EFC 1 – Regressão Linear

Rafael Alves Mayer, RA: 150992 rmayer@ifi.unicamp.br
Pedro Felipe G. de Vazquez, RA: 263293 p263293@dac.unicamp.br

Resumo

Para esta atividade foi aplicado o método de regressão linear para a predição de séries temporais. Os últimos 15% das amostras foram reservadas para teste e o restante foi dividido entre treinamento e validação, utilizando um método de validação cruzada em séries temporais (*time series split*). Afim de otimizar os hiper-parâmetros do modelo, para os diferentes blocos de teste e validação, obtivemos o melhor valor do número de amostras de entrada (K) que resultaram no menor erro de validação com a métrica Raíz do Erro Quadrático Médio (RMSE). Ademais, ilustramos a predição da série temporal com a melhor escolha de K, resultando num RMSE de 0.048.

Na segunda parte, os dados foram normalizados e foi explorada uma técnica chamada de *Extreme Learning Machines* (ELM), onde é explorada um mapeamento não-linearidade do vetor de entradas a partir de elementos aleatórios. Utilizando o mesmo método de validação cruzada em acordo com o método de regularização Ridge Regression foram otimizados os melhores valores do hiper parâmetro de regularização (α) e a dimensão do novo vetor mapeado (V). Por fim, utilizamos os melhores hiper-parâmetros, K, α, e V, para a predição da série temporal em questão. Obtivemos um RMSE de 0.036, que é ligeiramente melhor do que aquele obtido na primeira parte.

Introdução e validação-cruzada

A previsão de séries temporais é um importante problema do ponto de vista tecnológico e econômico. Neste EFC iremos estudar a série de Mackey-Glass, uma série caótica, não-linear com aplicações na biologia. Os primeiros 500 pontos da série podem ser observados na figura 1, onde p representa a densidade de células brancas do sangue.

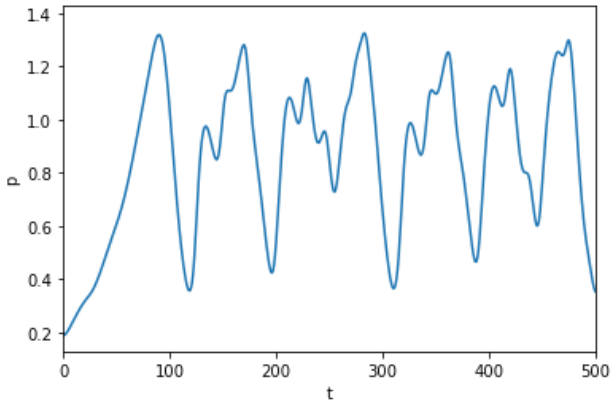

Figura 1 – Dados da série de Mackey-Glass retirados da base fornecida para o exercício.

Para a previsão da série, utilizaremos o modelo de regressão linear. Conforme observado na Figura 2a, para prever o resultado de p(t) na posição t=n, utilizaremos um vetor **x**(n), composto de K amostras da série, distanciadas por L amostras de t=n. Para a validação cruzada, os dados temporais foram estruturados em forma de matriz, conforme a Figura 2b, onde cada linha compõe um vetor **x**(n), e o vetor coluna de target (**y**) é dado pelos valores a serem previsto. Exemplificando para K =4 e L=5, o vetor de entrada X_1 será composto pelos valores *p* cujos índices são [0,1,2,3] e y1 pelo índice [8]. O segundo X_2 é dado pelos índices [4,5,6,7] e y2 por [12], e assim por diante. Note que dessa forma, todos os elementos do vetor *p* são utilizados.

