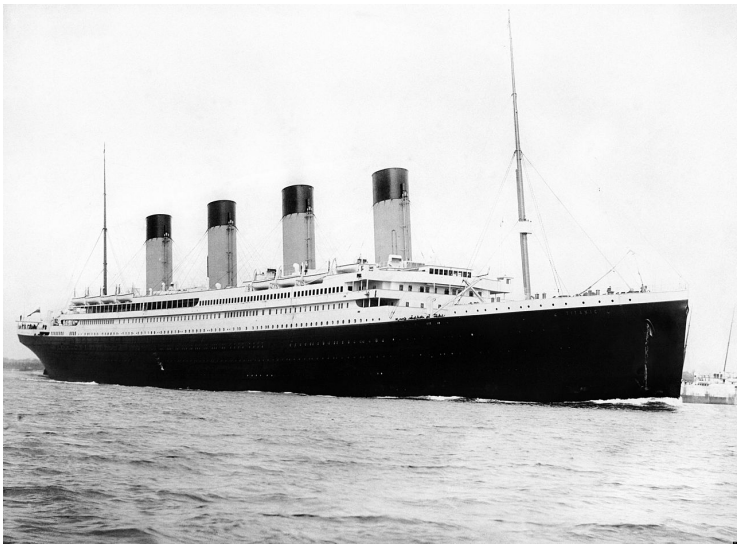
EE4483 Artificial Intelligence and Data Mining

Mini Project (Option 2)

Due: Friday (11:59pm), 13 Nov 2020



RMS Titanic departing Southampton on 10 April 1912.
[Source: https://commons.wikimedia.org/wiki/File:RMS_Titanic_3.jpg]

On 15 April 1912, during her maiden voyage from Southampton to New York, RMS Titanic, the largest commercial ship afloat at that time, collided with an iceberg and sank to the bottom of the Atlantic Ocean. More than 1500 of the 2,224 passengers and crew on board perished in this infamous tragedy. You are required to build a classifier to predict which passengers on board of Titanic would survive the tragedy.

1. **Reading in data**: You will need to load the data from the provided csv files. Two data sets are available, namely the training set ("train.csv") and test set ("test.csv"). The attributes of the data are as follows:
   - PassengerID – No. 1 – 1309 ID
   - Survived – 1=Yes, 0=No (to predict in testing) Label
   - Pclass – Ticket class, 1 = 1st, 2 = 2nd, 3 = 3rd
   - Name – Name of passenger
   - Sex – Male, Female
   - Age – Age in years
   - SibSp – Number of siblings/spouses on board
   - Parch – Number of parents/children on board
   - Ticket – Ticket no.
   - Fare – Passenger fare
   - Cabin – Cabin no.
   - Embarked – Port of embarkation, C = Cherbourg, Q = Queenstown, S = Southampton

   10 attributes

   The "ground truth" for the class of interest "Survived" of each sample is provided in the training set.

2. **Data processing**: You need to convert the raw data into appropriate feature format. For example, you can convert the Pclass into a one-shot vector of length three, as there are 3 types of tickets in total.

3. **Model Selection:** You are going to conduct a classification model to predict which passengers on board of Titanic would survive the tragedy. You can use either **support vector machine** or **neural networks** for this task.

4. **Training**: You will need to try different model parameters to obtain good classification results, e.g., feature dimension, weights, initialization, and learning rate. In the training stage, your algorithm should take as only the input attributes available in the training set, excluding attribute "Survived", which is the output of the classifier.

5. **Prediction:** By setting up the correct learning algorithm, you can predict which passengers on board of Titanic would survive. You need to report the predicted survival of passengers for the testing data in the file ("submission.csv").

Submit your report to **answer the following questions**:

(a) Understand the training and testing datasets that we provided. Make a table to describe the two datasets, including their feature dimension, number of samples, mean and variance of the values in each attribute. (**5% marks**)

(b) Select at least one appropriate model (e.g., neural network, support vector machine, etc.) to build your classifier. Clearly describe the model you use, including the input and output dimensions, structure of the model, loss function(s), training strategy, etc. Include your code as well if you are solving the problem by programming. (**15% marks**)

(c) Discuss how you consider and determine the parameters (e.g., learning rate, etc.) / settings of your model as well as your reasons of doing so. (**15% marks**)

(d) Discuss how you handle attributes with missing values and what attributes are excluded from your classifier(s), and why. (**10% marks**)

(e) Apply the classifier(s) built to the test set ("test.csv") which contains the data of 418 passengers not included in the training set. Submit the "submission.csv" with the results you obtained. (**20% marks**)

(f) How many of these 418 passengers are classified as survivors from the Titanic tragedy? Among these classified survivors, how many of them were (i) female, (ii) below 18 years old, (iii) without any other family members on board? Passengers from which (iv) ticket class and (v) port of embarkment had the least chance of surviving the tragedy? (**10% marks**)

(g) Build one more classifier. Discuss and compare the results obtained from different classifiers. (**10% marks**)

(h) Extend your classification algorithm, or build another network (e.g., CNN) for image classification for **Cifar-10** dataset (https://www.cs.toronto.edu/~kriz/cifar.html). Describe details about the dataset, classification problem that you are solving, and how your algorithm can tackle this problem. Report your results for the testing set of **Cifar-10**. (**15% marks**)

[Hint: an image can be considered as a matrix. If you wish to extend your algorithm for images, you can vectorize an image matrix into a vector, and apply PCA to reduce the vector dimension to fit your network input.]

Notes:

- You can choose any programming language / platform that you like to complete the task.
- If you couldn't obtain any meaningful results or answers to the questions above, you may describe what you have done and attach the relevant working, codes, or screenshots, if available.
- You can work in group with another student who is **in the same TA group**. However, each of you must submit your own report, and the report must include independent answers to all of the questions. If you choose to work in group, clearly specify in your report who is your **project partner** and your **respective contributions** to the project.
- You should clearly cite all the references and sources of information used in your report.
- You are expected to uphold NTU Honour Code.
- Submit your report and the file "submission.csv" with your results to the assigned TA via NTULearn by the **deadline**: Friday (11:59pm), 13 Nov 2020.