Predict which Titanic passengers will survive based on gender, age, passenger class, etc.

Problem

Here is an example of the data. Many of the columns/features are self-explanatory, some are not.

**Pclass**: 1 = upper class, 2 = middle class, 3 = lower class **SibSp**: # of siblings and spouses aboard

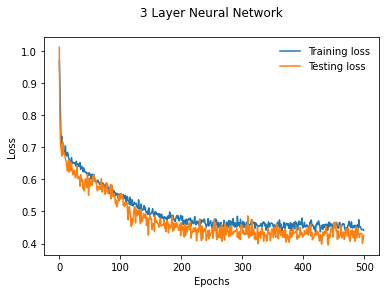
**Parch**: # of parents and children aboard **Embarked**: port the passenger embarked from

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
| 892 | 3 | Kelly, Mr. James | Male | 34.5 | 0 | 0 | 330911 | 7.8292 | Null | C |
| 893 | 3 | Wilkes, Mrs. James | Female | 47 | 1 | 0 | 363272 | 7 | Null | C |
| 894 | 2 | Myles, Mr. Thomas Francis | Male | 62 | 0 | 0 | 240276 | 9.6875 | B45 | S |
| 895 | 3 | Wirz, Mr. Albert | Male | 27 | 0 | 0 | 315154 | 8.6625 | Null | Q |
| 896 | 3 | Hirvonen, Mrs. Alexander | Female | 22 | 1 | 1 | 3101298 | 12.2875 | E31 | S |

Analysis

Data

Plot of Test/Train Loss for NN Model



Results

When testing the linear regression model against a split train data (using Sklearn train\_test\_split()), it had around 80% accuracy. Although, the actual Kaggle score was about 0.65.

When training the NN model, the highest test accuracy was about 83%. The Kaggle score improved to about 0.78.

Approach

Two main models were used: Logistic Regression and a 3-layer Neural Network.

**Preprocessing**

* Missing *Age* and *Fare* values were filled with median values, and missing *Embarked* values were filled with the mode value.

* *Id, Name,* and *Ticket* columns were removed as those columns are not important to the models.
* A large majority of *Cabin* values were null; thus, that column was removed as well.
* Lastly, the categorial columns: *Sex* and *Embarked*, are converted to dummy variables using one hot encoding.

**Neural Network**

* Uses Relu after first two steps and log\_softmax after last step. Adam optimizer and cross entropy loss functions used.

Linear Regression Kaggle score: 0.65

3-Layer Neural Network score: 0.78

Overall, the NN increased the score by about 0.13 points, which is a fairly significant improvement.

Conclusions

Github: https://TLT034/TitanicNN

Source Code