ABIYAANTRIX AND SAPIENCE ACADEMY

A

PROJECT REPORT

ON

"IRIS DATASET"

FOR THE COURSE MACHINE LEARNING

SUBMITTED BY:

MEGHANA A S

VARSHA KRISHNA

AISHWARYA SIMHA C P

GUIDED BY:

PROF KRISHNA KASHYAP

SUBMITTED TO:

ABIYAANTRIX AND SAPIENCE ACADEMY

FOR THE ACADEMIC YEAR

2018

CONTENTS

INTRODUCTION:

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

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: Setosa, Versicolour, 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

The data set contains three classes of 50 instances each, where each class refers to a type of iris plant. Each class is linearly separable from the other two classes. The attribute information will include sepal length, sepal width, and petal length and petal width. All of them have the same unit, *cm*.

1.4 Using K-means algorithm to achieve clustering

K-means algorithm was used for clustering 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, K-means is a classic clustering algorithm. In this project, K-means algorithm can be implemented with the Python programming language.

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 analysing is necessary.

K-means clustering

As mentioned earlier in this thesis, machine learning consists of many kinds of learning algorithms for different learning methods. In this thesis, the classification information is assumed to be unlabeled. In this case, the best choice in unsupervised learning is the K-means clustering algorithm.

Introduction to clustering

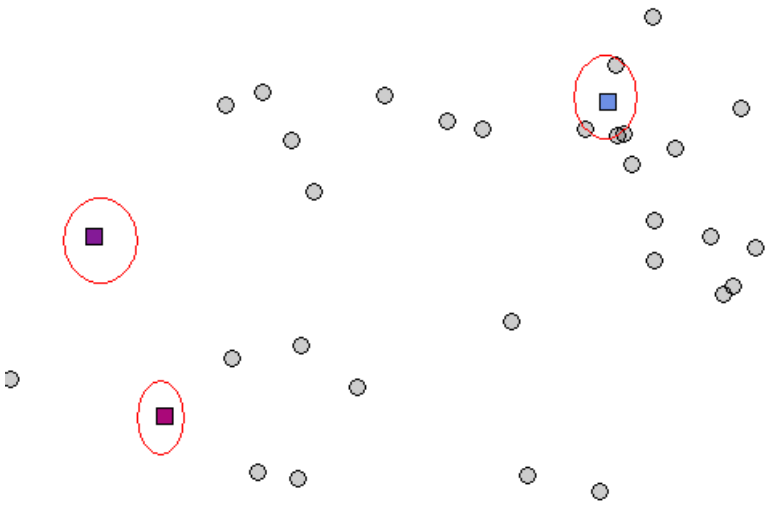
The K-means clustering algorithm is one of the most popular clustering algorithms in the world. Clustering aims to classify data from the whole data space. The difference between each data object in the same class is similar. However, the difference between each data objects in different classes is large. Clustering belongs to the unsupervised learning method and it can automatically sort data sets.

Basically, the result of clustering algorithm is to find the same classification of different data in the whole data sets. For example, the data set contains monkey, lion, banana, apple, four different data units. After clustering, these four data will be divided into two main sections. One section includes monkey and lion representing the class of animals. The other section includes apple and banana, this section representing the class of fruits.

A clustering algorithm groups all the same kind of data into one single class. The computer will recognize the specific features of all data so that it can separate data to the proper classes.

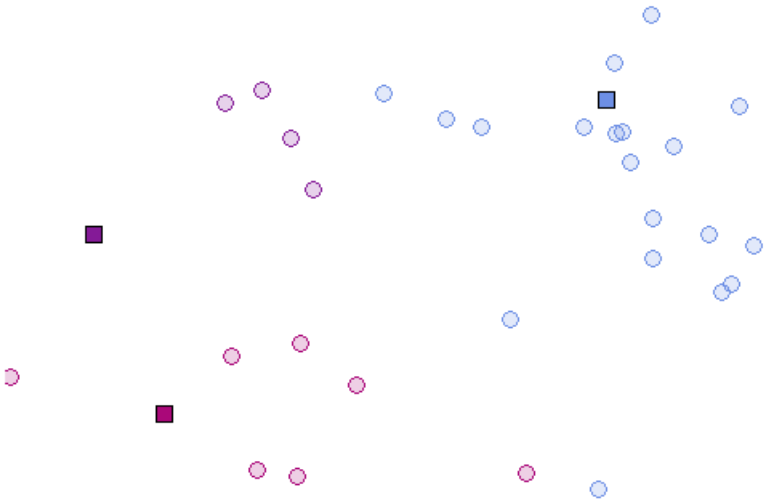
K- MEANS CLUSTERING STEPS

The following tables show a sample of workflow of K-means. The dataset contains 30 samples and the number of clusters is 3



