University of Denver

University College

ICT Program

Cathie Wilson, Instructor

ICT 4410 – Data Warehousing Design

Assignment 3

**NAME: \_Frederick McGovern\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Use the material from the assigned readings, notes, and outside sources to answer questions in this assignment.**

**SECTION 1 (25 points)**

**Directions**: The answer to each of the following questions is either True or False. Answer questions by placing a T or F in the cell immediately to the right of each question.

**Answer ALL questions**

|  |  |  |
| --- | --- | --- |
|  | **QUESTION** | **T/F** |
| 1 | Neural computing involves processing information by means of changing states of networks formed by interconnecting extremely large numbers of simple processing elements that interact with one another by exchanging signals. | T |
| 2 | The purpose of a state function is to consolidate the weights of the various inputs to the neurode into a single value that can be passed to the transfer function for processing. | T |
| 3 | The transmission of signals from one neuron to another in an artificial neural network occurs at the neural synapse via a complex chemical process in which specific substances, called neurotransmitters, are released from one side of the synaptic junction. | F |
| 4 | The basic structure a typical neurode consists of is a set of weighted input connections, a bias input, a state function, a nonlinear transfer function, and an output connection. | T |
| 5 | Fuzzy logic deals with the likelihood that something has a particular property while probability deals with the degree to which a particular property is present in something. | F |
| 6 | The GA’s smallest unit is called a gene. The gene represents the smallest unit of information in the problem domain and can be thought of as the basic building block for a possible solution. | T |
| 7 | The primary purpose of the learning paradigm is to serve as the vehicle by which the summed information is passed on as output. | F |
| 8 | Moving web-based data into a data warehouse rigorous cleansing, conversion- and reduction. | T |
| 9 | When continuous variables are involved, traditional rule-based systems tend to provide more accurate results than fuzzy logic. | F |
| 10 | The spiral development methodology used in data warehouse building, is considered to be a development methodology. | T |

**SECTION 2 (25 points)**

**Directions**: Answer each of the following questions by indicating the letter that corresponds to the correct answer. There is only one right answer for each question. Place answer letter in the cell immediately to the right of each question.

**Answer ALL questions**

|  |  |  |
| --- | --- | --- |
|  | **Question** | **Answer** |
| 1 | The ANN involves an interconnected system of nodes called *\_\_\_\_\_\_\_\_\_*that are associated with one or more *weighted connections* that are equivalent to human neural synapses, inside the memory of a digital computer.  a. transmitters  b. neurides  c. neurodes  d. synapses | C |
| 2 | Which of the following is not true of fuzzy logic?  a. Fuzzy logic focuses on gradation.  b. Fuzzy logic utilizes precise distinction.  c. Fuzzy logic allows for the partial description of a rule.  d. None of the above. | B |
| 3 | The \_\_\_\_\_\_\_\_\_\_ in an ANN is where the processing and transformation of the input signal takes place and where the type of output signal is determined.  a. hidden layer  b. output layer  c. input layer  d. neural layer | A |
| 4 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_ is a method of reasoning that allows for the partial description of a rule. Combining this approach to reasoning with the realm of digital processors has results in a class of computer applications that can “learn” from their mistakes and can “understand” the vagaries commonly found in human thought.  a. Backward propagation  b. Forward propagation  c. Fuzzy logic  d. Neural computing | C |
| 5 | Which of the following is a disadvantage of fuzzy systems?  a. Fuzzy systems decrease system autonomy.  b. Objects in a fuzzy system can belong to multiple sets.  c. Fuzzy systems lack memory.  d. All of the above. | C |
| 6 | An orderly hierarchy of items and item categories that divides each item into a basket analysis is referred to as a:  a. cluster analysis.  b. taxonomy.  c. multi dimensional market basket analysis.  d. None of the above. | B |
| 7 | Which of the following is not a benefit derived from neural computing?  a. Reduced need for experts.  b. Highly verifiable, especially for complex problems.  c. Allows for generalization from specific information content.  d. All of the above are benefits. | B |
| 8 | Which of the following is not an advantage of market basket analysis?  a. Identifying which products sell together can help manage inventory.  b. It is more preferable to marketers to market to existing customers.  c. Market basket analysis can sometimes produce results that are due to prior marketing campaigns.  d. All of the above are advantages. | C |
| 9 | Which of the following is a limitation of data mining?  a. Identification of missing data  b. Data noise  c. Missing values  d. All of the above. | D |
| 10 | Which of the following factors could be used in determining whether to store a document in in an unstructured warehouse or a reference to the location of such a document :  a. The number of relevant documents.  b. The size of the documents.  c. Accessibility of the documents if not stored in the data warehouse.  d. All of the above. | D |

**SECTION C (50 points)**

**Directions:** Provide complete answers to ALL the following questions based on research and investigation of the topic. Use in-text citations to references and properly cite all references in CMS format.

