Previsão de Séries Temporais com Programação Genética

(versão #4 atualizada em 24/04/2020)

Este exercício tem o objetivo de utilizar a Programação Genética para gerar um modelo para a previsão de séries temporais. Os dados fornecidos se referem ao ativo PETR4 (Petrobrás, ações preferenciais) negociadas na BOVESPA durante o período de 16/05/2018 a 17/04/2019. A partir da série temporal dos valores negociados do ativo minuto-a-minuto são calculados os seguintes indicadores:

- indicador SMA – *Simple Moving Average* de 10 períodos (1 se valor de fechamento do minuto for maior e -1 o contrário) - (https://www.investopedia.com/terms/s/sma.asp)
- Indicador WMA – *Weighted Moving Average* de 10 períodos (1 se valor de fechamento do minuto for maior e -1 o contrário) - (https://www.investopedia.com/articles/technical/060401.asp)
- Indicador MACD – *Moving Average Convergence Divergence* (1 se valor de fechamento do minuto for maior e -1 o contrário) - https://www.investopedia.com/terms/m/macd.asp
- Indicador RSI – *Relative Strength Index*, indica mercado sobrecomprado X sobrevendido (1 se valor de fechamento do minuto for maior e -1 o contrário) - https://www.investopedia.com/terms/r/rsi.asp
- Indicador MOM – *Momentum* (1 se valor de fechamento do minuto for maior e -1 o contrário) - https://www.investopedia.com/terms/m/momentum.asp
- TEND – Tendência do mercado para o próximo momento: -1 (DOWN) se a houver tendência em descer o valor do ativo; 0 (STABLE) se a tendência for ficar estável o valor do ativo; +1 (UP) se houver a tendência em subir o valor do ativo.

Os dados de 16/05/2018 (11:37) a 28/12/2018 (55000 pontos) devem ser utilizados para a criação do modelo (**TREINAMENTO**) através da PG, enquanto que os dados de 02/01/2019 a 17/04/2019 (26000 pontos) devem ser utilizados para **TESTE** do modelo.

O objetivo é aplicar o paradigma da Programação Genética para encontrar uma função que, dado o conjunto de indicadores no instante *t* (5 variáveis preditoras independentes), possa prever a tendência do ativo no instante *t+1* (variável dependente). Isto é:

$$TEND(t+1) = f(SMA(t), WMA(t), MACD(t), RSI(t), MOM(t))$$

Como não se pode inferir previamente nenhuma sazonalidade nos dados, fica a critério do aluno utilizar ou não amostras de instantes *t-1, t-2*, etc. Os conjuntos de funções e de terminais são de livre escolha do aluno. É permitido criar novas funções ou utilizar as funções matemáticas básicas. O conjunto de terminais deve incluir pelo menos as 5 variáveis preditoras e a variável dependente. Constantes arbitrárias podem ser utilizadas a critério do aluno. Considerando que a PG pode criar uma equação que gera valores reais e a saída desejada é ternária *{+1, 0, -1}*, é necessário criar uma função *wrapper* para reinterpretar a saída contínua da PG convertendo-a para os valores esperados.

Fica também a critério do aluno utilizar parte dos dados de treinamento, caso o tempo de processamento seja muito elevado. Não há restrição para a profundidade das árvores criadas e evoluídas pela PG. Porém, a solução final deve ser restrita em até **50** nós. Observe que quanto maior o número de nós, mais complexa será a árvore. Assim, o número de nós deve ser ajustado com parcimônia. Após o treinamento da PG e obtenção da equação, simplificá-la de modo a ficar **humanamente compreensível** e utilizável para a próxima etapa.

Deve-se aplicar a equação obtida a todos os dados de TESTE (26000 pontos) para contrastar o resultado obtido pela PG com o valor real. Sumarizar os resultados com uma matriz de confusão (TEND real versus TEND predito pela PG), conforme mostrado a seguir.

		Previsto		
		+1(sobe)	0 (estável)	-1(desce)
Real	+1(sobe)			
	0 (estável)			
	-1(desce)			

Preferencialmente utilizar o software Lil-GP, porém, outro software similar poderá ser utilizado.

Preparar um relatório objetivo contendo:

1. Informar os diversos conjuntos parâmetros, funções e terminais que foram testados na PG, e aqueles que efetivamente obtiveram o melhor ajuste.
2. A melhor expressão-S obtida (com identação, para melhor entendimento).
3. A equação correspondente **já simplificada**.
4. A matriz de confusão e a acurácia para cada uma das três classes de saída.
5. Conclusão sobre o uso da PG para a previsão de séries temporais.