Notes on Forecasting/Sequence Modeling

Sameer Kesava PhD

Contents

Framework

A set of questions needs to be clearly answered to understand and frame the time-series forecasting problem.

1. What are the Inputs and Outputs for the forecast?
2. What are the Endogenous and Exogenous variables?
3. Is the data Structured or unstructured?
4. Is it a Regression or Classification problem?
5. Is it a Univariate or Multivariate problem?
6. Does the problem require a Static or Dynamic model?
7. Is the data Contiguous or Discontiguous?

If the problem is non-stationary, i.e. if the data is seasonal, then it needs to be made stationary either by gathering data over a time scale where it is stationary or using differencing methods.

Improving the Accuracy

2.1 Issues

- Long term dependencies is a problem.

2.1.1 Vanishing Gradient

LSTM or GRU architecture helps resolve this problem.

2.1.2 Exploding Gradient

1. Gradient clipping
2. Nonlinear activation functions such as relu6.

2.2 Weights

Parameter sharing is the main feature of RNNs. The parameters, i.e. same weights, are shared across several time steps.

1. Sample or Class weights can be used to improve the prediction accuracy.
2. Alternatively, a user-defined function can be supplied to compute the weights.

2.2.1 ML

2.2.2 DL

(a) A good starting values for the initializers are Glorot-Xavier or Orthogonal for the weights.

(b) For the bias, zeros suffice.

2.3 Brute-Force

Brute-force algorithm is effective for small datasets.
Query time grows as O[pn], where p is the number of features/predictors/dimensions and n is no. of data points.

2.4 Small Dataset

n: number of samples
p: number of features/predictors

2.4.1 When p is comparable to n

1. Use shrinkage methods such as Lasso or L1 norm.
2. Check Brute-Force section 2.3.

2.4.2 $p \ll n$

1. Use k-fold CV.

2.5 Embedding

Embedding of characters and words in character/language modeling is a popular way of encoding in sequence modeling.

2.6 Statefulness

- Stateful = True, i.e. in many-to-many sequence modeling, where there is one-to-one mapping between row i of batch j with row i of batch j+1.

2.7 Norm of the Gradient

- During training, the norm of the gradient clipped can also be dynamically adjusted (from NIPS:Oral Session 4 -Ilyasuts).

2.8 Curriculum Learning

To help stabilize the model's open-loop output.

2.9 Teacher Forcing

Used in Encoder-Decoder section 4.6 where the target word is passed as the next input to the decoder.

2.10 Hyperparameter Tuning

- Hyperparameter optimization based on Gaussian Process Regression and Bayesian Optimization
- keras tuner in keras
- GridSearchCV or RandomSearchCV in scikit-learn
- RandomSearch can be used as the baseline against which optimization algorithms can be evaluated.

ML Methods

3.1 ARIMA

- Implemented in statsmodel.
- Cannot model seasonal trends. Hence, differencing should be carried out to remove this.

3.2 XGBoost

Highly popular classifier and regressor.

- Gradient boosting method
- Absolute loss and Huber loss more robust to outliers.

DL Methods

All operations, e.g. loss functions, activation functions, in neural networks must be differentiable.

- If the sequences are not of the same size, e.g. Sentiment Analysis, either padding or masking can be used.

4.1 Binary Classification

- Should end with a dense layer of 1 unit and a sigmoid activation function.
- The loss function should be binary cross-entropy.

4.2 CNN-LSTM

Hyrbid model.

- Invariance is hard-coded in CNNs.(?)

4.3 ConvLSTM

Hybrid model.

4.4 TreeRNN

4.5 Bidirectional RNN

- Helps RNN learn long range dependencies in the sequences.

- For language modeling.

4.6 Encoder-Decoder

1. Bahdanau et al. - RNN encoder, RNN decoder with Attention, i.e. contextual output from the encoder as input to the decoder.

2. In Neural Machine Translation (NMT), where the above method was applied, the encoder output is trained to capture the semantic information, which is then passed to the decoder.

Loss Functions

5.1 Sparse Categorical Cross Entropy

To be used when the labels are not one-hot encoded.

5.2 Binary Cross Entropy

1. If the labels have only 2 unique values, preferably [0,1] or [-1,1].
2. Use sigmoid activation function for [0,1] and tanh for [-1,1].

5.3 KL Divergence

If each label is a probability distribution, e.g. image segmentation.

5.4 Perplexity

A function used in language modeling: e^{Loss}

Metrics

6.1 BLEU score

Used in language modeling, e.g. Bahdanau et al. paper; Encoder-Decoder section 4.6.

Optimizers

1. If the training and validation accuracies with epoch jump around, then learning rate must be reduced.

7.1 Stochastic Gradient Descent

- For training one sample at a time.
- For online learning.

7.2 Mini-batch Gradient Descent

- For mini-batches typically of the order of 2^x for GPU compatibility.
- Examples are Adam, Adagrad, RMSProp, etc.

Overfitting

8.1 DL

- L1, L2, L1L2 regularization
- Dropout
- Batch Normalization
- Reduce the complexity of the network especially if the data size is small.

8.1.1 Dropout

- For a RNN layer, there are 3 dropouts: for the input from the previous layer, output of the layer and the recurrent dropout.
- Dropout functions
    1. Gaussian Dropout