ARTIFICIAL INTELLIGENCE

COMP338

K-Nearest Neighbors Algorithm

(KNN)

Name: Naser Alden Masalma

ID: 1131232

Instructor: Dr. Majdi Mafarja

Machine Learning:

Machine learning is classified as a major branch of artificial intelligence, we can define it as a science that allows the computer to act and make decisions without being programmed to do that act or act openly and correctly, or it can be said that it learns how to respond to a particular event independently without being explicitly informed by the programmer.

This was often strange, because we used to teach the computer an algorithm and a series of instructions to solve a particular problem, for example, we provide the device with an algorithm to arrange the numbers up. But it may seem difficult and impossible to sort e-mail messages and get rid of annoying ads, because there is no fixed algorithm for doing so.

Machine learning processes are divided into two main types:

1- **Supervised Learning (Predictive Learning):** In this type of learning, the computer is trained with inputs that know what its output is, as a group of pre-classified mail messages is important or not important, and the computer is required to learn a particular method of connecting inputs and results, For the system the ability to predict the future result of any new entrance, there are many types fall under the supervised learning and depends on the output required from the machine learning system, the most important of these types:

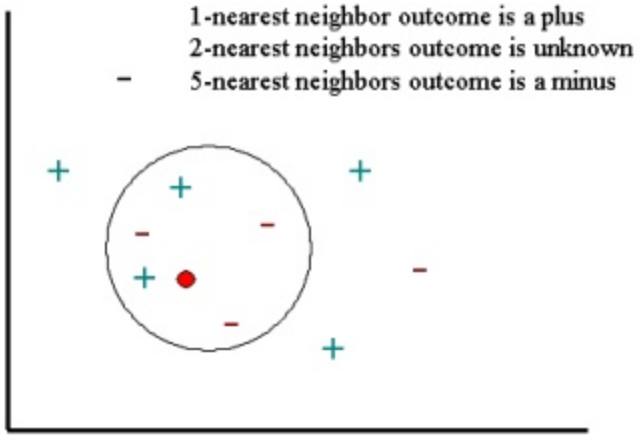
- **Classification:** It is the most commonly used type of machine learning. In this type, income is pre-categorized into two or more types. The goal of the learning process is to produce a model that can classify any new income into a species previously defined.
- **Regression:** This type is similar to classification, but it predicts continuous values rather than separate classes. This type is frequently used to predict stock prices and predict indoor temperature based on weather information, time and sensors.

2- **Unsupervised learning (Descriptive Learning**): In contrast to previous learning, this type of learning is taught by means of income data without any known results. The aim is to develop new models and hidden relationships between data. There is a lot of sub types inside this type:

- **Clustering:** In this type, income is sorted into previously unknown groups. One of the most important applications is learning the movements of the person standing in front of a camera to capture and record movements, later the system can recognize these movements and linked to appropriate reactions, such as games. They are also used to send advertising messages based on purchases and browsing behavior of users, resulting in a classification of groups of people to whom advertisements are sent accordingly.

KNN (k-Nearest Neighbors)

To demonstrate a k-nearest neighbor analysis, let's consider the task of classifying a new object (query point) among a number of known examples. This is shown in the figure below, which depicts the examples (instances) with the plus and minus signs and the query point with a red circle. Our task is to estimate (classify) the outcome of the query point based on a selected number of its nearest neighbors. In other words, we want to know whether the query point can be classified as a plus or a minus sign.



Confusion Matrix: it shows the number of correct and incorrect predictions, i.e. made by through the classification model compare to the actual outcome in the data. The matrix is N*N matrix, where N is the number of target values. Performance of these models is generally evaluated using data in matrix. Below is a 2*2 confusion matrix for two classes, i.e. positive and negative.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Accuracy is the proportion of the total number of predictions that were correct. So here the prediction and actual is true for both the cases in “a” and “d”. So, the total of (a+d) by total number of instances, i.e. a+b+c+d.

Feature selection

Sometimes the output from knn classifier is not determined by the complete set of the input features, instead, it is decided only by a subset of them, where. With sufficient data and time, it is fine to use all the input features, including those irrelevant features, to approximate the underlying function between the input and the output. But in practice, there are two problems which may be evoked by the irrelevant features involved in the learning process.

1- The irrelevant input features will induce greater computational cost.
2- The irrelevant input features may lead to overfitting. For example, in the domain of medical diagnosis, our purpose is to infer the relationship between the symptoms and their corresponding diagnosis. If by mistake we include the patient ID number as one input feature, an over-tuned machine learning process may come to the conclusion that the illness is determined by the ID number.

How I Implement KNN in Java Language

```
//breast cancer
KnnWork nn =new KnnWork();
nn.readFile();
nn.DoKnn();
```

```
After 30 iteration   0.9602380952380952
```

```
//M-of-n.dat
KnnWork2 nn1 =new KnnWork2();
nn1.readFile();
nn1.DoKnn();
```

```
After 30 iteration   0.8096666666666668
```

```
//Exactly2
KnnWork3 nn3 =new KnnWork3();
nn3.readFile();
nn3.DoKnn();
```

```
After 30 iteration    0.7370000000000001
```

Example of combination

```
a
a,b
a,b,c
a,b,c,d
a,b,c,d,e
a,b,c,d,e,f
a,b,c,d,e,f,g
a,b,c,d,e,f,g,h
a,b,c,d,e,f,g,h,i
a,b,c,d,e,f,g,i
a,b,c,d,e,f,h
a,b,c,d,e,f,h,i
a,b,c,d,e,f,i
a,b,c,d,e,g
a,b,c,d,e,g,h
a,b,c,d,e,g,h,i
a,b,c,d,e,g,i
a,b,c,d,e,h
a,b,c,d,e,h,i
a,b,c,d,e,i
a,b,c,d,f
a,b,c,d,f,g
a,b,c,d,f,g,h
a,b,c,d,f,g,h,i
a,b,c,d,f,g,i
a,b,c,d,f,h
a,b,c,d,f,h,i
a,b,c,d,f,i
a,b,c,d,g
a,b,c,d,g,h
a,b,c,d,g,h,i
a,b,c,d,g,i
a,b,c,d,h
a,b,c,d,h,i
```

For Feature Selection I made a brute-force algorithm to get the Best selection of feature for "Breastcancer"data and it take between 4 to 20 minutes from my computer to compute the results and this screen shot from the results. The data is 80% training and 20% testing. And with **5** iterations. And k == 5

Breastcancer

```
final   0.7642857142857142
Fearures is a,b,c,e,f,i with accuracy 0.9785714285714286
```

Feature number 1,2,3,5,6,and 8

```
2
max # of occurences: 3
Class of new instance is: 0
final   0.57
Fearures is a,c,e,g,i,k with accuracy 1.0
```

Take 1 hour

M-of-n

Feature number 1,3,5,7,9 and 11

Exactly2

```
Fearures is a,d,i,j with accuracy 0.835
```

Feature number 1,4,9, and 10

References:

[1] Ethem Alpaydin. 2010. Introduction to Machine Learning (2nd ed.). The MIT Press

[2] https://www.cs.cmu.edu/~kdeng/thesis/feature.pdf

[3] Wikipedia knn and feature selection

[4] Some code from https://github.com/nsadawi/KNN/blob/master/KNN.java