**Mini-Project Number 1.** **What Breast cancer will recur?**

Given Dec 3; due Dec. 24 (i.e.two weeks). (There will probably be a second project that I will assign around Dec 24 or Jan 1. (So don’t delay onthis.)

If submitted between Dec 24 and Dec 31 (-5 points). Between Jan 1 and Jan 7 (-15 points). Jan 7 FINAL deadline. (So submit on time by Dec 24.) If you don’t finish everything by Dec 24 submit a report on what you have done so far.

------------------------------------------------------------------------------------------------------------------------

**Collaboration:** 1) You should work in teams of two- three. I will try to speak with each member of the team. It is expected that all members of the team know **all** parts of the project **completely** since you are presenting it as your work.

2) **External Collaboration**. You may freely discuss the project with other members of the ***class***. I recommend you even arrange a zoom meeting between yourselves on e.g. Sunday evenings. You should not share codes, but you can ask about problems and hear others suggestions. **Of course you can not share *code* with anyone outside your team.** If you find help on the web or the internet, **\*YOU MUST REFERENCE IT\***. If a particular student helped you, reference him and explain what help you received. It is OK to get help from other students on how to download and us

*I wish to emphasize to you that in academia, one of the* ***worst*** *sins is to represent someone else's work as your own.* ***Another cardinal principle*** *is to always report truthfully even if results of an experiment aren’t as good as you hoped. Science only proceeds when people can trust the reports. (There have been full professors at famous universities who lost their reputations and even their positions by violating these principles.)*

---------------------------------------------------------------------------------------------------------------------------

This work will require you to download a "real" (i.e. not toy) data set relating to Breast Cancer recurrence prediction. The data set has values taken from a mass extracted from a patient suffering from breast cancer and then indicates whether or not the cancer recurs and how long it takes.

There are approximately 151 patients that did not recur; and 47 that recurred. (This is not a "balanced" set ;i.e. different amounts of data for each of the two cases, and you will have to take that into account when considering the quality of your results. Be sure to explain how you handled this part of the problem.)

Your goal is to implement some of our algorithms and report on how accurate they should be for generalization.

You will implement the following algorithms:

(1) perceptron algorithm . (Your own code. The sets are not linearly separable; but you should indicate how you discover this when running the perceptron.)

(2) **Adaline algorithm** . (Your own code. After training the adaline, you can impose a cut-off to decide which class the data point belongs to. For example, if you try to train the adaline to +1 for Recurring and -1 for Non-recurring then you can use 0 as a cut-off; i.e. a positive result means recurring and negative non-recurring.)

(3) **Backpropagation algorithm**. Here you have a **choice** between your own code and the code of a package. (i) For your own code, use only one hidden level; and experiment with how many neurons there will be here. Report on how it worked for each number of neurons and how much time it took. I suggest you use the same activation function on all neurons; but you can experiment on this if you wish. Be sure to report on the results of all your choices. (ii) Alternatively you can use a package on this problem from a package (like MATLAB or (for those who took the course in deep learning (Tensorflow)). Other packages are also OK probably (I am not sure what is in WEKA, for example.)

The data set can be found at

https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.data

Explanation regarding the dataset:

<https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.names>

note: you will need to filter the dataset using field (i.e. column) “2”:

R = recurrent, N = nonrecurrent

Additionally, you need to: leave out the first column (participant’s ID) and the third column (the recurrence time) from the data when implementing your algorithms.

Training and testing.

1. You should divide the set into two classes; say 66% for training and 33% for testing. (You need to think about the balancing problem here, describe what you did …) Indicate how long you trained and what the results were for the training class and the testing class.
2. Having accomplished 1) if you change the 33% choice 3 times, you obtain 3 results. This is called "cross-validation". You should report the average of these 3 results and the standard deviation.

**Reporting the results: (The quality of the REPORT will affect your grade significantly.)**

A detailed word/pdf report file along with documented code is expected. Besides the code I expect about 6 pages or so (not counting code). ***You will be graded on completeness and clarity of the report as well as the quality of the implementation****.*

In your report you should refer to the following items (i.e. **I would like to see about 4 sections in your report**):

1. Time it took to train the final model (in sec or mins on your PC/Laptop) i.e. after all the cross validation processes.
2. Ease (are the parameters can be adjusted intuitively or not) of parameter search/adaptation (e.g. learning parameter, architecture of feed-forward network (if any)) in order to find the best model for classification
3. Performance on the training set and on the testing set (in %) – explain the results
4. Summary and discussion of process/ problems.

Note: you might meet missing data (features) in some data points - you should think and decide what to do / how to deal with it and of course explain your decision (it doesn’t need to be some “scientific” decision).

The results should be reported clearly. Be sure to include a tables like the following two ways:

* Confusion matrix for each model evaluation

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | True Positive (mean+/-std) | False Negative |
| Actual: no | False Positive | True Negative |

**Any additional charts and graphs will help**. (For example, a graph of quality of results over iterations of training data.)

Include an appendix with your codes (i.e. printed out) so we can read it when checking it.