German University in Cairo
Faculty of Media Engineering and Technology
CSEN 1005: Neural Networks
Project

1. **Goals:**
   - To make students familiar with the new (applied/theoretical) emerging topics in the NN field.
   - To improve research, technical reporting, and presentation abilities of students.

2. **Procedure:** Student groups require to implement the pruning algorithm in the project paper (posted to the Course website). The group consists only of **two** students. The project consists of **three** milestones:

   M1: Implementation of the paper's algorithm using any programming language (preferably Matlab or R) and validating the results in the two toy examples of the paper.

   M2: *For milestones M2 and M3, the goal is to find the optimal hidden layer structure that provides the best generalization for each data set.*
   The algorithm is applied to **six** different **classification** problems (can be downloaded from the UCI ML Database (`archive.ics.uci.edu/ml/datasets.html`): Iris Plants dataset, Wisconsin-breast-cancer dataset (cancer), Hepatitis Domain dataset (hepatitis), Pima Indians Diabetes dataset (diabetes), Ionosphere dataset (ionosphere), and Wave form dataset (wave). The Table below shows how to use each data set in training and testing.

   Properties of six real datasets

Properties	Datasets					
	Iris	Cancer	Hepatitis	Diabetes	Ionosphere	Wave
# of classes	3	2	2	2	2	3
# of examples	150	699	155	768	351	5000
# of training examples	75	350	81	384	176	2700
# of testing examples	75	349	74	384	175	2300
# of attributes	4	9	19	8	34	40

   M3: The algorithm is then applied to **three prediction** pronblems:

   1- The Mackey-Glass time series prediction benchmark problem: The Mackey-Glass time series is generated from the following delay differential equation:

   $$\frac{dx(t)}{dt} = \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} - 0.1x(t)$$

   where $\tau = 17$ and $x(0) = 1.2$. The fourth-order method is used to generate 1000 data points ranging from $t = 118$ to 1117 so that the task involves predicting the value of $x(t+6)$ from the input vector $[x(t-18)\ x(t-12)\ x(t-6)\ x(t)]$ for any t. Therefore, the input–output data pairs for prediction are $[x(t-18)\ \ x(t-12)\ \ x(t-6)\ \ x(t); x(t+6)]$ where the first 500 data pairs are used as training data, and the last 500 data pairs are used as testing data. A matlab script for generating the time series can be found at `http://tinyurl.com/o6gtfgg`

2- Sunspots time series (`http://tinyurl.com/mgldqg8`). The objective of this test involves using $[x(t-10)\ x(t-9)\ \cdots x(t)]$ to predict $x(t+1)$, where t represents the year and $x(t)$ represents the number of sunspots in year t. The number of sunspots from 1700 to 1920 should be used as the training data, and the number of sunspots from 1921 to 1955 is used as the testing data.

3- Vehicle count (`http://tinyurl.com/o29bqyh`).
The data set was obtained from the hourly vehicle count for the Monash Freeway outside Melbourne in Victoria, Australia, beginning in August 1995. The objective of this experiment involves using $[x(t-15)\ x(t-14)\ldots x(t)]$ to predict $x(t+1)$.

(a) M1 will be evaluated on Nov 6, M2 on Nov 27, M3 and the project report (should follow the standard template of the IEEE Journal/Conference) are supposed to be evaluated on Dec 18, 2014.

3. **Evaluation of course project:**

M1 (Nov 6)	25 %
M2 (Nov 27)	20 %
M3 (Dec 18)	20 %
Quality of report (language, format appearance and analysis)	35 %