03_stacked_lstm_with_feature_embeddings_regression

September 29, 2021

1 Stacked LSTMs for Time Series Regression

We'll now build a slightly deeper model by stacking two LSTM layers using the Quandl stock price data (see the stacked_lstm_with_feature_embeddings notebook for implementation details). Furthermore, we will include features that are not sequential in nature, namely indicator variables for identifying the equity and the month.

1.1 Imports

[1]:
```
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
%matplotlib inline

from pathlib import Path
import numpy as np
import pandas as pd

from scipy.stats import spearmanr

import tensorflow as tf
from tensorflow.keras.callbacks import ModelCheckpoint, EarlyStopping
from tensorflow.keras.models import Sequential, Model
from tensorflow.keras.layers import Dense, LSTM, Input, concatenate, Embedding,␣
 ↪Reshape, BatchNormalization
import tensorflow.keras.backend as K

import matplotlib.pyplot as plt
from matplotlib.ticker import FuncFormatter
import seaborn as sns
```

[3]:
```
gpu_devices = tf.config.experimental.list_physical_devices('GPU')
if gpu_devices:
    print('Using GPU')
    tf.config.experimental.set_memory_growth(gpu_devices[0], True)
else:
    print('Using CPU')
```

```
Using CPU
```

[4]:
```python
idx = pd.IndexSlice
sns.set_style('whitegrid')
np.random.seed(42)
```

[5]:
```python
results_path = Path('results', 'lstm_embeddings')
if not results_path.exists():
    results_path.mkdir(parents=True)
```

1.2 Data

Data produced by the notebook build_dataset.

[6]:
```python
data = pd.read_hdf('data.h5', 'returns_weekly').drop('label', axis=1)
```

[7]:
```python
data['ticker'] = pd.factorize(data.index.get_level_values('ticker'))[0]
```

[8]:
```python
data['month'] = data.index.get_level_values('date').month
data = pd.get_dummies(data, columns=['month'], prefix='month')
```

[9]:
```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 1167341 entries, ('A', Timestamp('2009-01-11 00:00:00')) to ('ZUMZ',
Timestamp('2017-12-31 00:00:00'))
Data columns (total 66 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   fwd_returns  1167341 non-null  float64
 1   1            1167341 non-null  float64
 2   2            1167341 non-null  float64
 3   3            1167341 non-null  float64
 4   4            1167341 non-null  float64
 5   5            1167341 non-null  float64
 6   6            1167341 non-null  float64
 7   7            1167341 non-null  float64
 8   8            1167341 non-null  float64
 9   9            1167341 non-null  float64
 10  10           1167341 non-null  float64
 11  11           1167341 non-null  float64
 12  12           1167341 non-null  float64
 13  13           1167341 non-null  float64
 14  14           1167341 non-null  float64
 15  15           1167341 non-null  float64
 16  16           1167341 non-null  float64
 17  17           1167341 non-null  float64
 18  18           1167341 non-null  float64
```

```
 19   19           1167341 non-null  float64
 20   20           1167341 non-null  float64
 21   21           1167341 non-null  float64
 22   22           1167341 non-null  float64
 23   23           1167341 non-null  float64
 24   24           1167341 non-null  float64
 25   25           1167341 non-null  float64
 26   26           1167341 non-null  float64
 27   27           1167341 non-null  float64
 28   28           1167341 non-null  float64
 29   29           1167341 non-null  float64
 30   30           1167341 non-null  float64
 31   31           1167341 non-null  float64
 32   32           1167341 non-null  float64
 33   33           1167341 non-null  float64
 34   34           1167341 non-null  float64
 35   35           1167341 non-null  float64
 36   36           1167341 non-null  float64
 37   37           1167341 non-null  float64
 38   38           1167341 non-null  float64
 39   39           1167341 non-null  float64
 40   40           1167341 non-null  float64
 41   41           1167341 non-null  float64
 42   42           1167341 non-null  float64
 43   43           1167341 non-null  float64
 44   44           1167341 non-null  float64
 45   45           1167341 non-null  float64
 46   46           1167341 non-null  float64
 47   47           1167341 non-null  float64
 48   48           1167341 non-null  float64
 49   49           1167341 non-null  float64
 50   50           1167341 non-null  float64
 51   51           1167341 non-null  float64
 52   52           1167341 non-null  float64
 53   ticker       1167341 non-null  int64
 54   month_1      1167341 non-null  uint8
 55   month_2      1167341 non-null  uint8
 56   month_3      1167341 non-null  uint8
 57   month_4      1167341 non-null  uint8
 58   month_5      1167341 non-null  uint8
 59   month_6      1167341 non-null  uint8
 60   month_7      1167341 non-null  uint8
 61   month_8      1167341 non-null  uint8
 62   month_9      1167341 non-null  uint8
 63   month_10     1167341 non-null  uint8
 64   month_11     1167341 non-null  uint8
 65   month_12     1167341 non-null  uint8
dtypes: float64(53), int64(1), uint8(12)
```

```
memory usage: 498.8+ MB
```