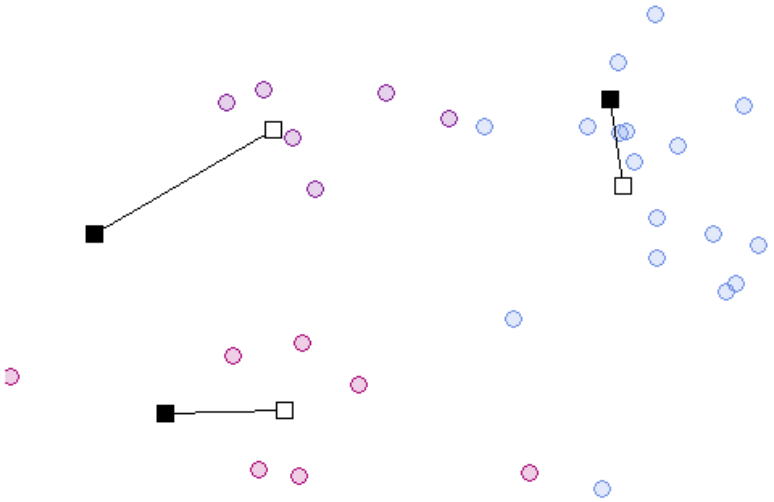
K- MEANS CLUSTERING STEP 1

Now the system generates three cluster points with randomly. There are three different colours: purple(Top left), blue(Top right), pink(Bottom left). These three colours stand for three different clusters.



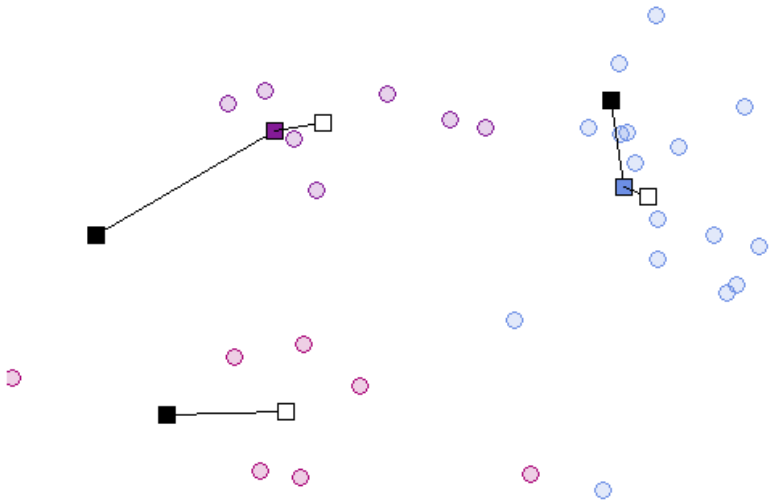
K- MEANS CLUSTERING STEP 2

With the initial the point of k, then the system should calculate the distance of each object to the cluster centres. The new blank box indicates the new cluster centre.



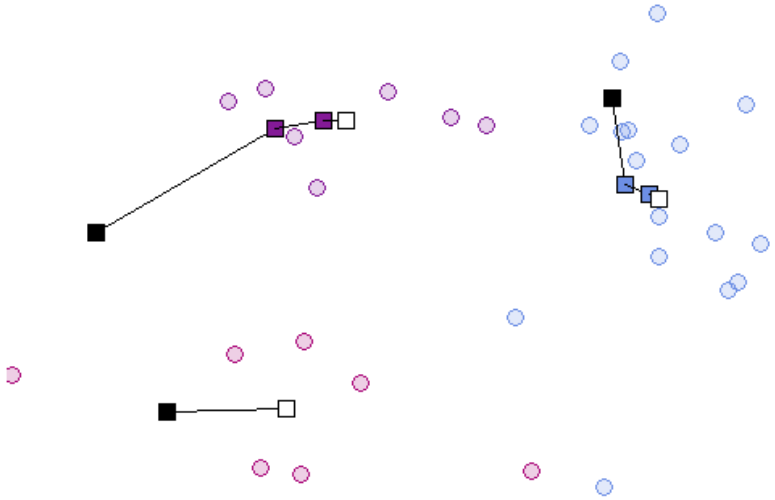
K – MEANS CLUSTERING STEP 3

If there are still some objects missing, then the system will continue to find the new centroid for each cluster until all the samples are grouped. In next figure, the k cluster centroids move to a new place and the calculation is continued.



