1 Homework 1 (due on 09/14/22)

1. The Titanic dataset.

The dataset is available in an R package `titanic`, which already has the data split into a training set and a test set. The training set `titanic_train` contains 891 subjects and 12 variables, in which `Survived` is the outcome variable. The test set `titanic_test` contains 418 subjects and 11 variables without the outcome variable. The R package also contains objects `titanic_gender_class_model` and `titanic_gender_model`. They are probably leftover objects from the package creator's own models. You can ignore them.

Note that the same data in .csv format are available from https://www.kaggle.com/c/titanic/data . Also note that R also has an object `Titanic`, which is a simple data tabulation that is slightly discrepant from the data in the `titanic` package: `Titanic` indicates there were 1316 people, while the `titanic` package contains 1309.

(a) Play with the Titanic dataset to predict the survival status. Do exploratory analyses and clean up the data when necessary. Do all model training and tuning on the training set. Turn in your final prediction model and the predictions on the test set (as a 418 × 2 object or a file with two columns: PassengerId, Survived).

(b) There are many analysis demonstrations of this dataset online, both in R and Python. Do not look at them. Try to get as much as possible yourself. Then look at other people's solutions online (by just googling). Summarize what you have learned from them, and if you did something better than most of online demonstrations, briefly highlight them.

2. Show the following:

(a) For binary outcomes, an ANN model without any hidden layer and with a softmax output layer is equivalent to logistic regression.

(b) For categorical outcomes with more than two outcome categories, an ANN model without any hidden layer and with a softmax output layer is equivalent to multinomial logistic regression.

(c) Simulate a small dataset (with $n = 1000$) with a binary outcome. Fit a logistic regression model and an ANN model without any hidden layer. Compare the results.