STAT430: Machine Learning for Financial Data

LarryHua.com/teaching

Spring 2019

Classification and regression - Examples

Using the multi-assignment (%<-%) operator

The datasets built into Keras are all nested lists of training and test data. Here, we use the multi-assignment operator (%<-%) from the zeallot package to unpack the list into a set of distinct variables.

```
imdb <- dataset_imdb(num_words = 10000)
c(c(train_data, train_labels), c(test_data, test_labels)) %<-% imdb
```

This could equally be written as follows:

```
train_data <- imdb$train$x
train_labels <- imdb$train$y
test_data <- imdb$test$x
test_labels <- imdb$test$y
```

- The multi-assignment version is preferable because it's more compact. The %<-% operator is automatically available whenever the R Keras package is loaded.
- In Keras, 0, 1, and 2 are reserved for padding, start of sequence, and unknown

Binary classification

- Sequences of words can be encoded as binary vectors (eg, 01000110000010000010000000000)
- Stacks of dense layers with **relu** activations can solve a wide range of problems
- In a binary classification problem, use a dense layer with one unit and a sigmoid activation, and the loss function should be **binary_crossentropy**
- Try R

Multiclass classification

- For N classes classification, end with a dense layer of size N
- Use a **softmax** activation to output a probability distribution over the N output classes
- Catergotical crossentropy as the loss function
- Two ways to handle labels in multiclass classification:
    - Encoding the labels via one-hot encoding, and using **categorical_crossentropy** as a loss function
    - Encoding the labels as integers, and using the **sparse_categorical_crossentropy** loss function
- If there are a large number of categories, avoid creating information bottlenecks with too small intermediate layers
- Try R

Scalar regression

- Mean Squared Error (MSE) is a common loss function for regression
- Mean Absolute Error (MAE) is a common regression metric
- **Feature** should be scaled independently
- When there is little data available, using K-Fold validation is a great way to reliably evaluate a model
- When little training data is available, it is preferable to use a small network with very few hidden layers (1 or 2), in order to avoid severe overfitting
- Try R
- Back to Course Scheduler