2024 /11 /12

20:00 - 21:00

11310CS460200 Group 23 Meeting Minutes

Topic	Progress Report
Place	Discord Voice Chat
Agenda	Discuss about our dataset collection and input format
In attendance	All present
Task Assigned	游松澤: collecting starting pitcher datasets 曾柏勳: collecting closer pitcher datasets 蕭以勝, 楊立慈, 賴允中: model design
Next meeting	Date: 11/19 Time: 8:00 ~ 9:00 pm Objective: Keep working on phase3 (finish training our first implementation of RNN and try to optimize it) Location: Discord Voice Chat

Meeting Summary:

1. I eventually find a python package called pybaseball, which is a platform that combines statcast, Baseball Reference, Baseball Savant, and FanGraphs data. Just easily import it and we can get the required dates data and features easily.

   Below is a partial screenshot of my data collection process. I selected **Shota Imanaga** as our starting pitcher and **Mason Miller** as our closer, gathering their complete data for the 2024 season. This resulted in **2,590 rows** for Imanaga and **995 rows** for Miller. In the future, we can expand by adding more features or extending the dataset to include additional years if needed.

Date	Pitcher Name	Batter ID	Batter Name	Pitch number	Pitch type	isStrike	Zone	Strike Detail	Description
2024-09-28	Miller, Mason	571745	mitch haniger	5	Slider	1	14	2	swinging_strike
2024-09-28	Miller, Mason	571745	mitch haniger	4	Slider	0	12	2	ball
2024-09-28	Miller, Mason	571745	mitch haniger	3	Slider	1	14	1	swinging_strike
2024-09-28	Miller, Mason	571745	mitch haniger	2	Slider	1	6	0	called_strike
2024-09-28	Miller, Mason	571745	mitch haniger	1	Slider	0	11	0	ball
2024-09-28	Miller, Mason	593871	jorge polanco	6	Slider	0	14	2	swinging_strike_blocked
2024-09-28	Miller, Mason	593871	jorge polanco	5	Slider	0	12	2	foul

2. We discussed and tried to build the prototype of our RNN model, maybe we can use the scraped data into training next week if everything works properly.
   Because some length of pitch sequences are not the same as sequence_length , so we want to use padding length to the sequence_length. To ignore the padding values, we use masking layer to train based on the real value. Each time step we need to get the probability of pitch type,so return_sequences ==True. We think we need more layers to build up some model and join more features.

```
from tensorflow.keras.optimizers import Adam
model = Sequential([
    Masking(mask_value=0.0, input_shape=(sequence_length, 1)),  # Masking 層忽略填充的 0
    SimpleRNN(10, activation='relu', return_sequences=True),  # 10 個神經元，並返回每個時間步的輸出
    Dense(4, activation='softmax')  # 輸出層，使用 softmax 進行分類，4 類（1 到 4）
])

# 編譯模型
model.compile(optimizer=Adam(), loss='sparse_categorical_crossentropy', metrics=['accuracy'])

# 訓練模型
model.fit(x_train, y_train, epochs=50, batch_size=16, validation_data=(x_val, y_val), verbose=1)

# 測試模型
test_loss, test_acc = model.evaluate(x_test, y_test, verbose=2)
print(f"Test accuracy: {test_acc * 100:.2f}%")
```

A group photo of the discussion session: