EA072 –Exercícios de Fixação de Conceitos 1 (EFC1)

Seleção de Variáveis e Predição de Séries Temporais – Abordagens Lineares e Não-Lineares

Peso da Lista: 4 || Data de entrega: 01/10/2015

1 Introdução

Em qualquer ramo de atuação profissional, percebe-se um crescimento acentuado na demanda por predições e detecção de tendências junto a variáveis de interesse. Ferramentas computacionais capazes de fornecer predições automáticas acerca de valores futuros de certas variáveis que estão sendo monitoradas já fazem parte do dia-a-dia de muitas empresas privadas e órgãos governamentais, e têm contribuído para o sucesso na definição de políticas estratégicas, em processos de tomada de decisão e em todo tipo de planejamento de curto e médio prazo.

Predição de séries temporais é uma área de pesquisa muito ampla, com desdobramentos na Estatística, na Computação e na Matemática. Serão tratados neste EFC1 apenas alguns aspectos básicos e fundamentais, visando fazer com que o aluno se familiarize com técnicas de síntese de preditores lineares e não-lineares, em particular técnicas de inteligência computacional. Um aprofundamento do estudo pode ser buscado junto à literatura recomendada.

Técnicas de seleção de variáveis também serão trabalhadas aqui, também envolvendo abordagens lineares e não-lineares. Definir o melhor subconjunto de entradas, dentre as candidatas, é uma etapa de grande relevância em tarefas de aprendizado de máquina, incluindo predição de séries temporais. Como a seleção de variáveis pode ser considerada uma etapa de pré-processamento de dados, ela é objeto de estudo dos primeiros exercícios de fixação de conceitos do curso.

2 Seleção de variáveis empregando filtros e *wrappers*

A primeira atividade envolve uma comparação entre filtros e *wrappers*. Serão considerados como filtros o coeficiente de correlação de Pearson, que só captura correlações lineares entre pares de variáveis, e um índice denominado *distance correlation*, capaz de considerar correlação não-linear entre pares de variáveis. Serão considerados como *wrappers* as técnicas de *forward selection* e de *backward selection*, ambas empregando um modelo linear sem regularização, visando reduzir o custo computacional. As aplicações envolvem duas tarefas de regressão, embora a metodologia seja igualmente válida para tratar problemas de classificação.

(1) Recorra ao paper [Guyon, I.; Elisseeff, A. "An introduction to variable and feature selection", Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003] para explicar a diferença entre filtro (usado na Questão 1) e *wrapper*.

(2) Explique como funcionam as abordagens *forward selection* e *backward selection* e apresente a razão pela qual elas não garantem encontrar a melhor combinação de entradas para a tarefa.

(3) **Caso de estudo 1**: Série temporal SUNSPOT, do ano de 1749 ao ano de 2013 [http://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Data/sunspot.long.data]. São considerados 20 atrasos candidatos, são incluídas 5 entradas aleatórias e é empregado 10-folds cross-validation. Todas as variáveis excursionam no intervalo [0,+1]. O modelo do preditor é linear e sem regularização.

   a. Descreva as principais características desta série temporal;
   b. Ordene as entradas candidatas de acordo com as abordagens de filtro linear e de filtro não-linear, comparando os resultados. Use os programas [filtro_lin.m('dados1')] e [filtro_nlin.m('dados1')].
   c. Use os programas em Matlab [prog1.m], para *forward selection*, e [prog2.m], para *backward selection*, fornecidos pelo professor, para obter os resultados junto à série temporal. Execute 5 vezes cada programa e apresente numa tabela as entradas que foram selecionadas em cada abordagem (que levam ao mínimo erro quadrático de validação), lembrando que as 10 pastas são redefinidas a cada execução. Compare os resultados das abordagens *forward selection* e *backward selection*.
   d. Explique por que nem sempre as entradas de maior correlação linear ou não-linear são as primeiras a serem selecionadas na abordagem *forward selection*;
   e. Explique por que nem sempre as entradas de menor correlação linear ou não-linear (ainda sem considerar as 5 entradas aleatórias) são as primeiras a serem podadas na abordagem *backward selection*;
   f. Explique por que as 5 entradas aleatórias, também consideradas como entradas candidatas, não são as últimas a serem selecionadas na abordagem *forward selection* e as primeiras a serem podadas na abordagem *backward selection*.