1. *Explain the similarities and the differences between fuzziness and probability.*

As Marakas points out, it is easy to confuse the concepts of probability and fuzzy logic. There are in fact similarities. He points out that both use the same “continuous number range” and that both fuzziness and probability assign the representation of False to a value of 0.0 and the representation of True to a value of 1.0.

Utilizing the statement “Dan is very tall”, Marakas points out that in probability, one assigns a value to the statement. For example, the probability of Dan being tall is 80%. With fuzzy logic the interpretation would be “Dan’s degree in membership within the set of tall people is .80” (Marakas pg 128).

In summary, probability measure the likelihood that Dan is in fact tall (ie a particular property). Probability is making an estimation in the existence of a property. Fuzzy Logic assumes the property exists and provides a strength value in how it relates to other members in the set. Per Marakas, “With probability, we can only determine membership in a set, but with fuzzy logic, we can determine where on the continuum of membership a particular set member resides” (Marakas pg 129).

1. *List and discuss at least three limitations of neural computing*

Marakas lists numerous limitations of neural computing in table 4.3 on pg 138. The following limitations were listed on this table:

* ANN’s cannot “explain” their inferences.
* The “black box” nature of ANN’s make accountability and reliability issues difficult.
* The repetitive training process is often time consuming.
* Highly skilled machine learning analysts and designers are still a scarce resource.
* Neural computing technologies generally push the limits of existing hardware.
* ANN’s require that a certain amount of faith be imparted to the final output.

The first two points and the final point are related. There is a comprehensive statistical algorithm at work where the weighted sum is passed in as an input. The learning by the ANN is incremental and there is limited software on the market that allows for the user to watch the progress of the ANN and see the decisions that were made at each neurode.

The third point refers to the fact that the ANN needs to be trained and with each iteration and inaccurate guess an adjustment to the input weight is done. As Marakas points out, the adjustment is a “complex mathematical procedure”. He continues stating “This process continues for each hidden layer in the network until all neurodes have been subjected to the sensitivity aanalysis and the net is ready for another trial” (Marakas pg 135). With the initial connection weights set at random values, it takes a lot of time before we start having a high degree of accuracy in predicting the output.

1. *What are some of the limitations to market basket analysis?*

The main limitation in a market basket analysis is that a large number of transactions must be available for any result to be meaningful. Additionally, all products need to occur with similar frequency or else the results will be skewed resulting in inaccurate or unsubstantiated data.

Another issue is that previous market basket campaigns can influence the results from a new market basket analysis. Marakas points out that if pizza coupons are provided with frozen pizza due to a previous analysis, there will most likely be a strong relationship between the two items in the current market basket analysis. This can actually cause relationships that actually exist to be hidden by the existing promotion. This can be a lost opportunity as the real relationship (such as beer and pizza) might over better revenue that that of the existing campaign (soda and pizza) (Marakas pg 89).

1. *Discuss some of the characteristics of web-based data as they relate to data warehousing. How do we deal with the very granular nature of web-based data?*

Web based data comes from the collection of user activity into a log. This data is typically referred to as clickstream data. This data would include such things as what a user has looked at, what a user thought of purchasing and what was actually purchased. Using this data, an organization can also determine what products were not looked at. This knowledge can help to driver future advertising and product creation.

Clickstream data is extremely granular in nature and therefore needs to be processed before entering the data warehouse. Software, called a Granularity Manager, eliminates extraneous data, edits bad data, combines similar data, summarizes and aggregates data and creates a singular record from multiple clickstream data logs. According to Inman, approximately 90% of raw clickstream data is either discarded or summarized by the Granularity Manager before being passed into the data warehouse (Inman pg 291).

References

Inmon, William H. *Building the Data Warehouse*. 4th ed. Indianapolis, IN: Wiley Pub., 2005.

Marakas, George M. *Modern Data Warehousing, Mining, and Visualization: Core Concepts*. Upper Saddle River, NJ: Prentice Hall, 2003.