Predicting Fertility in Men

PROGRAMMER:	Andrew Ybarra
COURSE:	CSCI 4391: Intro to Machine Learning
DATE:	4/11/2018
PROGRAMMING ASSIGMENT:	Term Project
ENVIRONMENT:	Java 1.8 running on NetBeans
OBJECTIVE:	Implement classification or regression on data set
SCOPE:	Due April 24, 2018
LIMITATIONS:	None
INPUT:	Inputs are according to the menu/instructions
PRECONDITIONS:	Data set already hard coded.

Introduction:

Male infertility rates have been in an increase for the past twenty years. According to WebMD, studies show most test performing infertility are split equally between women and men. The data set collected online consisted of attributes that would help predict abnormal sperm morphology. In other words, the result would be able to predict if a male would be placed into the category of having abnormal sperm morphology or not having abnormal sperm morphology. The results do not confirm infertility in men but do help predict reasons behind infertility rates in men. Many environmental and biological reasons have been accused of contributing to abnormal sperm morphology, however, the data set selects only a few to help predict our classification.

Methods and Data Set:

The data set taken from the University of California, Irvine's contains 9 attributes and 1 output. The data set below contains the following attribute information taken exactly from the archives:

Season in which the analysis was performed. 1) winter, 2) spring, 3) Summer, 4) fall. (-1, -0.33, 0.33, 1)

Age at the time of analysis. 18-36 (0, 1)

Childish diseases (ie , chicken pox, measles, mumps, polio) 1) yes, 2) no. (0, 1)

Accident or serious trauma 1) yes, 2) no. (0, 1)

Surgical intervention 1) yes, 2) no. (0, 1)

High fevers in the last year 1) less than three months ago, 2) more than three months ago, 3) no. (-1, 0, 1)

Frequency of alcohol consumption 1) several times a day, 2) every day, 3) several times a week, 4) once a week, 5) hardly ever or never (0, 1)

Smoking habit 1) never, 2) occasional 3) daily. (-1, 0, 1)

Number of hours spent sitting per day ene-16 (0, 1)

Output: Diagnosis normal (N), altered (O)

When trying to approach this dataset, testing usually classification seemed reasonable with only one output and multiple inputs. To test the data set, only 20 samples were taken and used to create the weights.

There are no special packages needed to install. Just java running in the latest update (1.8)

In order to operate the program, run the porgram and select an option. Before the program presents to the screen any options for selection, all computations are done and the weights are set. The weights are all initialized to 1 and filled with each iteration of the update to the weights. A menu will appear to the screen as shown below:

MENU

Input Patient Info [1]
Patient Lookup [2]
Accuracy Results [3]

EXIT [0]

To input information and fill your own results, enter 1 to select Input Patient Info. Here questions will appear one by one and ask for you to enter the corresponding answer. All answers are generic and are computed to their true values according to the data set. After all information has been received the program enters the weights and the individuals values and determines whether you will be given a normal diagnosis or altered diagnosis. (Note: The output used in the data set was presented as a ‘N’ for normal and ‘O’ for altered. When the program finishes computing it will present a 1 or a 0 for normal or altered, respectfully. The program then shows the approximation for conceiving with a healthy women based on their results. After the result is given the user will be taken back to the menu.

To compare individual result’s you may look up patient lookup (option 2) which will allow you to type in the patient number (the location in the array). It will then present the output the

algorithm gives to the patient and also the true result for that patient. You will then be taken back to the menu

Option 3 presents you with the data set being compared to the last 30 patients in the dataset. For this subset, there is a .90 accuracy at predicting the correct case when given the inputs. The program then shows the last 30 patients and their predicted and true results. Before the program returns the user is presented with the equation developed from the weights.

Analysis:

At first when computing the dataset, the overall accuracy was extremely low. I soon realized this was because I was taking a small sample from the first twenty or so samples that didn't have much variance between them. I then searched for the best in sample size and subset which the following(column 1 – 9 are inputs and column 10 is the classification):

1	0.67	0	0	1	0	0.8	-1	0.25	1
1	0.75	1	0	0	0	0.6	0	0.25	1
1	0.67	1	1	0	0	0.8	-1	0.25	1
1	0.69	1	0	1	-1	1	-1	0.44	0
1	0.56	1	0	1	0	1	-1	0.63	1
1	0.67	1	0	0	0	1	-1	0.25	1
1	0.67	1	0	1	0	0.6	-1	0.38	0
1	0.78	1	1	0	1	0.6	-1	0.38	0
1	0.58	0	0	1	0	1	-1	0.19	1
1	0.67	0	0	1	0	0.6	0	0.5	0
1	0.61	1	0	1	0	1	-1	0.63	1
1	0.56	1	0	0	0	1	-1	0.44	1
1	0.64	0	0	0	0	1	-1	0.63	1
1	0.58	1	1	1	0	0.8	0	0.44	1
1	0.56	1	1	1	0	1	-1	0.63	1
-1	0.78	1	1	0	1	0.6	-1	0.38	1
-1	0.78	1	0	1	0	1	-1	0.25	1
-1	0.56	1	0	1	0	1	-1	0.63	1
-1	0.67	0	0	1	0	0.6	0	0.5	0
-1	0.69	1	0	0	0	1	-1	0.31	1

The subset was taken from the dataset given and was used due to it containing multiple occurrences of 1's and 0's for outputs. The data also has lows and highs for most of the inputs and ended up resulting in the best output.

The amount of iterations performed varied as I tested the sample size but was eventually assigned to seventy.

