AI

Mini project: IRIS Flower Classification

By student: Youssef Ahmed Abbas

Arafa Arafa Abd El-Mawgod

Abdallah Mamdouh Mohamed Alhaddad

Presented to: Dr. Hesham El-Deeb

T.A Yomna Fathy

Chapter 1: Introduction

1. Introduction:

Deep Learning as a powerful approach to achieve Artificial Intelligence, has been widely used in pattern recognition, a very basic skill for humans but a challenge for machines. Nowadays, with the development of computer technology, pattern recognition has become an essential and important technique in the field of Artificial Intelligence. The pattern recognition can identify letters, images, voice or other objects and also can identify status, extent or other abstractions.

1.1 Background

Since the computer was invented, it has begun to affect our daily life. It improves the quality of our lives; it makes our life more convenient and more efficient. A fascinating idea is to let a computer think and learn as a human. Basically, machine learning is to let a computer develop learning skills by itself with given knowledge. Pattern recognition can be treated like computer being able to recognize different species of objects. Therefore, machine learning has close connection with pattern recognition. In this project, the object is the Iris flower. The data set of Iris contains three different classes: Sentosa, Versicolor, and Virginica. The designed recognition system will distinguish these three different classes of Iris.

1.2 Objectives

After the project has been settled, the computer should have the ability to aggregate three different classifications of Iris flower to three categories. The whole workflow of machine learning should work smoothly. The users do not need to tell the computer which class the Iris belongs to, the computer can recognize them all by itself. The final purpose of this project is to let everyone who read this thesis have a basic understanding of machine learning. Even through someone never touched this field, they can realize that the machine learning algorithm will become more popular and useful in the future. Moreover, the case study

of Iris recognition will show how to implement machine learning by using Scikit-learn software.

1.3. Collecting data set

Fisher's Iris data base is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two; the latter are not linearly separable from each other.

The data base contains the following attributes:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm

Classes:

- Iris Setosa
- Iris Versicolour
- Iris Virginica

1.4 Using many algorithms to achieve classification

Support Vector Machines, Naive Bayes, KNN, Decision Tree algorithm was used for classification Iris classes in this project. There are many different kinds of machine learning algorithms applied in different fields. Choosing a proper algorithm is essential for each machine learning project. For pattern recognition

1.5 Evaluating result

Evaluation will be the final part of this project. For each scientific project, the final result should be tested and evaluated if that is acceptable. The result will be automatically shown in the end of the program execution. For every machine learning algorithm, exceptions will always exist. In order to find the best result, result analyzing is necessary.

Chapter 2: Literature review

2.1 Basic introduction to machine learning

Learning is a very important feature of Artificial Intelligence. Many scientists tried to explain and give a proper definition for learning. However, learning is not that easy to cover with few simple sentences. Many computer scientists, sociologists, logicians and other scientists discussed about this for a long time. Some scientists think learning is an adaptive skill so that the system can perform the similar task better in the next time. Others claim that learning is a process of collecting knowledge. Even though there is no proper definition for learning skill, we still need to give a definition for machine learning. In general, machine learning aims to find out how the computer algorithms can be improved automatically through experience. Machine learning has an important position in the field of Artificial Intelligence. At the beginning of development of Artificial Intelligence (AI), the AI system does not have a thorough learning ability so the whole system is not perfect. For instance, a computer cannot do self-adjustment when it faces problems. Moreover, the computer cannot automatically collect and discover new knowledge. The inference of the program needs more induction than deduction. Therefore, computer only can figure out already existing truths. It does not have the ability to discover a new logical theory, rules and so on.