K – MEANS CLUSTERING STEP 4

The next table is the final result. The principle of K-means algorithm is to make all samples in one cluster to be closer to each other, but the distance of each clusters should be larger.



K- MEANS CLUSTERING STEP 5

IMPLEMENTATION

2.1 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++.
As a successful programming language, it has its own advantages:
Simple & easy to learn: The concept of this programming language is as simple as it can be. That makes it easy for everyone to learn and use. It is easy to understand the syntax.
Open source: Python is completely free as it is an open source software. Several of open source scientific computing storage has the API for Python. Users can easy to install Python on their own computer and use the standard and extend library.
Scalability: Programmers can write their code in C or C++ and run them in Python.

2.2 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 . SciKit-learn contains the K-means algorithm based on Python and it helps to figure out how to implement this algorithm in programming.

2.3 Numpy, SciPy and Matplotlib

In Python, there is no data type called array. In order to implement the data type of array with python, NumPy and SciPy are the essential libraries for analysing and calculating data. They are all open source libraries. Numpy is mainly used for the matrix calculation. SciPy is developed based on NumPy and it is mainly used for scientific research.
By using them in Python programming, they can be used with two simple commands:
>>> import NumPy
>>> import SciPy
Then Python will call the methods from NumPy and SciPy.
Matplotlib 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.

2.4 Preparing the Iris flower data set

The data set of Iris flower can be found in UCI Machine Learning Repositor (Bache & Lachman 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.
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.

The K-means algorithm and unsupervised learning does not use a training data set to compute the training sample. Therefore, there is no need to separate the dataset into a training data set and a testing data set. It can simply use this dataset to get the result of clustering.

2.5 Machine learning system design

In general, the principles of machine learning system design should follow two basic requirements :
- the model selection and creation and
- the learning algorithm selection and design.

In addition, different models can have different learning systems. On the other hand, the objective function is also different in different learning models. The objective function can help the machine to establish a learning system. Moreover, the accuracy and complexity of different algorithms would be the most important factor of the learning system. If the chosen algorithm is not very adaptive to the learning system, then the efficiency and result of the learning system would be reduced. The selection of training data set can have an influence on learning performance and feature selection.

2.6 Using Python to implement the program

For good implementation and good compatibility, Python version 2.7 will be in use. The Integrated Development Environment in this case will be PyScripter. By using the Scikit-learn software package, there is no need to write a program to implement the K-means algorithm. After the installation has been finished, the K-means algorithm source code can be found in sklearn library. The source code of K-means clustering of Iris recognition can be found in the official website of Scikit-learn.
First of all, we need to import the library of NumPy, dataset of Iris, K-means and

Axes3D into the program. These are needed for this program. Numpy can helps to implement the K-means algorithm, the Iris dataset is the main data to be analysed, Axes3d can make 3D outputs of this program, and the image will be more visual.

```
>>> import numpy as np
>>> import pylab as pl
>>> from mpl_toolkits.mplot3d import Axes3D
>>> from sklearn.cluster import KMeans
>>> from sklearn import datasets
```

Then, the program loads the Iris dataset and sets the centroid value and the number of clusters. In this program, the number of k clusters will be chosen as three and eight. In order to make a comparison, the third one will be the number of clusters 3, but with a bad initialization on the classification process. The initialization number has changed to 1. The default number is 10. Therefore the times the algorithm executes with different centroid seeds is reduced. This shows what happens to the result if the whole system has a bad initialization.

```
>>> np.random.seed(5)
>>> centers = [[1, 1], [-1, -1], [1, -1]]
>>> iris = datasets.load_iris()
>>> X = iris.data
>>> y = iris.target
>>> estimators = {'k_means_iris_3': KMeans(n_clusters=3),
...               'k_means_iris_8': KMeans(n_clusters=8),
...              'k_means_iris_bad_init': KMeans(n_clusters=3, n_init=1,
...                                              init='random')}
...
>>>
```