(4) **Caso de estudo 2**: Conjunto de dados *Wine Quality* do *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml/datasets/Wine+Quality]. Foi considerado o caso com 1599 entradas e foram eliminadas as linhas 1361, 1364, 1441, 1443, 1477 e 1516, por conterem dados espúrios. São considerados 11 atributos de entrada, são incluídas 5 entradas aleatórias e é empregado 10-folds cross-validation. Todas as variáveis excursionam no intervalo [0,+1]. O modelo do regressor é linear e sem regularização.

   a. Descreva as principais características deste problema de regressão;
   b. Ordene as entradas candidatas de acordo com as abordagens de filtro linear e de filtro não-linear, comparando os resultados. Use os programas [filtro_lin.m('dados2')] e [filtro_nlin.m('dados2')].
   c. Use os programas em Matlab [prog11.m], para *forward selection*, e [prog21.m], para *backward selection*, fornecidos pelo professor, para obter os resultados junto aos dados. Execute 5 vezes cada programa e apresente numa tabela as entradas que foram selecionadas em cada abordagem (que levam ao

mínimo erro quadrático de validação), lembrando que as 10 pastas são redefinidas a cada execução. Compare os resultados das abordagens *forward selection* e *backward selection*.

3 Séries temporais e a tarefa de predição

Uma série temporal é dada pelos valores ao longo do tempo de uma variável de interesse, como em:

- ✓ Atividades vitais ou funções orgânicas de um indivíduo;
- ✓ Índices econômicos;
- ✓ Índices sociais;
- ✓ Variáveis climáticas;
- ✓ Variáveis de ecossistemas;
- ✓ Monitoramento de operação de plantas industriais;
- ✓ Monitoramento de processos químicos e físicos.

A frequência de amostragem depende da aplicação e geralmente é fixa. Os intervalos de amostragem mais comuns são ano, mês, semana, dia e hora.

Sendo a predição uma estimativa de valores futuros a partir do conhecimento do histórico de uma variável até o presente, surgem algumas questões:

- ✓ O histórico de uma variável até o presente é capaz de auxiliar na predição do seu comportamento futuro?
- ✓ Como geralmente a variável de interesse tem seu comportamento atrelado a uma grande quantidade de fatores e a uma complexa rede de inter-relações, como é possível predizer seu comportamento sem modelar os fenômenos complexos que regem o comportamento da variável e sem monitorar outras variáveis que influenciam nesse comportamento? Exemplo: Como predizer a vazão de um rio com base apenas em seu histórico, sem levar em conta valores atuais para a vazão dos afluentes, o nível de chuvas na cabeceira do rio, o consumo de água para irrigação agrícola e o seu grau de assoreamento, dentre outros fatores?

Existem séries temporais que não admitem predição, seja pela independência estatística intrínseca entre o que já se conhece da série e o valor futuro que se pretende predizer, seja pelo fato de que aspectos relevantes do processo estão sendo negligenciados, como em séries multidimensionais em que alguma(s) variável(is) não está(ão) sendo monitorada(s). Nesses casos, geralmente as iniciativas de predição tendem a produzir resultados decepcionantes.

Por outro lado, existem séries temporais que apresentam uma dependência estatística significativa entre o histórico da série e o valor futuro a ser predito. Nesses casos, as predições tendem a produzir resultados de predição de alta qualidade, mesmo que não se conheçam especificidades do processo que gera a série temporal.

