Applied Project Proposal

mshapiro[1], poulomipal[1], and mengxiaren[1]

[1]Affiliation not available

September 10, 2019

Project Proposal

CSCI 575: Introduction to Machine Learning

Team Name

Poulomi Pal

Mengxia Ren

Megan Shapiro

Applied Project

Preference: [✓ High] / Low

Topic Area: Correlating Box Office Success to Filmmakers, Producers, and Actors

Project Name: Movie Box Office Predictor

Problem Statement: Every movie that a studio chooses to produce is a gamble. While there are many factors that influence a movie's success at the box office, can these features be used to accurately predict how well a film will perform? If it is possible to calculate a movie's success before it is released, movie studios will produce films that will guarantee higher profits.

Proposed Solution: We will use linear regression to correlate box office profits with features such as popularity of the main actors, popularity of the director, which movie studio it is produced by, and the film's initial production budget.

Supervised learning will allow the machine to generate patterns from features with known numerical quantities from movies released in the past, and by analyzing the problem we proposed, we consider to use deep learning or random forest algorithm to set up our models. To test the learning set, we will use movies that have been released in the last year.

Data: Data provided by The Movie Database (TMDb) https://www.kaggle.com/tmdb/tmdb-movie-metadata

Applied Project

Preference: High / [✓ **Low**]

Topic Area: Image recognition

Project Name: Handwriting Recognition

Problem Statement: Though using a computer to type is popular now, some people still prefer to submit a handwritten document. For example, some students submit their handwritten homework on canvas. However, sometimes it is difficult for instructors to recognize the characters in these handwritten files and takes time to read and grade assignments. It will be useful if we can recognize the characters automatically using machine learning.

Proposed Solution: Instead of recognizing a sentence or a word completely, the project is to reliably recognize images of the 26 English characters (lowercase only), 10 numbers and 4 symbols.

In the project, data will be classified in 39 labels. From the categories listed above, the letter 'o' and the number '0' have been combined to avoid mislabeling. We will use supervised learning (classification) to recognize these characters and symbols. Firstly, we classify the data set, get some training data and some data for validation, and transfer the image data to the vectors. Then we will use the training data to train the model, such as KNN. Finally we will use validation data to verify the accuracy of the model we trained and modify the parameters of the model until we maximize the accuracy rate.

Data: Data provided by `https://www.kaggle.com/vaibhao/handwritten-characters`

Timeline

Work Breakdown Structure (WBS)

1. Import Data in Python
   (a) Install Python environment
   (b) Collect data from data source website and transfer it into documents that can import into python such as text or csv
   (c) Import data and check the integrity and correctness of imported data
2. Define Hyperparameters
   (a) According the problems proposed in the project, set up corresponding models
   (b) Determine features of data to be used in the model
3. Train System from Original Dataset
   (a) Dividing data set to get training data set and validation data set
   (b) Use the training data to train the model and use cross validation to modify some parameters
4. Test System on new Dataset
   (a) Import the new data set to our project
   (b) Apply the trained model and evaluate performance
5. Finalize Results and Write Presentation
   (a) Collect results
   (b) Write Project Progress report
   (c) Based on the report and codes, write slide for presentation

Critical Path The critical path of the project with expected completion dates of each task is:

- 1.3 Import and Verify Data Set - (09/20/19)
- 3.2 Cross-train and Validate Model - (09/28/19)
- 5.2 Write Progress Report - (10/5/19)
- 4.2 Test Model and Evaluate Performance - (10/20/19)
- 5.3 Write Final Presentation - (11/27/19)