1.3 Train-test split

To respect the time series nature of the data, we set aside the data at the end of the sample as hold-out or test set. More specifically, we'll use the data for 2017.

[10]:
```
window_size=52
sequence = list(range(1, window_size+1))
ticker = 1
months = 12
n_tickers = data.ticker.nunique()
```

[11]:
```
train_data = data.loc[idx[:, :'2016'], :]
test_data = data.loc[idx[:, '2017'],:]
```

For each train and test dataset, we generate a list with three input arrays containing the return series, the stock ticker (converted to integer values), and the month (as an integer), as shown here:

[12]:
```
X_train = [
    train_data.loc[:, sequence].values.reshape(-1, window_size , 1),
    train_data.ticker,
    train_data.filter(like='month')
]
y_train = train_data.fwd_returns
[x.shape for x in X_train], y_train.shape
```

[12]:
```
([(1035424, 52, 1), (1035424,), (1035424, 12)], (1035424,))
```

[13]:
```
# keep the last year for testing
X_test = [
    test_data.loc[:, list(range(1, window_size+1))].values.reshape(-1,␣
↪window_size , 1),
    test_data.ticker,
    test_data.filter(like='month')
]
y_test = test_data.fwd_returns
[x.shape for x in X_test], y_test.shape
```

[13]:
```
([(131917, 52, 1), (131917,), (131917, 12)], (131917,))
```

1.4 Define the Model Architecture

The functional API of Keras makes it easy to design architectures with multiple inputs and outputs. This example illustrates a network with three inputs, as follows:

- A two stacked LSTM layers with 25 and 10 units respectively
- An embedding layer that learns a 10-dimensional real-valued representation of the equities
- A one-hot encoded representation of the month

This can be constructed using just a few lines - see e.g., - the general Keras documentation, - the LTSM documentation.

Make sure you are initializing your optimizer given the keras-recommended approach for RNNs

We begin by defining the three inputs with their respective shapes, as described here:

[14]:
```
K.clear_session()
```

[15]:
```
n_features = 1
```

[16]:
```
returns = Input(shape=(window_size, n_features), name='Returns')
tickers = Input(shape=(1,), name='Tickers')
months = Input(shape=(12,), name='Months')
```

1.4.1 LSTM Layers

To define stacked LSTM layers, we set the `return_sequences` keyword to `True`. This ensures that the first layer produces an output that conforms to the expected three-dimensional input format. Note that we also use dropout regularization and how the functional API passes the tensor outputs from one layer to the subsequent layer:

[17]:
```
lstm1_units = 25
lstm2_units = 10
```

[18]:
```
lstm1 = LSTM(units=lstm1_units,
             input_shape=(window_size,
                          n_features),
             name='LSTM1',
             dropout=.2,
             return_sequences=True)(returns)

lstm_model = LSTM(units=lstm2_units,
             dropout=.2,
             name='LSTM2')(lstm1)
```

1.4.2 Embedding Layer

The embedding layer requires the `input_dim` keyword, which defines how many embeddings the layer will learn, the `output_dim` keyword, which defines the size of the embedding, and the `input_length` keyword to set the number of elements passed to the layer (here only one ticker per sample).

