and calculate the error rate for that category.

Office	Yes	No
H	3	2
S	2	3

The error rate is $\frac{4}{10}$.

Party	Yes	No
D	2	3
R	3	2

The error rate is $\frac{4}{10}$.

State	Yes	No
NY	1	2
NJ	1	3
CT	3	0

The error rate is $\frac{2}{10}$.

Since the state category has the minimum error rate of $\frac{2}{10}$, we will us it to generate our rules.

```
If (State == NJ) then NO; else yes;
```

and if we plug the instance Office = S ; Party = D ; State = NJ we will get **NO**.

2.2 Exercise 3

we first need to convert the continues numerical category of age int sections to be able to work with it :

19	27	29	35	38	39	40	41	42	43	43	43	45	55	55
Y	N	Y	Y	Y	Y	Y	Y	N	Y	N	Y	N	N	N

after finding the separation lines which in our case its only one line between 41 and 42. we start building our frequency tables and calculate the error rate for each category.

Age	Yes	No
19-41	7	1
42-55	2	5

The error rate is $\frac{3}{15}$.

Sex	Yes	No
Male	3	5
Female	6	1

The error rate is $\frac{4}{15}$.

Salary	Yes	No
Tres Bon	2	0
Bon	1	3
Moyen	4	1
Faible	2	2

The error rate is $\frac{4}{15}$.

Assurance	Yes	No
Y	3	0
N	6	6

The error rate is $\frac{6}{15}$.

we conclude that Age is the one category to use and the rules for it are :

```
If (Age >=19 && <= 41 ) then YES;
else if (Age >=42 && <= 55 ) then NO;
```

If we plug our instance Age= 50 ; Sexe = M ; Salaire= Bon ; Assurance = Y. the answer will be **NO**

Bibliography

[1] Minitab Blog Editor. *So Why Is It Called Regression Anyway?* en. URL: https://blog.minitab.com/en/statistics-and-quality-data-analysis/so-why-is-it-called-regression-anyway (visited on 10/24/2021).

[2] Peter High. *Carnegie Mellon Dean Of Computer Science On The Future Of AI*. en. URL: https://www.forbes.com/sites/peterhigh/2017/10/30/carnegie-mellon-dean-of-computer-science-on-the-future-of-ai/ (visited on 10/24/2021).

[3] Ezra Lazuardy. *How Machine Learn*. en. July 2020. URL: https://ezralazuardy.medium.com/how-machine-learn-c2f73f60ef14 (visited on 10/24/2021).

[4] Gil Press. *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. en. URL: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/ (visited on 10/24/2021).

[5] Philipp Gerbert and Martin Sam Ransbotham Reeves David Kiron. *Reshaping Business With Artificial Intelligence*. en-US. URL: https://sloanreview.mit.edu/projects/reshaping-business-with-artificial-intelligence/ (visited on 10/24/2021).