Final project: Predicting Survival on the Titanic Using a Neural Network

Dr. Ahmed El Kerim*

Objective

This project aims to create a classification neural network that predicts whether a person will survive or perish based on the Titanic dataset. This will involve data manipulation using the Pandas library and implementing a neural network for classification using a deep learning framework such as TensorFlow or PyTorch.

Dataset

We will use the Titanic dataset. The dataset includes the following columns:

- `PassengerId` - Unique identifier for each passenger
- `Survived` - Survival (0 = No, 1 = Yes)
- `Pclass` - Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- `Name` - Name of the passenger
- `Sex` - Gender of the passenger
- `Age` - Age of the passenger
- `SibSp` - Number of siblings/spouses aboard the Titanic
- `Parch` - Number of parents/children aboard the Titanic
- `Ticket` - Ticket number
- `Fare` - Passenger fare
- `Cabin` - Cabin number
- `Embarked` - Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

*Laboratory of Mathematics in Interaction with Computer Science, CentraleSupélec, Paris-Saclay University, Gif-sur-Yvette, France.

Task 1: Data Loading and Exploration

- Load the dataset using Pandas.
- Display the first few rows of the dataset to understand its structure.

Task 2: Data Preprocessing

- Convert categorical variables (`Sex` and `Embarked`) into numerical values.
- Handle missing values in the `Age`, `Fare`, and `Embarked` columns by filling them with median values.

Task 3: Correlation Study

- Drop columns that are not useful for correlation: `Name`, `Ticket`, `Cabin`.
- Calculate the correlation matrix for the dataset using the Pandas command: `dataframe.corr()`.
- Visualize the correlation matrix.
- Analyze the correlation to identify which variables have the least correlation with the survival status.
- Correlation Coefficient Values:
  - The correlation coefficient ranges from -1 to 1.
  - A coefficient close to 1 indicates a strong positive correlation: as one variable increases, the other also increases.
  - A coefficient close to -1 indicates a strong negative correlation: as one variable increases, the other decreases.
  - A coefficient around 0 indicates no correlation: the variables do not have a linear relationship.

Task 4: Building the Neural Network

- Split the dataset into training and testing sets.
- Define a neural network architecture using a deep learning library such as TensorFlow or PyTorch.
- Train the neural network using the training set.
- Evaluate the neural network using the testing set.

Deliverables

- Python code implementing all sections.
- Documentation describing How to execute the code think about a Readme.
- Presentation slides summarizing the project goals, methodology, results, and potential extensions.
- Optional: Report with detailed results.