Esta dependência estatística (entre o histórico da série e o valor a ser predito) pode ser estimada a priori. Com isso, antes mesmo de se iniciar o projeto do preditor já é possível ter uma boa noção de como pode vir a ser o seu desempenho naquela tarefa de predição. Além disso, análises de dependência estatística auxiliam na definição da estrutura do preditor, mais especificamente na definição do vetor de regressão, ou seja, o vetor de entrada do preditor. Para tanto, geralmente são respondidas duas perguntas:

- Qual é o tamanho da janela de valores passados a serem considerados na entrada do preditor?
- Quais valores passados dentro desta janela devem ser considerados?

Coeficientes de correlação linear (contextualizados por Francis Galton em 1885 e devidamente formalizados por Karl Pearson em 1895) (RODGERS & NICEWANDER, 1988; STIGLER, 1989) são muito utilizados para responder a essas perguntas, mas deve-se salientar que eles supõem o uso subsequente de preditores lineares. Um índice mais consistente, capaz de capturar dependências lineares e não-lineares, é a informação mútua, derivado da Teoria de Informação (COVER, 1991; REZA, 1994; WESTIAN, 1990).

3.1 Metodologia

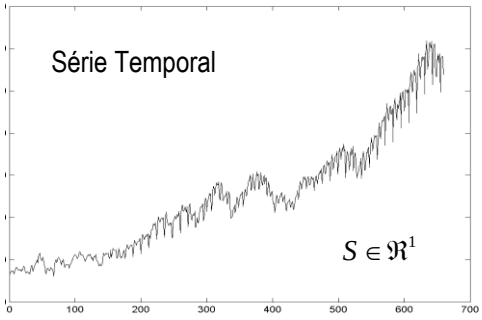
As primeiras tentativas no campo da predição de séries temporais foram efetuadas nos anos 20, quando YULE (1927) aplicou um modelo autorregressivo linear no estudo de manchas solares. Nos anos 50, DOOB (1953) prosseguiu a investigação com a análise teórica de séries temporais estacionárias. Já nos anos 70, foram propostas as técnicas e metodologias que obtiveram maior destaque a partir de então, reunidas no trabalho de BOX & JENKINS (1976).

Os métodos de Box & Jenkins baseiam-se na proposição de que o valor atual da série temporal é a combinação de p valores precedentes e q impactos aleatórios anteriores, mais o impacto atual. Os p valores precedentes formam o *componente autorregressivo* e os q impactos prévios formam o *componente de média móvel* da série. Obtêm-se assim os bem conhecidos modelos ARMA (do inglês *autoregressive moving average*). A modelagem de uma série temporal tem por objetivo, então, a determinação dos valores de p e q, seguida da estimação dos respectivos coeficientes da combinação linear (que é um problema de regressão linear).

Nos últimos anos, considerável atenção tem sido dedicada a métodos alternativos para o estudo de séries com dependências não-lineares, destacando-se a utilização de redes neurais artificiais. O emprego das arquiteturas MLP e RBF trouxe resultados muito positivos no campo da predição de valores futuros em séries temporais, em virtude do caráter essencialmente não-linear dessas estruturas.

Para o emprego de uma rede neural artificial como preditor de um passo à frente, é necessário definir quais valores passados da série serão utilizados na definição da entrada da rede neural. Feito isso, o problema de síntese do preditor se transforma em um problema de treinamento supervisionado, onde o que se deseja é obter um mapeamento multidimensional não-linear de entrada-saída, como indicado na sequência de passos abaixo.

Passo 1: Obter a série temporal, ou seja, os valores históricos da variável a ser predita um passo à frente. Se necessário, normalize os dados (média zero e variância unitária), evitando que o intervalo de excursão dos valores seja qualquer. Outros tipos de pré-processamento, como diferenciar a série temporal, também podem ser considerados visando eliminar tendências e sazonalidades, por exemplo.



$$S = s_1 , s_2 , s_3 , s_4 , s_5 , s_6 , s_7 , s_8 \ldots , s_N$$

Passo 2: Definir quais valores passados da série serão considerados na predição. Suponha aqui que L valores passados consecutivos sejam considerados. Com isso, monte a tabela a seguir, a qual retrata o comportamento desejado do preditor.

$\boldsymbol{s}_{t-L+1}$	$\boldsymbol{s}_{t-L+2}$	$\cdots$	$\boldsymbol{s}_t$	$\boldsymbol{s}_{t+1}$
s_1	s_2	$\cdots$	s_L	s_{L+1}
s_2	s_3	$\cdots$	s_{L+1}	s_{L+2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s_{N-L}	s_{N-L-1}	$\cdots$	s_{N-1}	s_N
$\Downarrow$	$\Downarrow$		$\Downarrow$	$\Downarrow$
X_1	X_2		X_L	Y

Passo 3: Separe os dados da tabela acima em 3 conjuntos: conjunto de treinamento, conjunto de validação e conjunto de teste. Não é necessário que a separação seja sequencial.
Passo 4: Treine a rede neural com o conjunto de treinamento (ela vai produzir um mapeamento do $\Re^L$ no $\Re^1$) e pare o treinamento quando for atingido o valor mínimo do erro quadrático médio para o conjunto de validação.
Passo 5: Avalie o preditor recém-obtido junto aos dados de teste.

3.2 Atividade a ser desenvolvida

3.2.1 Síntese de um preditor linear

Supondo um modelo autorregressivo de predição linear na forma:

$$x(k) = b_1 x(k-1) + b_2 x(k-2) + ... + b_L x(k-L) + b_{L+1},$$

monte a matriz $A = \begin{bmatrix} X_1 & \cdots & X_L & \vec{1} \end{bmatrix}$, onde $\vec{1}$ é um vetor-coluna de 1's e resolva o seguinte sistema linear:

$$A\vec{b} = Y,$$

produzindo o vetor de coeficientes $\vec{b} = \begin{bmatrix} b_1 & b_2 & \cdots & b_L & b_{L+1} \end{bmatrix}^T$ na forma:

$$\vec{b} = \left(A^T A\right)^{-1} A^T Y.$$

Para o caso não-regularizado acima, use apenas o conjunto de treinamento, deixando o conjunto de validação para comparação de desempenho com o preditor não-linear.

Proponha também coeficientes regularizados, na forma:

$$\vec{b} = \left(A^T A + cI\right)^{-1} A^T Y$$

onde a matriz identidade I tem a mesma dimensão de $A^T A$, e encontre um valor adequado para c no conjunto $\{2^{-24}, 2^{-23}, ..., 2^{+24}, 2^{+25}\}$, monitorando o erro junto ao conjunto de validação. O que se busca resolver aqui é o problema de quadrados mínimos regularizado, como segue:

$$\vec{b} = \arg \min_{\vec{b} \in \Re^{L+1}} \left\| A\vec{b} - Y \right\| + c \times \left\| b \right\|^2$$

Regularizar, portanto, implica em reduzir o módulo do vetor $\vec{b}$, ou seja, reduzir o grau de flexibilidade da combinação linear. O caso não-regularizado implica em tomar $c \to 0$.

3.2.2 Síntese de uma rede neural MLP

Treine uma rede neural MLP, usando os conjuntos de treinamento e validação e o programa em Matlab fornecido pelo professor. Defina o número n de neurônios na camada intermediária (Sugestão: Escolha um valor entre 1 e 20). A função de ativação é $g(u) = \tanh(u)$. Repare que $\frac{dg}{du} = 1 - g(u)^2$.

A rede neural é apresentada na figura a seguir, sendo que $m = L$ (número de entradas) e $r = 1$ (número de saídas).

