Challenge report

November 26, 2018

0.1 Mácio Matheus Santos de Arruda

0.1.1 Desafio

Objetivo: previsão do IPCA (Índice Nacional de Preços ao Consumidor Amplo) do próximo mês.

Prêmio: a metodologia que obtiver o menor MAE para as previsões de Jan/2017 a Set/2018 (base de dados de teste) ganhará um livro de Inteligência Artificial! Rode o seu método 30 vezes e obtenha a média e o desvio-padrão dos resultados.

0.1.2 Relatório da execução do desafio

1 - Foram plotados gráficos sobre o dataset

* Serie
* Densidade
* Lag

2 - Arquiteturas

A arquitetura LSTM(long short time memory), é um tipo especial de RNN, capaz de aprender dependências de longo prazo, ou seja, sempre que o fator "tempo"ou até mesmo "ordem (sequência)" estejam presentes nos dados, é possível conseguir resultados interessantes, pois recordar a informação por longos períodos de tempo é praticamente o seu comportamento padrão, não é algo que a arquitetura luta para aprender.

E essse comportamento é exatamente o que estava buscando na resolução do problema IPCA, no entanto o resultado obtido (0.21) não foi melhor do que a RandomForest (0.12).

A RandomForest é um método conjunto no qual um classificador ou regressor (no problema do desafio) é construído pela combinação de vários classificadores/regressores independentes de base.
Essa técnica é conhecida como bagging ou agregação de bootstrap. Uma outra característica da Random Forest, é diminuir o risco de overfitting, pois ela analisa subamostra aleatória de seus dados.

3 - Pipeline de execução

3.1 Os dados foram montados com uma janela de 3(meses)
3.2 Os dados foram normalizados em uma escala entre 0 e 1
3.3 Os dados foram separados de acordo com a exigência (Jan/2017 a Set/2018) para teste e o restante para treinamento.
3.4 Foi realizado o treinamento dos dados na LSTM utilizando Keras
3.5 Foi realizado o treinamento dos dados usando uma Random Forest Regressor do Sklearn
3.6 Foram executados 30 experimentos para validar relevância estatística de cada uma das técnicas
3.7 Foram armazenadas as métricas por experimento numa lista
3.8 Foram calculadas as métricas e plotados gráficos de alguns dos experimentos

4 - Detalhes dos parâmetros da arquitetura LSTM

4.1 LSTM
    4.1.1 A rede possui apenas 1 camada densa
    4.1.2 O otimizador utilizado foi o SGD (gradient descent)
    4.1.3 A função de loss escolhida foi a MSE
    4.1.4 O treinamento foi realizado em 200 épocas

4.2 Random Forest Regressor
    4.2.1 A profundidade máxima da árvore foi de 50
    4.2.2 O critério que mede a qualidade da divisão foi o MAE
    4.2.3 Número de estimators foi 200

5 - Técnicas experimentadas

5.1 LSTM
5.1 RandomForest Regressor (Melhor para o problema)

6 - Resultados - LSTM

6.1 media MAE: 0.21505316556209608
6.2 desvio padrão: 0.011281130460127181
6.3 variância: 0.0001272639044584093

7 - Resultados - RandomForest Regressor (Melhor resultado)

7.1 Mean of MAE: 0.12421896387683362
7.2 Standard deviation: 0.0030893803868903267
7.3 Variance: 9.544271174902625e-06