The result is shown as a table with three feature vectors. The feature vectors consists of petal width, sepal length and petal length. The output table will be three-dimensional.

```
>>> fignum = 1
>>> for name, est in estimators.iteritems():
...     fig = pl.figure(fignum, figsize=(4, 3))
...     pl.clf()
...     ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)
...     pl.cla()
...     est.fit(X)
...     labels = est.labels_
...     ax.scatter(X[:, 3], X[:, 0], X[:, 2], c=labels.astype(np.float))
...     ax.w_xaxis.set_ticklabels([])
...     ax.w_yaxis.set_ticklabels([])
...     ax.w_zaxis.set_ticklabels([])
...     ax.set_xlabel('Petal width')
...     ax.set_ylabel('Sepal length')
...     ax.set_zlabel('Petal length')
...     fignum = fignum + 1
...
```

Then the program will show the standard plot of K-means clustering of Iris flower in supervised learning technique. The standard result of clustering is labeled with three species.

```
>>> fig = pl.figure(fignum, figsize=(4, 3))
>>> pl.clf()
>>> ax = Axes3D(fig, rect=[0, 0, .95, 1], elev=48, azim=134)
>>> pl.cla()
>>> for name, label in [('Setosa', 0),
...                     ('Versicolour', 1),
...                     ('Virginica', 2)]:
...     ax.text3D(X[y == label, 3].mean(),
...               X[y == label, 0].mean() + 1.5,
...               X[y == label, 2].mean(), name,
...               horizontalalignment='center',
...               bbox=dict(alpha=.5, edgecolor='w', facecolor='w'))
...
<mpl_toolkits.mplot3d.art3d.Text3D object at 0x0000000016A7D4A8>
<mpl_toolkits.mplot3d.art3d.Text3D object at 0x0000000016A7D438>
<mpl_toolkits.mplot3d.art3d.Text3D object at 0x0000000016A7D588>
>>>
```

The next step is to reorder the labels with the matched colours for the cluster results. After that all of the figures will be shown on the screen.

```
>>> y = np.choose(y, [1, 2, 0]).astype(np.float)
>>> ax.scatter(X[:, 3], X[:, 0], X[:, 2], c=y)
<mpl_toolkits.mplot3d.art3d.Patch3DCollection object at 0x000000000EE22240>
>>> ax.w_xaxis.set_ticklabels([])
[]
>>> ax.w_yaxis.set_ticklabels([])
[]
>>> ax.w_zaxis.set_ticklabels([])
[]
>>> ax.set_xlabel('Petal width')
<matplotlib.text.Text object at 0x000000000F8E4BE0>
>>> ax.set_ylabel('Sepal length')
<matplotlib.text.Text object at 0x000000000F8EED30>
>>> ax.set_zlabel('Petal length')
<matplotlib.text.Text object at 0x000000000F8F0CC0>
>>> pl.show()
>>>
```

Evaluating results

The result is shown in four images for the clustering results. Figure 1 will be the result with eight clusters. Figure 2 shows the result with three clusters.