The sample inputs I tested on after my weights were determined were the last thirty samples in the dataset. Those samples are listed below (column 1 – 9 are inputs and column 10 is the classification):

-0.33	0.5	1	1	0	-1	0.8	0	0.88	0
0.33	0.69	1	0	0	1	1	-1	0.31	1
1	0.56	1	0	0	1	0.6	0	0.5	1
-1	0.5	1	0	0	1	0.8	-1	0.44	1
-1	0.53	1	0	0	1	0.8	-1	0.63	1
-1	0.78	1	0	1	1	1	1	0.25	1
-1	0.75	1	0	1	1	0.6	0	0.56	1
-1	0.72	1	1	1	1	0.8	-1	0.19	1
-1	0.53	1	1	0	1	0.8	-1	0.38	1
-1	1	1	0	1	1	0.6	0	0.25	1
-0.33	0.92	1	1	0	1	1	-1	0.63	1
-1	0.81	1	1	1	1	0.8	0	0.19	1
-0.33	0.92	1	0	0	1	0.6	-1	0.19	1
-0.33	0.86	1	1	1	1	1	-1	0.25	1
-0.33	0.78	1	0	0	1	1	1	0.06	0
-0.33	0.89	1	1	0	0	0.6	1	0.31	1
-0.33	0.75	1	1	1	0	0.6	1	0.25	1
-0.33	0.75	1	1	1	1	0.8	1	0.25	1
-0.33	0.83	1	1	1	0	1	-1	0.31	1
-0.33	0.81	1	1	1	0	1	1	0.38	1
-0.33	0.81	1	1	1	1	0.8	-1	0.38	1
0.33	0.78	1	0	0	0	1	1	0.06	1
0.33	0.75	1	1	0	0	0.8	-1	0.38	1
0.33	0.75	1	0	1	0	0.8	-1	0.44	0
1	0.58	1	0	0	0	0.6	1	0.5	1
-1	0.67	1	0	0	0	1	-1	0.5	1
-1	0.61	1	0	0	0	0.8	0	0.5	1
-1	0.67	1	1	1	0	1	-1	0.31	1
-1	0.64	1	0	1	0	1	0	0.19	1
-1	0.69	0	1	1	0	0.6	-1	0.19	1

The sample was spread out when it came to the types of inputs and was a good subset to be used to test on. The results of the analysis resulting in predicting the correct class 90% of the time.

One thing to be mentioned about the main dataset was the fact that few patients resulting in having a classification of a 0. This made the set hard to work with and was why I selected smaller sample sizes in order to best represent an even set. When testing with a smaller subset that did not contain many classifications of 0, all inputs were classified as a 1 and vice versa when not many iterations where done the amount of 0's were overwhelming compared to the dataset.

According to an article published by Healthline and another published by BabyCenter, the average rate for sperm morphology low in men but creates a big impact. On average a normal count for fertile sperm found in men is 14% while being considered to have abnormal sperm morphology will result in 5% - 10% of sperm being considered fertile. This results in taking twice the amount of time to conceive since on average half of the sperm count is able to fertilize. With data found from BabyCenter I was able to compare on average the amount of time taken if a male was placed into the 0 or 1 classification the their respected timeline for conceiving given that the women is healthy.

Conclusion and Challenges:

The biggest challenge for working on this project was the dataset. The dataset was taken from the archives presented on UC Irvine's website but was in actual a subset of a much larger dataset from another experiment. The downside of this was the outputs remained the same while the number of attributes contributing to those outputs where extremely minimized. The original data set contained thirty-four attributes whereas the data set I took and worked with only contained nine.

The attribute information was poorly constructed and ambiguous. For some of the attributes, the questions were hard to determine exactly what kind of relation this was to the data set. For example, accident or serious trauma had no indication to what was involved. There was no way to tell what a correct answer could pertain so it is left up to the digression of the user. Along with that, the report that goes with the data set did not go into detail about some of the attributes.

Although this could be put aside since the classification deals with numerical data, the assigned values to these data that were given by the owners of the dataset were poorly set as well. For questions with five answers there were only two numerical representations for these answers. In example, the question for Frequency of alcohol consumption contained five answer choices which only had an output of (0,1). While observing the data I found data that were decimals therefore I assumed the just took the average in respect to the answer choice. Moreover, some positive attributes had positive values as well as negative attributes having positive values. There seemed to be no order.

I struggled with trying to collect the data and analyze it. It took many modifications and testing to see what numbers worked best such as sample size and iterations for testing. In the end I was able to at least have the concept down and was overall satisfied with the progress made in the project.

Bibliography:

"Infertility: It's Not My Fault." *WebMD*, WebMD, 3 Apr. 2000, www.webmd.com/men/features/infertility-its-not-my-fault.

Gil, David, et al. "Expert Systems with Applications." *Schematic Scholar*, pdfs.semanticscholar.org/876d/413220f24d564b6d99048a994e1e007cf5b2.pdf.

Gil, David, and Jose Luis Girela. "Fertility Data Set." *UCI Machine Learning Repository: Fertility Data Set*, archive.ics.uci.edu/ml/datasets/Fertility.

"Sperm Morphology: What Is It and How Does It Affect Fertility?" *Healthline*, Healthline Media, www.healthline.com/health/sperm-morphology.

Kelmon, Jessica. "How Long Does It Take to Get Pregnant?" *BabyCenter*, 3 Apr. 2018, www.babycenter.com/how-long-does-it-take-to-get-pregnant.