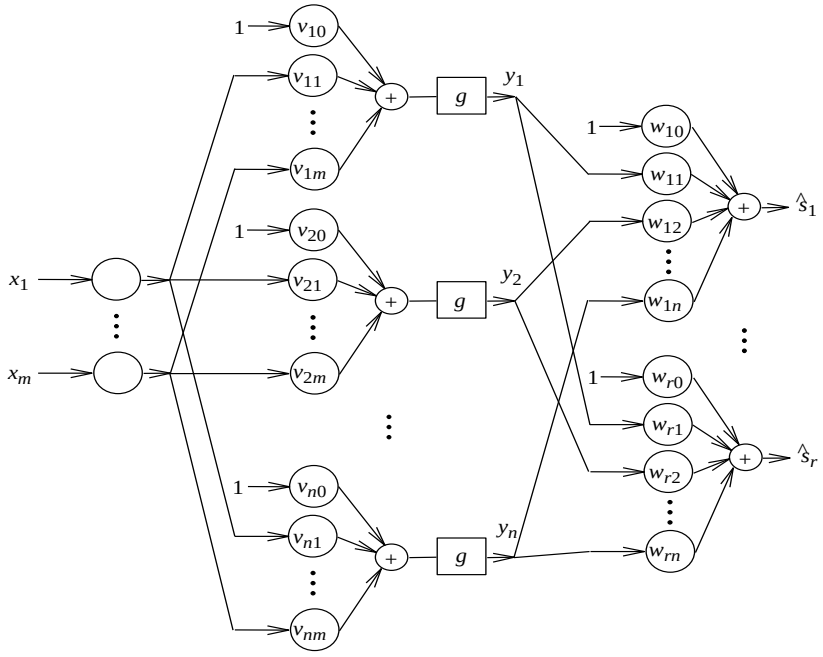


Figura 1 – Rede neural MLP com uma camada intermediária. Neste estudo, será considerada sempre uma única saída para a rede neural ($r = 1$).

3.3 Série temporal a ser tratada

Preditores lineares e não-lineares serão comparados junto à série temporal de casos confirmados de dengue, a cada semana, na cidade de São Paulo, no período de 2000 a 2014, sendo que esses dados foram disponibilizados pelo SUS (Sistema Único de Saúde). O preditor linear deve ser um modelo auto-regressivo e o preditor não-linear deve ser uma rede neural MLP com 10 neurônios na camada intermediária. Use o programa [calc_corr2.m] para obter a correlação entre o valor a ser predito e 20 valores passados da série temporal e constate que utilizar os 5 valores consecutivos mais recentes é uma escolha aceitável. Divida o conjunto de dados em 10 pastas e considere 9 delas para treinamento e a que sobrou para validação. Faça isso 10 vezes, de modo que todas as pastas sejam empregadas ao menos uma vez para validação. Apresente o desempenho médio obtido pelas 10 MLPs treinadas. Em seguida, componha as 10 MLPs num ensemble. Trabalhe com a média. Faça o mesmo para uma rede neural *extreme learning machine* (ELM) e para o preditor linear auto-regressivo, sendo que para esses dois casos empregue *ridge regression* para obter cada preditor, obtendo o parâmetro de regularização no conjunto de valores candidatos $\{0,2^{-24},2^{-23},...,2^{+24},2^{+25}\}$.

3.4 Algoritmo para treinamento da rede neural MLP

Está presente no toolbox fornecido pelo professor quando do anúncio deste roteiro de atividades.