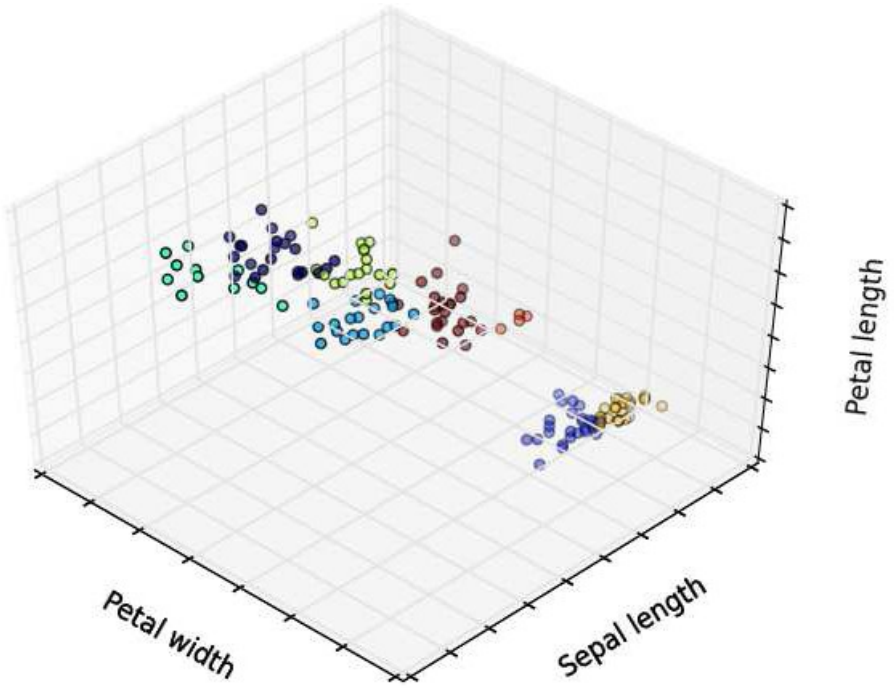


Figure 1 Clustering of Iris dataset with eight clusters

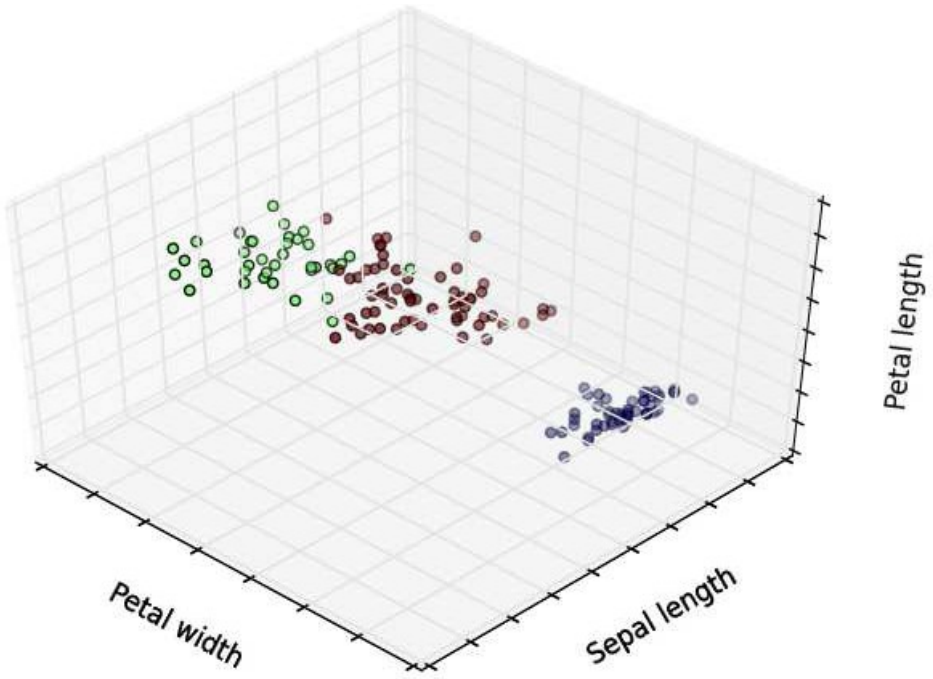


Figure 2 Clustering of Iris dataset with three clusters

As seen in Figure 1 and 2 , the whole dataset is separated into eight clusters in Figure 1 and three clusters are shown in Figure 2 with different colours. In Figure 1, most of the samples stick together, it is really hard to distinguish them very clearly. The differences between each sample is small. In this case, the cluster result is not acceptable. On the other hand, in Figure 2, it can be easily seen that the cluster result is much better than in Figure 1. Even though there are still some overlapping parts between green and purple, but it quite clear to see the difference between these three clusters. This case shows the importance of choosing the number of clusters for K-means algorithm. Sometimes for the real datasets, it is difficult to know how many data sets should be used. Therefore, it is quite hard to choose the number of clusters. One method is to use the ISODATA algorithm, through the merging and division of clusters to obtain a reasonable number of k.