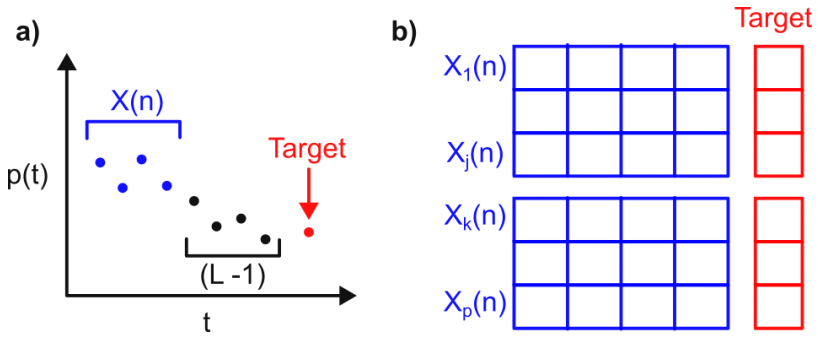

Figura 2 – a) Exemplo de série temporal, onde se utiliza o vetor x(n) de dimensão k=4, para se prever o resultado p(t) com um horizonte de predição L=5. b) Estrutura de dado em forma de matriz, onde se elimina os pontos entre x(n) e o "target". Os blocos representam k-folds para a validação cruzada. Note que dessa forma é respeitado o aspecto temporal dos dados.

Para ainda preservar o aspecto temporal dos dados, utilizamos a validação cruzada baseada na função *TimeSeriesSplit*, da biblioteca *sklearn* em python, com parâmetro 5 divisões, como mostrado na Figura 3. A escolha desse parâmetro significa que os conjuntos de treinamento e validação são divididos em 5 blocos, com incremento sucessivo, ou seja, no segundo bloco, o modelo usa para treinamento dados de validação do primeiro bloco. Pela natureza desse método de divisão, é levado em conta a ordem das amostras na série temporal. No final, do modelo chegamos a 5 valores de *scoring* (função que nos diz o quão próximos estamos dos resultados desejados), dos quais retiramos a média e o desvio padrão. O desvio padrão servirá como base para observar o intervalo de confiança das curvas de otimização. O *scoring* utilizado aqui é o de *Root Mean Squared Error* (RMSE).

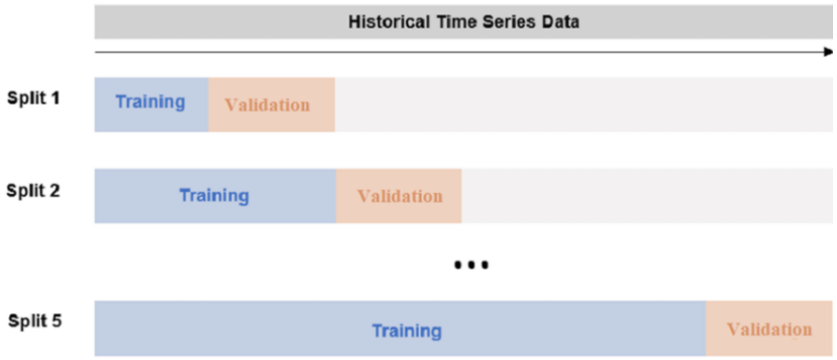

Figura 3 – Exemplo da divisão de blocos em treinamento e validação, para 5 divisões, onde a cada bloco se acrescenta os dados de validação no treinamento do bloco seguinte.

Parte 1

Na parte 1 do exercício, reservamos os 15% finais dos dados para teste. Além disso, estruturamos os dados de treino conforme descrito acima, e aplicamos a validação cruzada onde observamos a média de RMSE e seu desvio padrão. Para buscar o hiper parâmetro que melhor ajusta com o modelo, aplicamos a regressão linear para cada valor de K, conforme mostrado na figura 3.

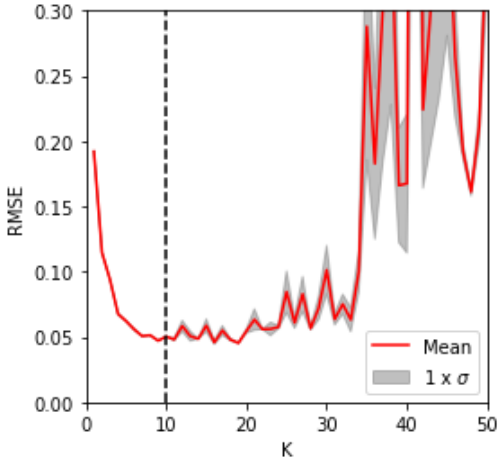

Figura 4 – Busca do hiperparâmetro K, onde aplicamos a regressão linear para cada valor de K e calculamos seu RMSE. A linha tracejada indica a posição K em que resulta no modelo mais simples e otimizado. A parte em cinza no gráfico representa o desvio padrão resultante da validação cruzada.

