Algorithm Projects

Table of Contents

Introduction

In this trimester, an extension to the previous trimester's works have been carried out with three new algorithm projects namely Cancer detection, GBM ranking model and Blood lactate model.

Project team members include
Cancer detection – Shashvat Joshi
Incentive allocation & GBM Ranking – Hyun Dong Kim
O2 (Blood lactate) – Karl Birti

These projects have been completed through Sprint 1 and 2 which involves sufficient research and machine learning trials using python. All projects have been documented with proper evidence, hence, the future students can revise and refer back to what's been carried out thus far.

Sprint 1 projects (Completed)

Incentive allocation model – revision (Completed)

Project team members:
Shashvat
Hyun Dong Kim

An overall revision on last trimester's work on GAM methods & incentive allocation models. Updated documents have been uploaded in the file's section for a review.

O2 Model

Project team members:
Shashvat
Karl

O2 model comparison with competitor's models in the market

After building a mode to predict O2, we compared the performance of the model to that of models built by competitors in the market. Research was conducted to investigate the model structures and features used to predict oxygen uptake.

Results & Conclusion

For cycling, it was found that the best alternate model was constructed by FirstBeat with an accuracy of 92%. Our best model which was the decision tree regressor had an accuracy of 99%. This was tested with new data, and we saw a decrease in accuracy. The model performance was affected by sample size so keeping the same method for building the model but increasing the sample size increased robustness of the model in testing. Other models for predicting VO2 are out there in the market and provide a strong prediction accuracy. In comparison to our current model, competitors have not developed a model as effective as ours as the accuracy of our model is 99% after tuning the machine learning algorithm parameters.

Sprint 2 Projects (Completed)

GBM Ranking (Light GBM) - Completed

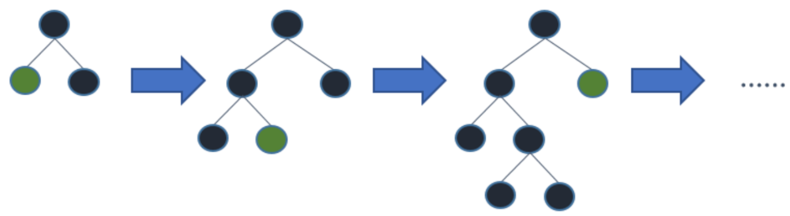
Member: Hyun Dong Kim

Over the past few years, ranking algorithms have been growing in popularity in many industries and companies are finding its merits in their application. The research focuses on lightGBM which is fast, distributed and high-performance gradient boosting tree-based learning model that ca be used in various applications including regression, classification, and ranking. As the name suggests, it is based on the gradient-based learning model that uses an ensemble model of decision trees.

In each iteration, the algorithm learns from a decision tree based on residual error that brings such advantages as below:

- Faster training speed and higher efficiency
- Lower memory usage
- Better accuracy
- Support for parallel and GPU learning
- Capable of handling large-scale data (Big data)

This method splits the tree leaf wise whereas other boosting algorithms splits the tree depth or level wise. The leaf-wise algorithm can perform better than the level-wise algorithm that results in higher accuracy as a measurement.



Leaf-wise tree growth

Key evaluation methods

One of the most common ways of measuring accuracy is mean absolute error (MAE), and root mean square error (RMSE). To support such decision, decision support metrics are used which includes Precision, Recall and F1 score. Rank-aware evaluation metrics consists of MRR (Mean Reciprocal Rank), MAP (Mean Average Precision), and NDCG (Normalized Discounted Cumulative Gain).

Implementation
Although it requires further research on how to implement on Redback Operation's coin system and gaming, GBM ranking system is sound and effective as it provides deep intuition

in deriving an efficient way of scoring a user/player’s scores based on which rank does he/she belong, hence entitlement of the coins.

Cancer project

In recent years the image processing mechanisms are used widely in several medical areas. Cancer cure depends on early detection and treatment stages, in which the time factor is very important. If doctors can discover the disease in the patient as possible as fast. The use of AI in medical science has been attracting the attention of medical and healthcare industry in the latest years because of its high prevalence allied with the difficult treatment.

According to the ACS Guidelines on Nutrition and Physical Activity for Cancer Prevention, getting more physical activity is associated with a lower risk for several types of cancer, including breast, prostate, colon, endometrium, and possibly pancreatic cancer.
This project is aimed at developing a model capable of detecting cancer and in future to be integrated in games project, so the users who our part of our Open exercise community can utilize this system in our games to detect cancer or reduce risk for several types of cancer via means of exercise.

Methodology

- We will load the data by kera.preprocessing and build train_ds. The images will be labeled using the directry/folder structure.
- The number of images are less here.
- We will first run the base model.
- If the accuracy is less and loss factor is higher, we have to apply the augmentation technique to increase the number of images.
- Hence the next the model will go through couple of iterations to achieve the good accuracy.
- For model building we will use CNN – Deep Learning Architecture.
- The CNN algorithm needs GPU infrastructure to tarin the model. Hence model is running in google colab with GPU instance.