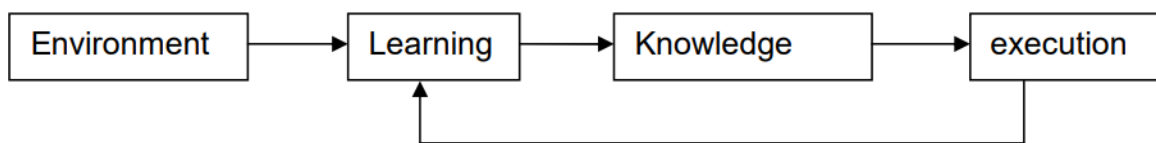## 2.11 Fundamental structure of machine learning system

Figure 1. Learning system structure

Figure 1 shows the basic work structure of machine learning. The structure of machine learning system consists of four main parts: Environment, Learning, Knowledge base and execute. The environment represents a combination of information from external information source. That would include any information from persons or references materials and so on. It is the learning source for the whole machine learning system. The environment is responsible for transferring data to the system. The quality of the data is very important. In the reality, the

data can be complex so it will be difficult for computer to process. In addition, the data can be incomplete, therefore the illation from the learning system is unauthentic. Learning is the procedure of transferring the information from the environment to knowledge. The environment will give the computer external information, and then the computer will go through all the information by using analysis, comprehensive induction and analogy to process this information to knowledge. At last, all the knowledge would be imported to the knowledge base. The knowledge base can be treated as the brain of the whole machine learning system. Different kinds of form and content of knowledge can have different influence on the designing of a machine learning system. Knowledge representation modes are eigen vector, First-order logic statements, production rule, and semantic system. Every mode has its own advantages and disadvantages. Therefore, when users want to design a machine leaning system, a good knowledge representation mode is very important for the whole system.

A proper knowledge representation mode should satisfy four basic requirements:

1. Strong expression
2. Easy theorization
3. Easy to modify the knowledge base
4. Easy to expand the knowledge representation

Moreover, a machine learning system cannot create new knowledge from nothing. It always needs original knowledge to understand the information from environment. Then the computer can use this information to learn new knowledge step by step. In conclusion, learning process in the whole system is a process of expansion and perfection of the knowledge base. Execution is the core of the whole machine learning system. Each part of the system aims to make a progress for the execution part. On the other hand, execution also has a connection to each part, especially the learning process. The purpose of a learning process is to make the execution perfect. At last, the complexity, feedback and transparency of execution also has an influence on the learning process.

Complexity:

The complexity of knowledge is different depending on the different learning tasks. Some tasks are quite easy, so the system does not need too much information. If the tasks are quite difficult, the system will need more information to learn.

Feedback:

After the execution, the execution system can evaluate the leaning task, and then give feedback information to the learning process. The learning process will try to decide whether to collect information from environment to modify or improve the knowledge in knowledge base or not based on the feedback. Transparency From the result of execution part, users can easily see the structure of the knowledge base and give the evaluation for it.