To combine the embedding layer with the LSTM layer and the months input, we need to reshape (or flatten) it, as follows:

[19]:
```
ticker_embedding = Embedding(input_dim=n_tickers,
                             output_dim=5,
                             input_length=1)(tickers)
```

```
ticker_embedding = Reshape(target_shape=(5,))(ticker_embedding)
```

1.4.3 Concatenate Model components

Now we can concatenate the three tensors and add fully-connected layers to learn a mapping from these learned time series, ticker, and month indicators to the outcome, a positive or negative return in the following week, as shown here:

[20]:
```
merged = concatenate([lstm_model,
                      ticker_embedding,
                      months], name='Merged')

bn = BatchNormalization()(merged)
hidden_dense = Dense(10, name='FC1')(bn)

output = Dense(1, name='Output')(hidden_dense)

rnn = Model(inputs=[returns, tickers, months], outputs=output)
```

The summary lays out this slightly more sophisticated architecture with 29,371 parameters, as follows:

[21]:
```
rnn.summary()
```

```
Model: "model"
________________________________________________________________________________
__________________
Layer (type)                    Output Shape         Param #     Connected to
================================================================================
==================
Returns (InputLayer)            [(None, 52, 1)]      0
________________________________________________________________________________
__________________
Tickers (InputLayer)            [(None, 1)]          0
________________________________________________________________________________
__________________
LSTM1 (LSTM)                    (None, 52, 25)       2700        Returns[0][0]
________________________________________________________________________________
__________________
embedding (Embedding)           (None, 1, 5)         12445       Tickers[0][0]
________________________________________________________________________________
__________________
LSTM2 (LSTM)                    (None, 10)           1440        LSTM1[0][0]
________________________________________________________________________________
__________________
reshape (Reshape)               (None, 5)            0           embedding[0][0]
________________________________________________________________________________
__________________
```

```
Months (InputLayer)            [(None, 12)]         0

--------------------------------------------------------------------------------
------------------
Merged (Concatenate)           (None, 27)           0            LSTM2[0][0]
                                                                 reshape[0][0]
                                                                 Months[0][0]

--------------------------------------------------------------------------------
------------------
batch_normalization (BatchNorma (None, 27)          108          Merged[0][0]

--------------------------------------------------------------------------------
------------------
FC1 (Dense)                    (None, 10)           280
batch_normalization[0][0]

--------------------------------------------------------------------------------
------------------
Output (Dense)                 (None, 1)            11           FC1[0][0]
================================================================================
==================
Total params: 16,984
Trainable params: 16,930
Non-trainable params: 54

--------------------------------------------------------------------------------
------------------
```

1.5 Train the Model

[22]:
```
optimizer =tf.keras.optimizers.Adam()

rnn.compile(loss='mse',
            optimizer=optimizer)
```

[23]:
```
lstm_path = (results_path / 'lstm.regression.h5').as_posix()

checkpointer = ModelCheckpoint(filepath=lstm_path,
                               verbose=1,
                               monitor='val_loss',
                               mode='min',
                               save_best_only=True)
```

[24]:
```
early_stopping = EarlyStopping(monitor='val_loss',
                              patience=5,
                              restore_best_weights=True)
```