3.5 Como evitar sobreajuste no treinamento

Uma técnica bem conhecida para se evitar o sobreajuste consiste em separar os padrões em três conjuntos: treinamento, validação e teste (PRECHELT, 1997; PRECHELT, 1998). O conjunto de treinamento é usado para ajustar os pesos sinápticos. Após cada época de ajuste de pesos, calcula-se o erro junto ao conjunto de validação. Assim, quando este valor apresentar uma tendência definida de aumento, será um indicativo de sobreajuste e o treinamento pode ser interrompido. O conjunto de teste, por sua vez, é utilizado para indicar como ficaria o desempenho da rede neural em operação. É suposto que os três conjuntos contêm amostras independentes e são todos capazes de representar bem o problema que está sendo abordado. Por exemplo, espera-se que um bom desempenho junto ao conjunto de validação implique em um bom desempenho junto ao conjunto de teste. Na prática, particularmente quando há recursos computacionais disponíveis, o mais indicado é sobretreinar a rede neural e ir armazenando o conjunto de pesos associado ao valor mínimo do erro junto ao conjunto de validação. Como o comportamento do erro de validação pode ser errático, detectar quando se atingiu o mínimo pode ser pouco confiável, sendo mais indicado forçar o sobretreinamento e tomar o conjunto de pesos associado ao instante em que o treinamento deveria ter parado, embora tenha seguido adiante. É assim que o programa fornecido pelo professor opera.

4 Sugestões para a elaboração do relatório

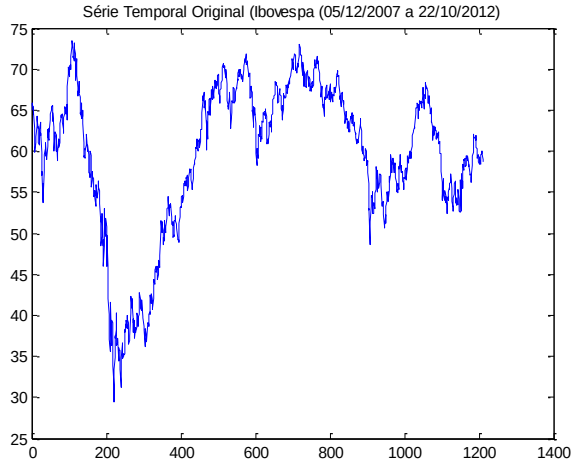
- Para todos os preditores lineares da Seção 3, além da análise de desempenho, apresentar o conjunto de coeficientes (vetor $\vec{b}$ da seção 3.2.1).
- Nos casos em que devem ser propostos preditores lineares e não-lineares, comparar o desempenho de ambos. Usar o erro de predição junto ao conjunto de validação. Para tanto, empregar a raiz do erro quadrático médio (rEQM), na forma:

$$\text{rEQM} = \sqrt{\frac{\sum_{j=1}^{N} (\text{valor\_obtido}_j - \text{valor\_esperado}_j)^2}{N}}$$

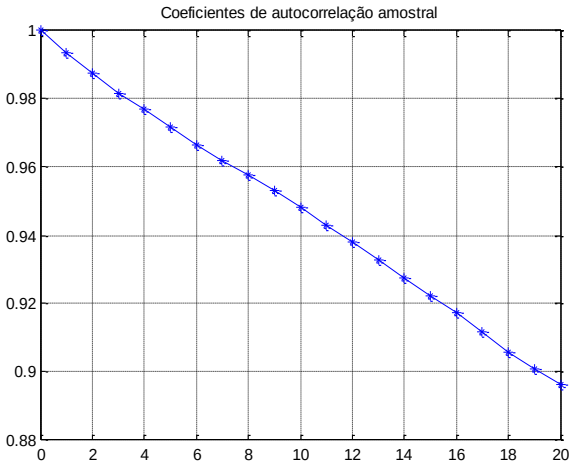
- Lembre-se que o erro quadrático (EQ) representa a soma dos erros ao quadrado, sendo, portanto, diferente do erro quadrático médio. Tanto EQ quanto EQM dependem do intervalo de excursão da série temporal, mas apenas EQ depende do número de amostras.
- Apresentar o método escolhido para definir o número de neurônios na camada intermediária de cada rede MLP usada como preditor. Sugestão: Sintetizar preditores variando de 1 a 20 neurônios na camada intermediária, selecionando a rede neural que produzir o menor erro junto ao conjunto de validação.
- Não basta apenas apresentar os resultados obtidos, é necessário analisá-los.
- Incluir uma seção final, com comentários conclusivos acerca das atividades realizadas.
- Evitar que o relatório fique muito extenso. Recorrer a tabelas e gráficos que sumarizem os resultados.
- Incluir referências bibliográficas.