Chapter 3: Classification Algorithms

3.1 Deep Learning Model :

Input Layer ∈ $\mathbb{R}^4$ Hidden Layer ∈ $\mathbb{R}^6$ Output Layer ∈ $\mathbb{R}^3$
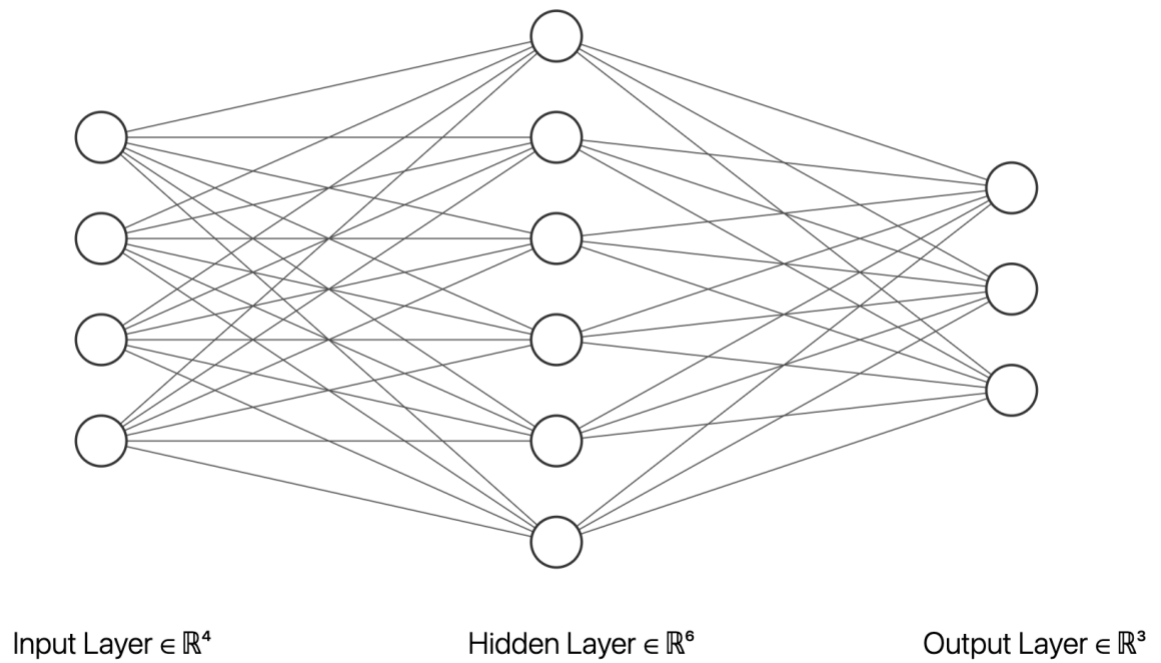
Figure 1 Deep Learning Model with 2 H.Layer

```
_____
Layer (type)                    Output Shape              Param #
===============================================================
dense_1 (Dense)                 (None, 6)                 30
_____
dense_2 (Dense)                 (None, 3)                 21
===============================================================
Total params: 51
Trainable params: 51
Non-trainable params: 0
_____
```

```
┌─────────────────────────────┐
│      140698933090896        │
└─────────────────────────────┘
              │
              ▼
┌──────────────────┬──────────┬──────────────┐
│                  │ input:   │ (None, 4)    │
│ dense_1: Dense   ├──────────┼──────────────┤
│                  │ output:  │ (None, 6)    │
└──────────────────┴──────────┴──────────────┘
              │
              ▼
┌──────────────────┬──────────┬──────────────┐
│                  │ input:   │ (None, 6)    │
│ dense_2: Dense   ├──────────┼──────────────┤
│                  │ output:  │ (None, 3)    │
└──────────────────┴──────────┴──────────────┘
```

Chapter 4: Implementation

4.1 Python

Python is a programming language created by Guido van Rossum in 1989. Python is an interpreted, object-oriented, dynamic data type of high-level programming languages. (Python Software Foundation 2013). The programming language style is simple, clear and it also contains powerful different kinds of classes. Moreover, Python can easily combine other programming languages, such as C or C++.

4.2 SciKit-learn

Scikit–learn is an open source machine learning library for the Python programming language. It features various classification, regression, and clustering algorithms and is designed to interoperate with the Python numerical libraries NumPy and SciPy (Pedregosa et al. 2011). SciKit-learn contains the Kmeans algorithm based on Python and it helps to figure out how to implement this algorithm in programming.

4.3 Numpy and Matplotlib:

In Python, there is no data type called array. In order to implement the data type of array with python, numpy are the essential libraries for analyzing and calculating data. They are all open source libraries. Numpy is mainly used for the matrix calculation.

Mathplotlib is a famous library for plotting in Python. It provides a series of API and it is suitable for making interactive mapping. In this case, we need to use it to find the best result visually.

4.4 Preparing the Iris flower data set:

The data set of Iris flower can be found in UCI Machine Learning Repositor (Bache & Lichman 2013). In this thesis, the famous Fisher's Iris data set will be used. The data set of Iris flower can be also found in the Scikit-learn library. In sitepackages, there is a folder named sklearn. In this folder, there is a datasets subfolder to contain many kinds of data sets for machine learning study. The data set can be found in Appendix 1. In the species of this table, 0 represents setosa, 1 represents versicolor, 2 represents virginica. In the process of preparing a training data set and a testing data set, the greatest problem is how to find the most appropriate way to divide the data set into training data set and testing data set. In some cases, by using sampling theory and estimation theory, we can separate the whole data set into training data set and testing data set. However, sometimes, the method would be changed. The attributes and the property of the data set would be different in various machine learning objects. Thus, in this kind of situation, in order to achieve a better result of machine learning, the data set will be separated according to the property of attributes of the data set.

4.5 Result:

	Train	validation
Accuracy	0.9833333392937978	0.9666666785875956
Loss	0.1688777469098568	0.22880137463410696