[25]:
```
training = rnn.fit(X_train,
                   y_train,
                   epochs=50,
                   batch_size=64,
```

```
                   validation_data=(X_test, y_test),
                   callbacks=[early_stopping, checkpointer],
                   verbose=1)
```

```
Epoch 1/50
16174/16179 [============================>.] - ETA: 0s - loss: 0.0097
Epoch 00001: val_loss improved from inf to 0.00157, saving model to
results/lstm_embeddings/lstm.regression.h5
16179/16179 [==============================] - 157s 10ms/step - loss: 0.0097 -
val_loss: 0.0016
Epoch 2/50
16179/16179 [==============================] - ETA: 0s - loss: 0.0029
Epoch 00002: val_loss improved from 0.00157 to 0.00155, saving model to
results/lstm_embeddings/lstm.regression.h5
16179/16179 [==============================] - 155s 10ms/step - loss: 0.0029 -
val_loss: 0.0015
Epoch 3/50
16179/16179 [==============================] - ETA: 0s - loss: 0.0029
Epoch 00003: val_loss did not improve from 0.00155
16179/16179 [==============================] - 155s 10ms/step - loss: 0.0029 -
val_loss: 0.0016
Epoch 4/50
16173/16179 [============================>.] - ETA: 0s - loss: 0.0028
Epoch 00004: val_loss did not improve from 0.00155
16179/16179 [==============================] - 156s 10ms/step - loss: 0.0028 -
val_loss: 0.0015
Epoch 5/50
16178/16179 [============================>.] - ETA: 0s - loss: 0.0028
Epoch 00005: val_loss did not improve from 0.00155
16179/16179 [==============================] - 155s 10ms/step - loss: 0.0028 -
val_loss: 0.0016
Epoch 6/50
16178/16179 [============================>.] - ETA: 0s - loss: 0.0028
Epoch 00006: val_loss improved from 0.00155 to 0.00154, saving model to
results/lstm_embeddings/lstm.regression.h5
16179/16179 [==============================] - 154s 10ms/step - loss: 0.0028 -
val_loss: 0.0015
Epoch 7/50
16179/16179 [==============================] - ETA: 0s - loss: 0.0028
Epoch 00007: val_loss did not improve from 0.00154
16179/16179 [==============================] - 145s 9ms/step - loss: 0.0028 -
val_loss: 0.0016
Epoch 8/50
16177/16179 [============================>.] - ETA: 0s - loss: 0.0028
Epoch 00008: val_loss did not improve from 0.00154
16179/16179 [==============================] - 145s 9ms/step - loss: 0.0028 -
val_loss: 0.0015
```

```
Epoch 9/50
16179/16179 [==============================] - ETA: 0s - loss: 0.0028
Epoch 00009: val_loss improved from 0.00154 to 0.00154, saving model to
results/lstm_embeddings/lstm.regression.h5
16179/16179 [==============================] - 149s 9ms/step - loss: 0.0028 -
val_loss: 0.0015
Epoch 10/50
16178/16179 [============================>.] - ETA: 0s - loss: 0.0028
Epoch 00010: val_loss did not improve from 0.00154
16179/16179 [==============================] - 144s 9ms/step - loss: 0.0028 -
val_loss: 0.0015
Epoch 11/50
16177/16179 [============================>.] - ETA: 0s - loss: 0.0028
Epoch 00011: val_loss did not improve from 0.00154
16179/16179 [==============================] - 148s 9ms/step - loss: 0.0028 -
val_loss: 0.0015
Epoch 12/50
16175/16179 [============================>.] - ETA: 0s - loss: 0.0028
Epoch 00012: val_loss did not improve from 0.00154
16179/16179 [==============================] - 144s 9ms/step - loss: 0.0028 -
val_loss: 0.0016
Epoch 13/50
16173/16179 [============================>.] - ETA: 0s - loss: 0.0028
Epoch 00013: val_loss improved from 0.00154 to 0.00154, saving model to
results/lstm_embeddings/lstm.regression.h5
16179/16179 [==============================] - 147s 9ms/step - loss: 0.0028 -
val_loss: 0.0015
Epoch 14/50
16177/16179 [============================>.] - ETA: 0s - loss: 0.0027
Epoch 00014: val_loss did not improve from 0.00154
16179/16179 [==============================] - 148s 9ms/step - loss: 0.0027 -
val_loss: 0.0016
Epoch 15/50
16176/16179 [============================>.] - ETA: 0s - loss: 0.0027
Epoch 00015: val_loss did not improve from 0.00154
16179/16179 [==============================] - 145s 9ms/step - loss: 0.0027 -
val_loss: 0.0015
Epoch 16/50
16179/16179 [==============================] - ETA: 0s - loss: 0.0027
Epoch 00016: val_loss did not improve from 0.00154
16179/16179 [==============================] - 149s 9ms/step - loss: 0.0027 -
val_loss: 0.0015
Epoch 17/50
16174/16179 [============================>.] - ETA: 0s - loss: 0.0027
Epoch 00017: val_loss did not improve from 0.00154
16179/16179 [==============================] - 146s 9ms/step - loss: 0.0027 -
val_loss: 0.0016
Epoch 18/50
```

```
16177/16179 [============================>.] - ETA: 0s - loss: 0.0027
Epoch 00018: val_loss did not improve from 0.00154
16179/16179 [==============================] - 149s 9ms/step - loss: 0.0027 -
val_loss: 0.0015
```

