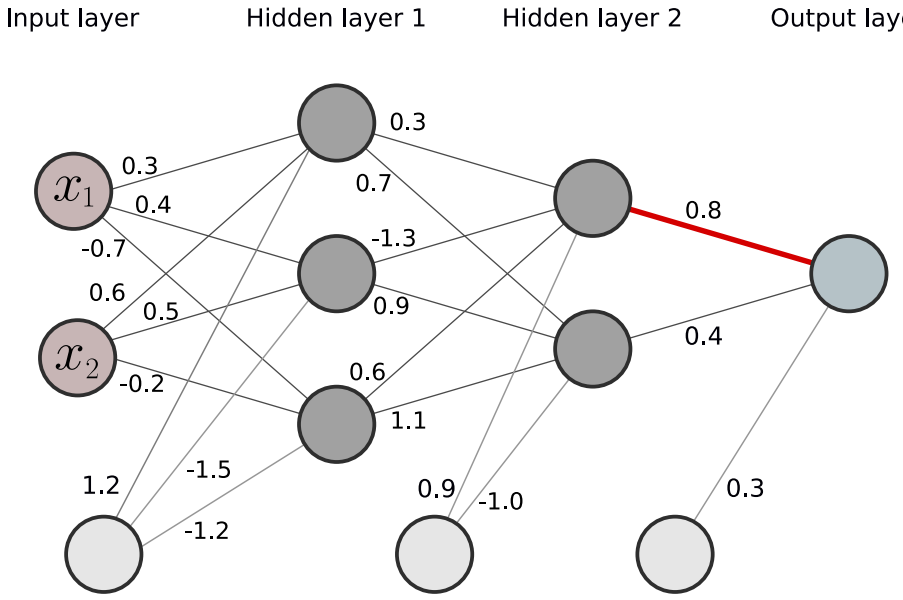
# STAT40970 – Machine Learning & A.I. (Online)

**Assignment 1**

**Deadline - Tuesday 4th March 2025 at 17:00**

## Exercise 1

The figure below shows a multiple layer neural network deployed to predict the outcome of a 2-class categorical target variable $y$. The activation function in both hidden layers is the rectified linear unit (ReLU) and the output layer uses the sigmoid output activation function. The weight and bias parameters are shown along the edges in the network. Because it is a binary classification problem, only the node corresponding to $y = 1$ is reported in the output layer.




1. Consider an input data point as stored in the vector $\mathbf{x} = (x_1, x_2)$. Perform a forward propagation calculation through the network for the input observation vector $\mathbf{x} = (-1, 1)$ *(20 marks)*

2. The target variable class label for the current input vector $\mathbf{x}$ is $y = 1$. Calculate the loss associated with this training instance using an appropriate loss function. *(10 marks)*

3. Denote with $w$ the weight corresponding to the red edge in the network. Compute the value of the gradient of the loss $E = -y \log o - (1 - y) \log(1 - o)$ with respect to this weight $w$. Note that $o$ denotes the output value of the network. Use the quantities computed in (a) for the calculations. *(20 marks)*

## Exercise 2 – Data analysis

The file `data_assignment_1_bats.RData` contains data concerning numerical features characterizing the acoustic properties of calls of a large sample of bats species in Mexico. The task is to predict the bat family type using the acoustic numerical features, with the purpose of a better monitoring of biodiversity change and a better characterization of the species living in a region. The dataset is a subset of a large database, more information about the aspects of the data is available here: `https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12556`.

The target variable `Family` is a categorical variable indicating the following 4 bat families: `emba` (Emballonuridae), `morm` (Mormoopidae), `phyl` (Phyllostomidae), and `vesp` (Vespertilionidae).

The file includes the data matrix `data_bats`, containing the observations on the target variable `Family` and 72 numerical input variables derived from the acoustic signal of a bat's call.

Consider the following classifiers in order to predict the family a bat call belongs to:

- $C_1$ – Multinomial logistic regression classifier + PCA dimension reduction with $Q$ coordinate vectors.
- $C_2$ – Neural network with single hidden layer.

**1.** Implement an appropriate cross-validation procedure for comparing and tuning the two classifiers, and select the best one. In doing so:

- For $C_1$, consider a range of values of the number of coordinate vectors $Q$ such that the proportion of cumulative explained variance is approximately around the range 80% - 90%.
- For $C_2$, tune appropriately the number of hidden units considering a sensible range. No need to tune/consider different activation functions, one is sufficient, but you need to justify briefly your choice.
- Discuss and justify clearly and concisely the various decisions taken in all stages of the process.

*(40 marks)*

**2.** Evaluate the generalized predictive performance of the selected model by assessing its accuracy on some appropriately prepared test data. Comment concisely on the specific ability of the model in predicting calls from the family `emba` (Emballonuridae).

*(10 marks)*

## Submission rules and instructions

- Write a short and tidy report and submit it as a single pdf file (approximately max 8 pages, code excluded).
- Include the R code used for the data analysis in the report. The report can be produced using R Markdown (or similar tool), with the code included in the main text or as an appendix. **The code must be working and the analysis must be reproducible in all parts**.
- In general, for full marks you **must explain** concisely and clearly **all reasoning**, as well as **show all steps** and **computations** in your answers. Correct answers alone will not achieve full marks.
- For the data analysis task, the use of the `caret` package (or packages with similar functionalities) is not allowed. You can use package `nnet` to implement the multinomial logistic regression model.
- For the data analysis task, submitting only code without any output, commentary, or discussion will not receive any marks.
- Multiple submissions before deadline are allowed and only the latest one will be considered for marking.
- **Submission after deadline will incur in penalization as UCD rules**. See "Module details" document under the "General information" tab on Brightspace.
- **Plagiarism is strictly prohibited and will result in severe penalties**. See "Module details" document and "SMS academic integrity protocol" under the "General information" tab on Brightspace to review which actions constitute plagiarism and further information.
- **Any instance of plagiarism** will result in a **zero grade** in this assessment component **for all students involved**.
- By submitting this assignment, you confirm that you have read and understood the regulations and policies regarding assessment and plagiarism outlined here, in the "Module Details" document, the "SMS Academic Integrity Protocol", and related documents. You also agree to abide by all the regulations stated in these documents.