University at Buffalo

MGS 662 Optimization Methods for Machine Learning

Project 1

Anna Liu - aliu54@buffalo.edu

Kally Yu – yyyu@buffalo.edu

Objective

In this project, the objective is to perform a machine learning process in R that the system should learn features over multiple modalities. Through the learning process, the system is expected to find the correlation between the labels and the features that it should later be able to identify the labels of the objects given the features. To find the correlation, linear regression will be performed on every model, and the system is expected to calculate the MSE for identifying which feature model provides a better learning method.

Data

The data are given in two groups. The first group is the objects with image and text features, contributing as the independent variable. The second group is the three CSV files provided with a pre-identified label for the same objects given in the first group, meaning the object is labeled as “Tree”, “Animal”, “Mythological” or none. The labels are in the form of 0 and 1, and they will be serving as the dependent variable when performing linear regression.

The data are then processed into three experiment groups: image features, text features, and both image and text features. Within the three experiment groups, the data are broken down further into train data and test data. The system will be learning through different train datasets to identify the labels of the objects and performing evaluation on the test data to see how well it can predict the correct labels of the objects.

Experiment and Observation

Image features and text features are extracted into data points or variables, so the system can process them. During the process of image extraction, the system will read an image as a matrix of pixel values, meaning the size of the image contributes to the number of pixel values. For example, resizing the image to a width and length of 50 will result in fewer pixel values compared to resizing the image to a width and length of 100. However, having more pixel values doesn’t necessarily mean that the system can identify the labels better, as some values could be distracting and unnecessary.

Before the system can identify the labels of the object, it has to go through the learning process. In this case, the learning process is to perform linear regression on the train dataset. Through linear regression, the system should be able to search through data points within the test data and find the correlation between data points and the label.

The learning process is broken down into three categories: the learning of tree labels, animal labels, and mythological labels. With some uncertainty during the identification of labels, there were disputes about the mythological label. Therefore, there are mythological labels 1 and 2. After performing linear regression on all three categories, the system encountered a problem when trying to evaluate its performance. When the independent variable, or predictor, is the extracted pixel data of the image features, the calculated MSE is very large and uncontrollable.

This is an indicator of linear regression might not be the best method for the learning process. Therefore, a decision is made to switch from linear regression to glmnet linear regression for a better-predicted result.

After using glmnet linear regression to simulate the learning process, the system evaluates its performance on the test data. MSE has been used for the evaluation that a smaller value indicates a smaller error. From the calculated MSE for all the experiments, there are only slight differences among the MSEs. For the tree labels, the MSE ranges from 0.222 – 0.236, with the lowest of 0.223 obtained from the experiment group using both image and text features as the predictor. The text feature model contributes the lowest MSE of 0.235 for the animal labels and the lowest MSE of 0.151 for mythological labels 2; while both image and text feature model contribute the lowest MSE of 0.236 for mythological labels 1. The other MSEs can be found in Table 1.

It can be argued that with a large dataset, a small difference in value could be meaningful. However, the calculated errors are similar that there is nearly no difference within the same category. This was not expected because given the pixel values if the system were to identify a tree label, there should be more pixel values of the color green. With the original expectation of significant differences among the MSEs, the obtained MSEs are not helpful when trying to differentiate the performance of the models.

If a decision were to be made based on the MSE, the text feature model and both the text and image feature model perform better compared to the image feature model. It seems that text feature data might be a better learning model for the system to correctly predict the labels.

Table 1

MSE	image	text	image_and_text
TreeLabels	0.235687546	0.223449372	0.222767601
AnimalLabels	0.247942964	0.235387958	0.246281222
MythologicalLabels_1	0.253035975	0.243891416	0.235990194
MythologicalLabels_2	0.15697706	0.150908335	0.154791311

Limitation

Given the dataset, all the dependent variables are in the form of 0 and 1s, which are binary outcomes. Using linear regression to model binary outcomes can sometimes cause the predicted values to be out of range. As mentioned above, performing linear regression can result in a very large MSE when using pixel values as the predictor. It was found that by restricting the number of pixel values by resizing the image, the MSE can be manipulated and become a much smaller number. Therefore, linear regression is not the best fit learning method with the given data. While glmnet does perform better, glmnet filtered the smallest MSE model during the evaluation, meaning the MSE provided is always the smallest MSE within the model. This results in the slight difference among all the MSEs and the inability to differentiate the performance of each model.