[26]:
```
loss_history = pd.DataFrame(training.history)
```

1.6 Evaluate model performance

[27]:
```
test_predict = pd.Series(rnn.predict(X_test).squeeze(), index=y_test.index)
```

[28]:
```
df = y_test.to_frame('ret').assign(y_pred=test_predict)
```

[29]:
```
by_date = df.groupby(level='date')
df['deciles'] = by_date.y_pred.apply(pd.qcut, q=5, labels=False,␣
↪duplicates='drop')
```

[30]:
```
ic = by_date.apply(lambda x: spearmanr(x.ret, x.y_pred)[0]).mul(100)
```

[31]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
MultiIndex: 131917 entries, ('A', Timestamp('2017-01-01 00:00:00')) to ('ZUMZ',
Timestamp('2017-12-31 00:00:00'))
Data columns (total 3 columns):
 #   Column   Non-Null Count   Dtype
---  ------   --------------   -----
 0   ret      131917 non-null  float64
 1   y_pred   131917 non-null  float32
 2   deciles  131917 non-null  int64
dtypes: float32(1), float64(1), int64(1)
memory usage: 3.1+ MB
```

[32]:
```
test_predict = test_predict.to_frame('prediction')
test_predict.index.names = ['symbol', 'date']
test_predict.to_hdf(results_path / 'predictions.h5', 'predictions')
```

[33]:
```
rho, p = spearmanr(df.ret, df.y_pred)
print(f'{rho*100:.2f} ({p:.2%})')
```

```
4.68 (0.00%)
```

[34]:
```
fig, axes = plt.subplots(ncols=2, figsize=(14,4))
sns.barplot(x='deciles', y='ret', data=df, ax=axes[0])
axes[0].set_title('Weekly Fwd Returns by Predicted Quintile')
axes[0].yaxis.set_major_formatter(FuncFormatter(lambda y, _: '{:.2%}'.
↪format(y)))
```

```
axes[0].set_ylabel('Weekly Returns')
axes[0].set_xlabel('Quintiles')

avg_ic = ic.mean()
title = f'4-Week Rolling IC | Weekly avg: {avg_ic:.2f} | Overall: {rho*100:.2f}'
ic.rolling(4).mean().dropna().plot(ax=axes[1], title=title)
axes[1].axhline(avg_ic, ls='--', c='k', lw=1)
axes[1].axhline(0, c='k', lw=1)
axes[1].set_ylabel('IC')
axes[1].set_xlabel('Date')

sns.despine()
fig.tight_layout()
fig.savefig(results_path / 'lstm_reg');
```