5 Predição do índice Ibovespa: curiosidades

Em muitas situações práticas, as séries temporais são não-estacionárias e apresentam tendências e sazonalidades que precisam ser eliminadas antes de iniciar a análise. Além disso, quando a variação da série é incremental entre instantes consecutivos, a predição de próximo passo não consegue capturar o incremento, pois a predição é dominada pelo valor atual da série, sendo o incremento insignificante. Uma situação típica pode ser ilustrada pelo Ibovespa (índice diário da Bolsa de Valores de São Paulo), apresentado na figura abaixo.

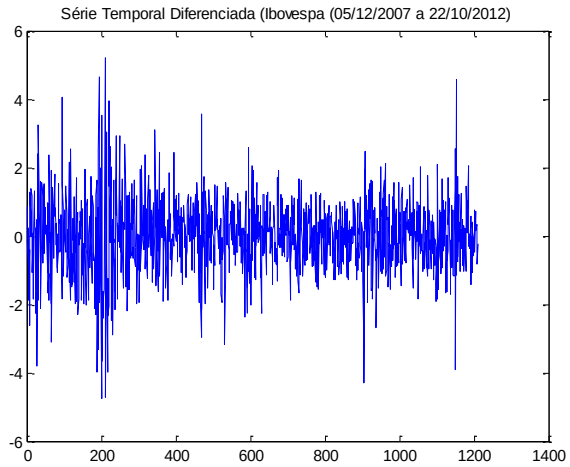


Os coeficientes de autocorrelação associados a esta série produzem:

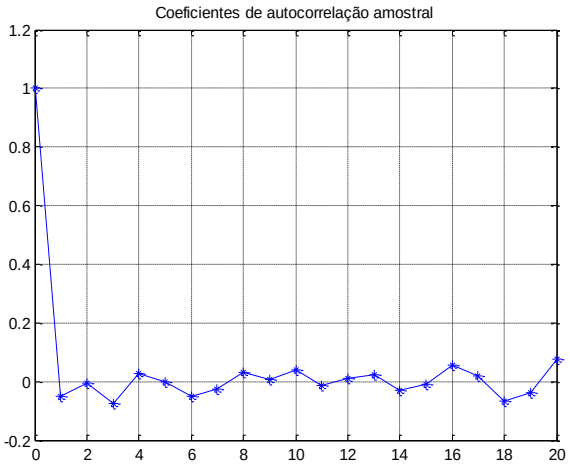


Como as variações diárias são incrementais em relação ao valor cheio do índice, as autocorrelações entre valores próximos no tempo são muito altas, se mantendo acima de 0,9 até o instante passado 19.

Faz-se necessário, então, diferenciar a série temporal, ou seja, obter uma nova série que é dada pela diferença entre valores consecutivos da série original. Para o Ibovespa, resulta:

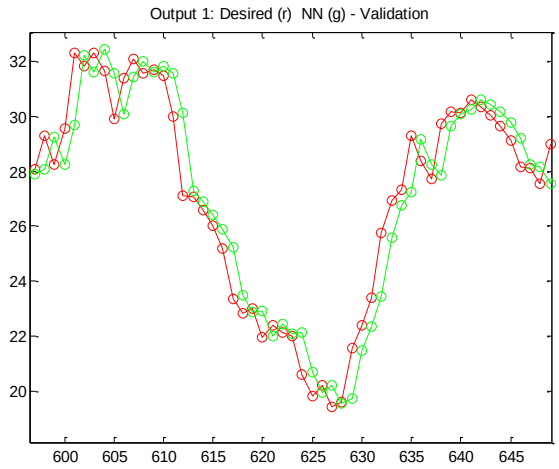


Embora a diferenciação da série tenha sido feita visando permitir capturar a variação diária (que é o que efetivamente interessa na prática), no caso particular do Ibovespa os novos coeficientes de autocorrelação produzem:



Este resultado deixa evidente que não há correlação significativa entre o valor diferencial (variação diária) que se quer predizer e os valores diferenciais passados da série.