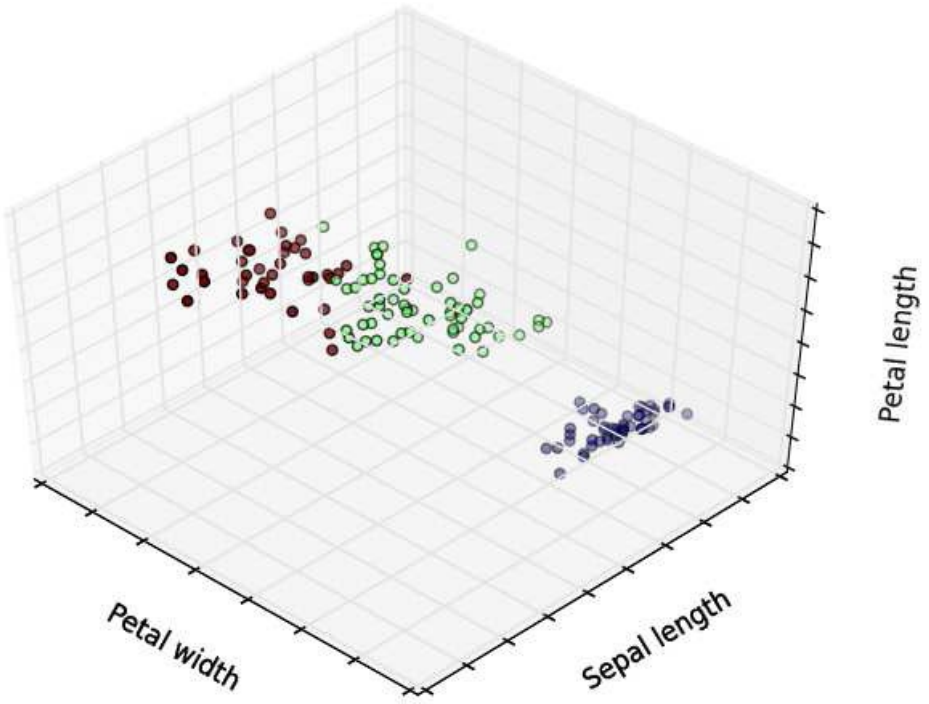


Figure 3 , shows the cluster result with three clusters but bad initialization.

We can see that some of the samples change their class compare to the Figure 1. With a random initialization number, the system will obtain different cluster results. Therefore, a random initialization number is very important for a good cluster result. However, we do not know what could be a good initialization number. In this case, in some machine learning systems, the scientists will choose GA(Genetic Algorithm) to have the initialization point.

Figure 4 below illustrates a standard result of K-means clustering of Iris recognition. The term “ground truth” refers to the classification of training datasets in supervised learning. The number of clusters are three and with a good initialization point. This is the best classification of all shown here. The whole dataset has been separated properly and each dataset has good differences. In Figure 2, it shows the standard result of classification in unsupervised learning. Compare to this figure, Figure 2 still has some small differences but it still works very well. Almost every data belongs to the right place.

