Machine learning II
Master Data Science
Winter Semester 2023/24 **Prof. Tim Downie**

Workshop 12
Fitting classification neural networks in R using Torch.

The subject of this week's lecture is neural networks for time series data. The practical aspects of this are not covered in this year's course, instead this week's worksheet follows on from Workshop 9, by focussing on classification MLPs using Torch in R.

If you are interested in NNs for time series data, there is a tutorial (which uses Keras) at Time Series Prediction in R Keras to get you started. There is also a Lab-Section 10.9.6 in James et.al. starting page 454 coverin recurrent neural networks. The code to implement this in Torch is available from their resources site: Torch-NN code staring at line 550

Exercise 1 Introduction to neural network classification: Star Type
The file `ML2_Wkshp12_NN_Classification.R` gives code template for all three exercises today.

This exercise follows a well written blog by Will Hipson at the web site
`https://willhipson.netlify.app/post/torch-for-r/torch_stars/`.
The code has been slightly modified in places, eg. you don't need to register with *Kaggle* and activate a Kaggle Token just to download the data, a data file is provided in Moodle. The graphic format for the confusion matrix, is nice but the given code does the same without having to install a new R package.

Open the website, read through the text, while working through the code in parallel.

Exercise 2 Mushroom species

The data set contains many characteristics for different species of mushroom. One variable is `poisonous`, labelled as `p`: poisonous and `e`: edible which we will use as the outcome in the classifier.

The code follows the website `https://blogs.rstudio.com/ai/posts/2020-11-03-torch-tabular/`. It has many similarities as in Exercise 1. Again an embedder is used for the non-numeric variables, but is slightly different as the data are imported as character variables (not factor) and the outcome variable is also character.

The data has been provided for you. The PCA Section at the end of the web page has been omitted.

Exercise 3 The diabetes dataset

In Week 1 we looked at the NHANES dataset for imputing missing data. Today we will use the same data set but use the quick and dirty method of dealing with missing data, i.e. to remove all observations which contain one or more missing value.

Because running a NN on a large data set with `R-torch` is slow, the data is down-sampled using the following method: all the diabetes cases are included in the downsampled data, and all the same number of patients without diabetes are chosen at random.

(a) Again work through the commands in the source code.

(b) Once you have fitted the NN and got the first results, investigate using a different number of nodes in the hidden layer, and multiple hidden layers. Try out other settings in the NN.