Este cenário dá origem a um fenômeno pouco explicado na literatura, que é chamado de predição atrasada. Se houver a iniciativa de predição da série original (não de sua versão diferenciada), o preditor vai apresentar um comportamento similar ao ilustrado na figura abaixo (série original em vermelho e predição em verde).



A interpretação é a seguinte: Como a correlação é muito forte com o passado da série original, mas é quase nula com o passado da série diferenciada, para reduzir o erro quadrático médio, o preditor tende a ignorar a diferença e faz a predição de um passo à frente com um valor muito próximo ao valor presente da série temporal, produzindo este efeito de atraso.

Conclusão: O preditor faz o melhor que ele pode, mas o melhor que ele pode é inútil, em termos práticos, pois ele prediz o futuro com o valor presente (ou quase isso).

6 Referências

BOX, G. E. P. & JENKINS, G. M. (1976). *Time Series Analysis, Forecasting and Control*. Holden Day.

BRADLEY, E. (1998). *Time-Series Analysis*. University of Colorado.

CHAKRABORTY, K., MEHROTRA, K., MOHAN, C. K. & RANKA, S. (1992). Forecasting the Behavior of Multivariate Time Series Using Neural Networks. *Neural Networks*, vol. 5, pp. 961-970.

COVER, T. M. & THOMAS, J. A. (1991) *Elements of Information Theory*, Wiley.

DOOB, J. (1953). *Stochastic Processes*. Wiley.

HAYKIN, S. (1999). *Neural Networks – A Comprehensive Foundation*. IEEE Press, 2nd edition.

PRECHELT L. (1997). Early Stopping - but when?, Technical Report URL: http://wwwipd.ira.uka.de/~prechelt/Biblio/stop_tricks1997.ps.gz

PRECHELT L. (1998). Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks*, vol. 11, no 4, pp. 761-767.

REZA, F. M. (1994) *An Introduction to Information Theory*, Dover.

RODGERS, J. L. & NICEWANDER, W. A. (1988) Thirteen ways to look at the correlation coefficient. *The American Statistician*, vol. 42, no. 1, pp. 59-66.

STIGLER, S. M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science*, vol. 4, no. 2, pp 73-79.

WEIGEND, A. S. & GERSHENFELD, N. A. (1993). *Time Series Prediction: Forecasting the Future and Understanding the Past*. Perseus Press.

WENTIAN, L. (1990) Mutual information functions versus correlation functions. *Journal of Statistical Physics*, vol. 60, nos. 5-6, pp. 823-837.

YULE G. (1927). On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philos. Trans. R. Soci.*, A226.

7 Pesquisadores de apoio

Além do professor da disciplina, os alunos da disciplina podem solicitar apoio técnico e tirar dúvidas com o doutorando da FEEC:

- Marcos Medeiros Raimundo [marcosmrai(_at_)gmail.com]

8 Agradecimento

Algumas figuras, dados, textos e referências são de autoria do pós-doutorando Wilfredo Jaime Puma Villanueva, que realizou a sua pós-graduação junto ao Programa de Pós-Graduação da FEEC/Unicamp.

O correr da vida embrulha tudo. A vida é assim, esquenta e esfria, aperta e daí afrouxa, aquieta e depois desinquieta. O que ela quer da gente é coragem. O que Deus quer é ver a gente aprendendo a ser capaz de ficar alegre e amar, no meio da alegria. E ainda mais no meio da tristeza. Todo caminho da gente é resvaloso. Mas, também, cair não prejudica demais – a gente levanta, a gente sobe, a gente volta!

João Guimarães Rosa – Grande Sertão: Veredas (1956)