Convolutional Neural Network (CNN)

Convolutional Neural Networks, or CNNs, are specialized architectures which work particularly well with visual data, i.e. images and videos. They have been largely responsible for revolutionizing 'deep learning' by setting new benchmarks for many images processing tasks that were very recently considered extremely hard. Although the vanilla neural networks (MLPs) can learn extremely complex functions, their architecture does not exploit what we know about how the brain reads and processes images. For this reason, although MLPs are successful in solving many complex problems, they haven't been able to achieve any major breakthroughs in the image processing domain.

CNN Architecture

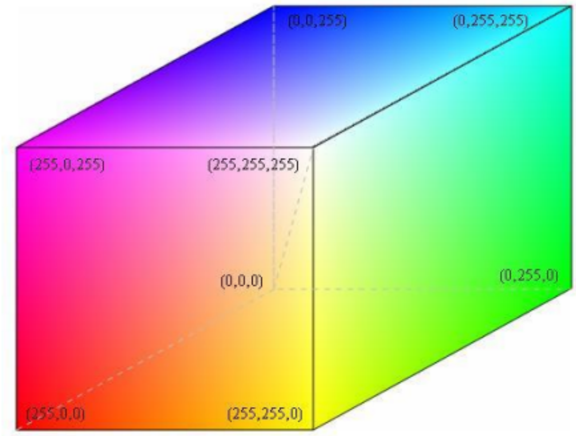


Figure 4: Pixel values in colour image

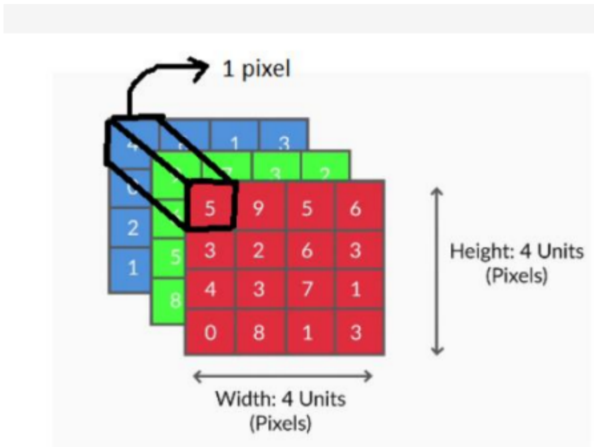


Figure 5: Color image represented as matrix

__Fig: 4 & 5__ - Represents the colour image of size 4x4x3, where height is 4, width is 4, the number of channels is 3(RGB). The number of pixels is 16 (height x width).

Convolution with colour image

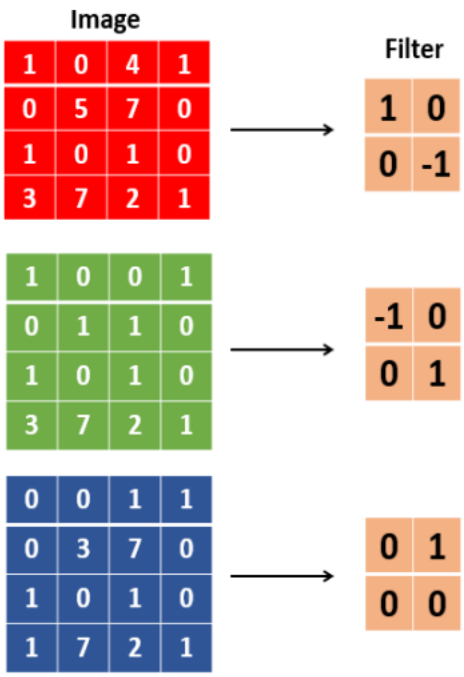


__Fig: 7 –__ RGB channel Images with filters

RGB channels of Image with corresponding filters. Image with RGB channel as in above figure and convolve with their corresponding filters.

The result is shown in below figure:

Result:

Visualizing Results

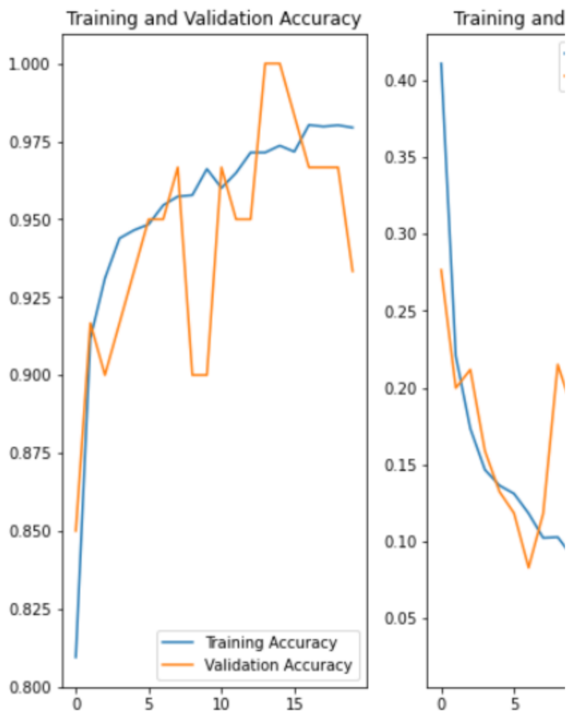
```
acc = history.history['accuracy']
val_acc = history.history['val_accuracy']

loss = history.history['loss']
val_loss = history.history['val_loss']

epochs_range = range(epochs)

plt.figure(figsize=(8, 8))
plt.subplot(1, 2, 1)
plt.plot(epochs_range, acc, label='Training Accuracy
plt.plot(epochs_range, val_acc, label='Validation Ac
plt.legend(loc='lower right')
plt.title('Training and Validation Accuracy')

plt.subplot(1, 2, 2)
plt.plot(epochs_range, loss, label='Training Loss')
plt.plot(epochs_range, val_loss, label='Validation L
plt.legend(loc='upper right')
plt.title('Training and Validation Loss')
plt.show()
```




```
loss, acc = model.evaluate(val_ds)
print("Accuracy", acc)
```

```
2/2 [==============================] - 0s 104ms/step - loss: 0.1742 - accuracy: 0.9333
Accuracy 0.9333333373069763
```

Blood Lactate Research and prediction model

Project team members:
Karl Birti

Research was conducted to analyse the importance of measuring blood lactate and the effects it has in the body under exercise conditions. A summary of the research detailing the methods as well as the benefits of collecting blood lactate data was included. Using a dataset with blood lactate as the target variable we explored the relationships of it with potential features with exploratory data analysis (EDA).
Features included were:

- Time
- Power
- VO2
- Cadence
- Respiratory Frequency
- Heart Rate

Feature importance was tested with the data, dropping variables that made trivial contributions to the prediction. Several models were built with hyperparameter tuning included to fine tune the accuracy.

Results

Model Results Summary Table

Model Type	Optimal Parameters (Grid Search)	R2 Score
Gradient Boosting Regressor	{'learning_rate': 0.1, 'max_depth': 9, 'n_estimators': 500, 'subsample': 0.7}	0.9946433172895832
MLP Regressor	MLPRegressor(activation='tanh', hidden_layer_sizes=(150, 100, 50), learning_rate='adaptive', max_iter=500)	0.987896403935139
Random Forest Regressor	{'max_depth': 400, 'n_estimators': 100}	0.9834825630374389

Gradient Boosting Regressor was the most effective machine learning algorithm once optimized to build a prediction model for blood lactate with an accuracy of 99.46%. The MLP Regressor and Random Forest Regressor respectively had accuracies of 98.79% and 98.34%. Overall, the 3 optimized models built here were highly accurate in predicting blood lactate in the test data.

Conclusion

In this research-based project we looked at the impact that several metrics (time, power, cadence, respiratory frequency, heart rate) have on the production of blood lactate (La). The ways in which blood lactate is measured are not practical or affordable for many people as method are typically invasive blood tests. The purpose of this study was to investigate machine learning algorithms with metrics commonly collected during cycling and cycling simulators to predict blood lactate levels for cyclists. Three models were chosen for this task and all of them achieved strong accuracy when tasked with predicting the test data. Given the results, it would be possible to apply this on new real data collected and refine the model further with greater samples collected.

Implementation

The blood lactate model has a couple of potential ways to be implemented, one would be to model the rider's effort during a workout as this gives valuable information for fatigue and difficulty. Another way is to integrate this into the incentive coin system, rewarding riders which maintain a rate of effort at or above the lactate threshold as this is most beneficial for increasing muscular endurance.

Overall project outcomes (Summary)

Findings

There are many algorithms that can be utilised to improve overall performance of Redback Operation. Since an actual data collection has not been practiced, many mock models have been developed and they are ready to be deployed once Gaming department & IoT sensor team launches a game for a demo run.

Implementations

Incentive system and GBM ranking system can be applied in recommending games, user's exercise model and many more when collaborated with O2, and blood lactate research and prediction model takes place.
More operational advancements can be made in an ad hoc manner when actual data sets are collected from each department.

Once data collection starts and outputs are gathered, analytics team can take their part and build a dashboard which will benefit the end users.

Limitations

Current limitation of algorithm team is subject to researches only as there is no actual data available for testing.

Recommendations

a) Collaborative research with game development team & IoT sensors team & Data Analytics team should be conducted in the following trimester
b) Further research to be carried on once actual data is gathered
c) Continue progressing on establishing an algorithm that connects all previous and future research