# Neural Network. Homework #6

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Please include (1) Rmd file; (2) Output from Rmd with answers to all the questions asked

# Due Tuesday at 9pm.

#### Neural Network with the Titanic Data – Classification Task

The Titanic.csv data we will use for our homework is taken from the Kaggle competition site (<https://www.kaggle.com/c/titanic>) where it was called the train.csv. **We shall treat this dataset as our entire data** because we do not know the survival status in the Kaggle test.csv data. Our Titanic data has 891 passengers and 12 variables:

* *PassengerId*: Passenger ID: 1- 891
* *Survived*: A binary variable indicating whether the passenger survived or not (0 = No; 1 = Yes); this is our response variable
* *Pclass*: Passenger class (1 = 1st; 2 = 2nd; 3 = 3rd)
* *Name*: A field rich in information as it contains title and family names
* *Sex*: male/female
* *Age*: Age, a significant portion of values are missing
* *SibSp*: Number of siblings/spouses aboard
* *Parch*: Number of parents/children aboard
* *Ticket*: Ticket number.
* *Fare*: Passenger fare (British Pound).
* *Cabin*: Cabin number
* *Embarked*: Port of embarkation (C = *Cherbourg*; Q = *Queenstown*; S = *Southampton*)

First, one must clean the data and decide which variables to exclude from our analysis. My recommendation is that we exclude *PassengerId, Name*, *Ticket*, and *Cabin* in the ensuing analysis. Next, please note that *Age* has many missing values – my suggestion is to delete those with missing values. Now after the data cleaning step, your task is to split the data randomly into training (75%) and testing (25%), first build the best neural network models to predict passenger survival using the training data, and then use these models to predict whether each passenger in the testing data survived or not. Please use the *neuralnet* package in R to build the various neural network classifiers.

**Note:** For this data set, we shall NOT perform data standardization (normalization) – just to make it easier for you. In the future, for your real world application, please do consider normalization for all machine learning methods, including the neural network.

Please review the following websites for related methods and concepts:

1. **neuralnet**:

<https://journal.r-project.org/archive/2010/RJ-2010-006/RJ-2010-006.pdf>

<https://www.geeksforgeeks.org/building-a-simple-neural-network-in-r-programming/>

2. **logistic regression** – a review, as we shall make comparisons:

<http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/>

1. For the entire dataset, please perform the data cleaning as instructed before; namely, exclude the variables *Name*, *Ticket*, and *Cabin* and delete missing values in the variable *Age*. Please report how many passengers are left after this step. Then please use the random seed 123 to divide the cleaned data into 75% training and 25% testing.

1. Please first build the best classifier to predict passenger survival using the training data and the Perceptron model with (i) no hidden layer, (ii) the default loss function of “sse”, and (iii) the default activation function of “logistic”. Please plot the perceptron model obtained using the training data. Please compute the Confusion matrix and report the sensitivity (that is, a passenger who survived is predicted to have survived), specificity (that is, a passenger who did not survive is predicted to not have survived), and the overall accuracy using the testing data.
2. Next we will build the best classifier to predict passenger survival using the training data and the Perceptron model with (i) no hidden layer, (ii) the loss function of “ce” (namely, cross-entropy, or the negative log likelihood), and (iii) the default activation function of “logistic”. Please plot the perceptron model obtained using the training data. Please compute the Confusion matrix and report the sensitivity (that is, a passenger who survived is predicted to have survived), specificity (that is, a passenger who did not survive is predicted to not have survived), and the overall accuracy using the testing data.
3. Now we shall build the best classifier to predict passenger survival using the training data and the Logistic Regression model. Please report the fitted logistic regression model obtained using the training data – and compare to the Perceptron models obtained in the plots of Question 2 and Question 3. Which Perceptron model better resembles the logistic regression model, and why? Please compute the Confusion matrix and report the sensitivity (that is, a passenger who survived is predicted to have survived), specificity (that is, a passenger who did not survive is predicted to not have survived), and the overall accuracy using the testing data.
4. Now we shall build the best classifier to predict passenger survival using the training data and the Perceptron model with (i) one hidden layer with 3 neurons, (ii) the default loss function of “sse”, and (iii) the default activation function of “logistic”. Please plot the perceptron model obtained using the training data. Please compute the Confusion matrix and report the sensitivity (that is, a passenger who survived is predicted to have survived), specificity (that is, a passenger who did not survive is predicted to not have survived), and the overall accuracy using the testing data. Please compare the performance in test data to that of Question 2.
5. Next we shall build the best classifier to predict passenger survival using the training data and the Perceptron model with (i) one hidden layer with 3 neurons, (ii) the loss function of “ce” (namely, cross-entropy, or the negative log likelihood), and (iii) the default activation function of “logistic”. Please plot the perceptron model obtained using the training data. Please compute the Confusion matrix and report the sensitivity (that is, a passenger who survived is predicted to have survived), specificity (that is, a passenger who did not survive is predicted to not have survived), and the overall accuracy using the testing data. Please compare the performance in test data to that of Question 3.