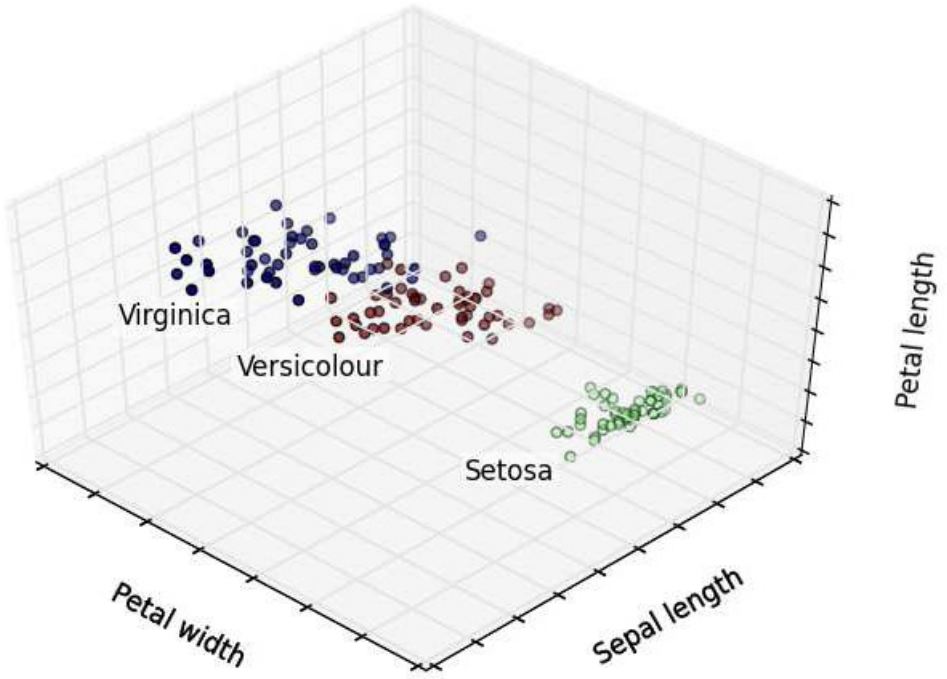


Figure 4 Clustering of Iris dataset in ground truth

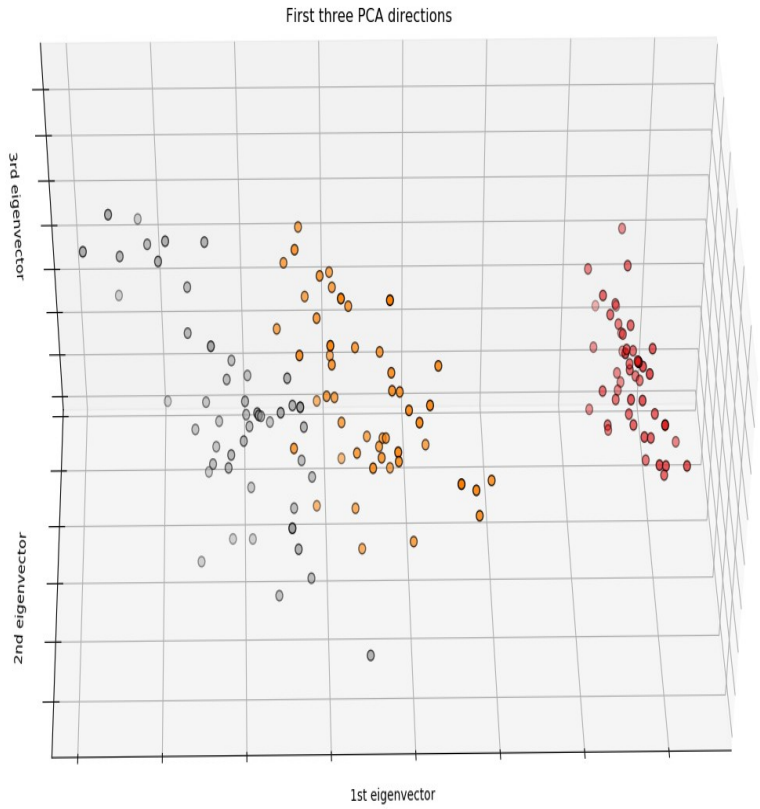
These results show the effect that the number of k and the random initialization number have on the clustering result. It is also possible to see the advantages and disadvantages of the K-means clustering algorithm.

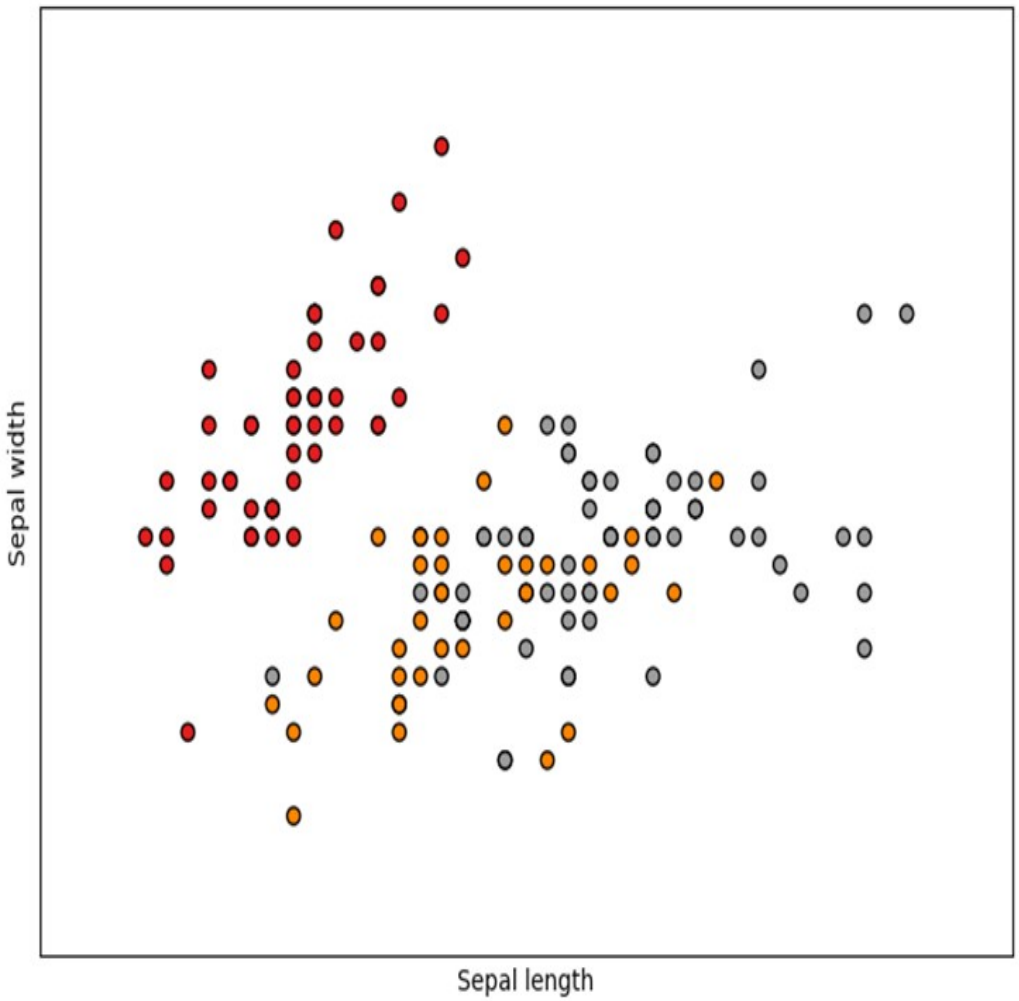
SOURCE CODE

```
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from sklearn import datasets
from sklearn.decomposition import PCA

iris = datasets.load_iris()
X = iris.data[:, :2]
y = iris.target

x_min, x_max = X[:, 0].min() - .5, X[:, 0].max() + .5
y_min, y_max = X[:, 1].min() - .5, X[:, 1].max() + .5

plt.figure(2, figsize=(8, 6))
plt.clf()

plt.scatter(X[:, 0], X[:, 1], c=y, cmap=plt.cm.Set1,edgecolor='k')
plt.xlabel('Sepal length')
plt.ylabel('Sepal width')
plt.xlim(x_min, x_max)
plt.ylim(y_min, y_max)
plt.xticks(())
plt.yticks(())

fig = plt.figure(1, figsize=(8, 6))
ax = Axes3D(fig, elev=-150, azim=110)
X_reduced = PCA(n_components=3).fit_transform(iris.data)
```

```
ax.scatter(X_reduced[:, 0], X_reduced[:, 1], X_reduced[:, 2], c=y,  cmap=plt.cm.Set1,
edgecolor='k', s=40)

ax.set_title("First three PCA directions")

ax.set_xlabel("1st eigenvector")

ax.w_xaxis.set_ticklabels([])

ax.set_ylabel("2nd eigenvector")

ax.w_yaxis.set_ticklabels([])

ax.set_zlabel("3rd eigenvector")

ax.w_zaxis.set_ticklabels([])


plt.show()
```

