Assignment 2

Question 1: 3 marks
Find an application of Genetic Algorithms in an area of your choice. Summarize the approach, and describe its advantages, disadvantages and results. (300 to 500 words)

I found a research paper published in 2020 on the application of Genetic Algorithm. The application identifies the groundwater pollution source. Genetic Algorithm is one of the approaches that have been widely in limelight and used in lot of areas. All the previous approaches that were tries and experimented with actual contaminated site, failed to yield actual true results. Genetic Algorithm is hoped to be able to solve the problems of method validity verification and then actual site application.

The 'source' data has the features of location, time and amount of pollution source release.

Approach:

Genetic Algorithm (GA) relies on natural selection and genetic laws. The basic principle that the paper talks about for GA is to use the existing relevant features of a contaminated source – this is like the training set for GA. Then GA via the biological evolution by three operators – selection, mutation and crossover – generates a new generation, hence identifying sources. This eliminates the results with lower fitness function values. There are largely four steps in the approach :

1. **Coding and generating initial populations**
   As mentioned – the parameters of the data, the source is location, release intensity and time of the pollution source. The first step includes having these characteristic parameters encoded using binary coding method. These are used as the decision variables and then in GA converted into genotype structure string that form chromosome string.
2. **Fitness function evaluation**
   The convergence speed and accuracy of the solution is determined by the selection of the fitness function. Using some relevant parameters, the groundwater solute transport model for calculation of the pollutant concentrations at different time and locations. Fitness function is what is defined as the difference between the simulated and measured value.

   $$z = \min \sum_{i=1}^{n} [C(X_0, M, T_0) - C(X, T)]^2$$

   C (X0, M, T0) Is the simulated value of pollutant concentration, X0 for time T0, .M is the quality of pollutant leakage.
3. **Population evolution**
   The data with best fitness values are selected first and used to create the next generation. Crossover operators are used on individuals with poor adaptability values. In case the previous two conditions can not be satisfied, that is when the mutation operator is used for generation of the next generation.
4. **Termination condition judgement**

There are various ways of terminating the conditions – like setting a maximum number of next generation/ iterations. After multiple iterations though, the final optimal individual is produced as the result.

Advantages

1. Easy convergence: Precise next generation individuals/evolution are produced as results which leads to efficient further steps to be taken. In cases like this where there is pollution and environment at stake, this is an important feature.
2. High accuracy: This avoids any wastage of time by false locations and results.

Disadvantages

1. Designing the function can be difficult without precise and accurate training/ base data
2. Computationally , this can be expensive and also time consuming.

Reference:

Link to the Research Paper

Han, K., Zuo, R., Ni, P., Xue, Z., Xu, D., Wang, J., & Zhang, D. (2020). Application of a genetic algorithm to groundwater pollution source identification. Application of a Genetic Algorithm to Groundwater Pollution Source Identification, 589. https://doi.org/10.1016/j.jhydrol.2020.125343

Question 2:
Compare Decision Trees and Genetic Algorithms. Look for commonalities or differences of the two approaches. Describe three things they have in common and describe three differences, as related to data mining. A table with an overview of these comparisons is fine.

Decision Trees

Decision Tree is one of the classifiers used in Data Mining or Machine Learning. The data with a series of conditions is split into two nodes creating a tree like structure. There are nodes in the trees, which are the pivot of breaking into two paths. And then there are leaves in the tree, that are the final decisions/results/outcomes.

Genetic Algorithms

GA is another algorithm used in machine learning for solving optimization kinds of problems. GA, as per the name, uses the terms and concepts of genetics like genes, chromosomes, and natural selection to provide solutions to problems.

Similarities between the two:

1. Both the algorithms in machine learning uses data in smaller parts. For instance, Decision tree clearly is good for classification problems. GA on the other hand uses certain said characteristics/features.
2. Both the algorithms uses a training model/base data to predict the next class or value of the target variable.
3. Both the methods can be used for categorical data

Differences between the two:

Decision Trees	Genetic Algorithms
Can not always get the optimize rule	Is specific for optimal cases
Breaks down the data based on characteristics to get outcomes	Uses certain relevant characteristics to create an optimal outcome
Can be used for solving regression and classification problems	Can be used to solve optimization problems
The decisions per node can be depended on the type of data	Uses mechanics of genetics and natural selection

Question 3

Investigate Model evaluation criteria discussed in the book. For two of the criteria (your choice), describe two different scenarios where the criteria works well and should therefore be applied, and two where it does not. The key is to connect the criteria to characteristics of the dataset. (300 to 500 words)

There are two methods discussed in the book under Model Evaluation – Holdout Method and Cross Validation.
Let's talk a out cross-validation.

Scenario when criteria should be applied

1. Cross validation is another predictive algorithm. One scenario where it can be used, in real life application is, replicability. This criteria ensures that it is avoiding fitting a model too closely to the set features of a dataset – which is what we call as overfitting. So, in scenarios where we need to replicate, like replicability in psychology, that is where cross validation can be used, as it mimics the advantages of an independent replication with the same amount of collected data.
2. Scenarios where cross validation can be used in predicting and having less biasness on unseen data are natural hazards. We know that there are existing systems for predicting these, but sometimes, these have faults in them, because there s an underlying biasness. I believe, by using cross validation, and some more systems such as remote sensing and stuff, we can do hazard analysis.

Scenario when criteria should not be applied

1. One criteria when cross fold should not be applied is when there is a time series in the dataset, this is because subsets might have correlation or dependencies with each other. This dependency will not match the requirement of having difference between the training set and the test data set.
2. Cross validation is good for big data sets, but not good when data is too complex. Difference between a big and complex data set is simply dependent on the number of data sets we need to make. For a complex data set, we need too many sets which might not really affect by a higher rate but would definitely be very computational and time consuming.