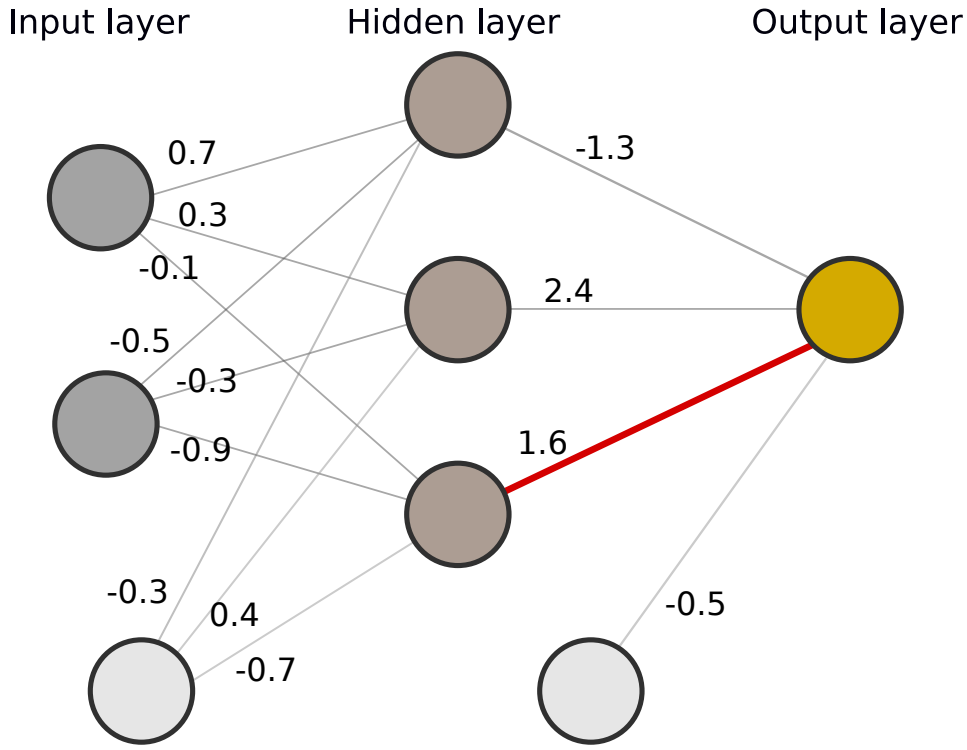
STAT40970 – Machine Learning & A.I. (Online)

Assignment 1

Deadline - Tuesday 12th March at 5pm

Exercise 1

The figure below shows a (single hidden layer neural) network deployed to predict the outcome of a numerical continuous target variable y. The activation function in the first hidden layer is the hyperbolic tangent function. The output layer uses an identity output function. The weight and bias parameters are shown along the edges in the network.



1. Perform a forward propagation calculation through the network for the input observation vector $\mathbf{x} = (2.1, 0.7)$. What is the value of the output unit o corresponding to this input vector? *(20 marks)*

2. The value of the target variable associated to the above input vector $\mathbf{x}$ is $y = -1.1$. Denote with w_{23} the weight corresponding to the red edge in the network. Compute the value of the gradient of the loss $E = \frac{1}{2}(y - o)^2$ with respect to this weight w_{23}. Note that o denotes the output value of the network. Show clearly all steps in your calculations. *(15 marks)*

3. Show how the output o for the input vector $\mathbf{x} = (2.1, 0.7)$ would change if the ReLU activation function were used in place of the hyperbolic tangent function. Justify your answer with minimal calculations. *(5 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.

(50 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 report and submit it as a single pdf file (approximately max 10 pages, code excluded).
- Include the R code used for the data analysis in the report. The report can be produced using R Markdown, 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**.
- 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).
- 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 does not provide any marks.
- **Plagiarism is strictly prohibited** and it will incur in severe penalties (see "Module details" document and "Information materials" tab).