OUTPUT SNAPSHOTS:




Sepal width
Sepal length

FUTURE SCOPE

The Iris recognition case study above shows that the Machine Learning algorithm works well in this pattern recognition. The speed of computing is fast and the result is acceptable. However, the K-means clustering algorithm is just one of the clustering algorithm in unsupervised learning. There are more algorithms for different work objectives in different scientific fields.

As it is mentioned above, Machine Learning can be separated into supervised learning and unsupervised learning. However, sometimes, a whole dataset have both labeled data and unlabeled data. In order to process this kind of dataset, a new learning method called Semi-supervised(SSL) Learning has become a research hotspot. Because of this learning method, both machine learning and pattern recognition have a new research direction. It saves a lot of time and human resource to label those large amounts of unlabeled data. The SSL is also significant on improving learning performance of a computer.

Moreover, a learning system always consists of two parts, learning and environment. The environment gives knowledge to the computer and the computer will transfer this knowledge and store them and select useful information to implements different learning objectives. Therefore, different learning strategies can also be separated into rote learning, learning from instruction, learning by deduction, learning by analog, explanation-based learning and learning from induction. All of them have different algorithms to process different work objectives.

The implemented case in this thesis is only a simple example of machine learning and pattern recognition. Moreover, the K-means algorithm used in this thesis is a basic algorithm. However, if the data set has many feature dimensions and it is complicated, and if the learning objective is not that simple, the K-means algorithm cannot be used.
Nowadays, GA (Genetic Algorithm), Artificial neural network and other machine learning algorithms have become more and more stable and useful.

CONCLUSION

With the rapid development of technology, AI has been applied in many fields. Machine learning is the most fundamental approach to achieve AI. This thesis describes the work principle of machine learning, two different learning forms of machine learning and an application of machine learning. In addition, a case study of Iris flower recognition to introduce the workflow of machine learning in pattern recognition is shown. In this case, the meaning of pattern recognition and how the machine learning works in pattern recognition has been described. The K-means algorithm, which is a very simple machine learning algorithm from the unsupervised learning method is used. The work also shows how to use SciKit-learn software to learn machine learning