Podemos observar que há uma queda acentuada no começo da curva, o que indica que o aumento de número de amostras para o vetor **x**(n), melhora o modelo. Após cerca K = 30, a média de RMSE aumenta, o que pode indicar um sobre-ajuste dos parâmetros do modelo. Visto que buscamos o modelo mais simples que atinja um bom nível de RMSE, selecionamos o hiperparâmetro K = 10 para a previsão.

Para prever a resposta dos dados de teste, ajustamos o modelo com o conjunto de dados de treino inteiro. Utilizando os coeficientes de pesos do modelo, prevemos o resultado da série temporal nos dados de teste com excelente concordância, conforme mostrado na figura 5. O RMSE obtido nesse modelo para os dados testes foi de 0.048.

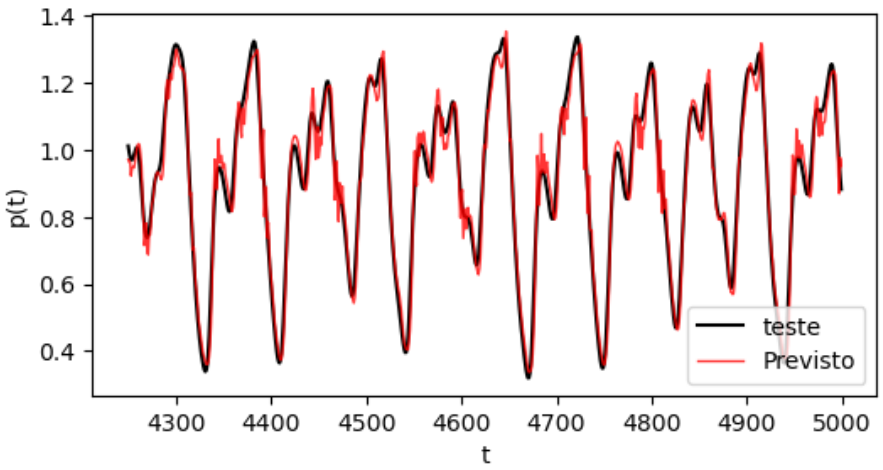

Figura 3 – Dados de teste (linha preta) e previsão do modelo (linha tracejada vermelha), utilizando os coeficientes do modelo de regressão linear ajustados nos dados de treino.

Parte 2

A parte 2 do EFC requer um mapeamento dos vetores **x**(n) de dimensão k, para um novo vetor **x'**(n) de dimensão V, onde cada componente k é descrita por:

$$x'_k(n) = \tanh\left(w_k^T x(n)\right) \quad (1)$$

Onde os elementos de w_k^T são vetores cujos elementos são gerados aleatoriamente por uma distribuição aleatória uniforme. Os dados primeiramente foram normalizados (Z-score) e em seguida os hiper parâmetros foram distribuídos entre o limite inferior (low) de -0.5 e superior (max) de +0.5, de forma que a $\tanh\left(w_k^T x(n)\right)$ ocupe todos os valores entre -1 e 1, como mostrado na figura 6.

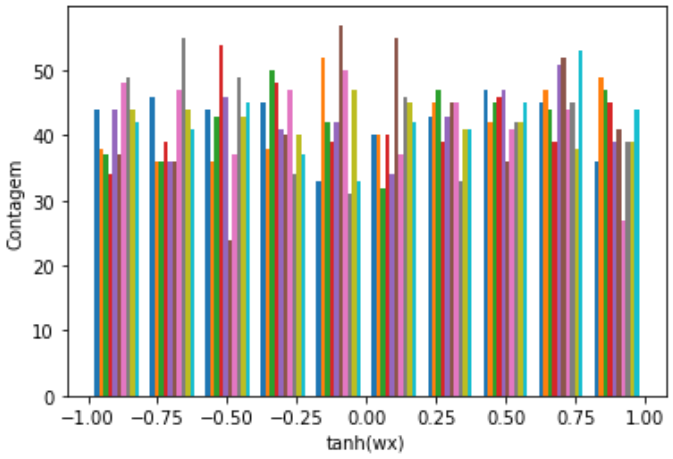

Figura 6 – Distribuição de valores de $\tanh\left(w_k^T x(n)\right)$, para V =10.

Para V = 200, low =-0.5, max = 0.5, K = 10, L = 7 foi investigado o melhor valor de α. Aqui podemos ver que para essa aplicação o valor de log(alpαha) próximo à -4 resulta no menor valor de RMSE.

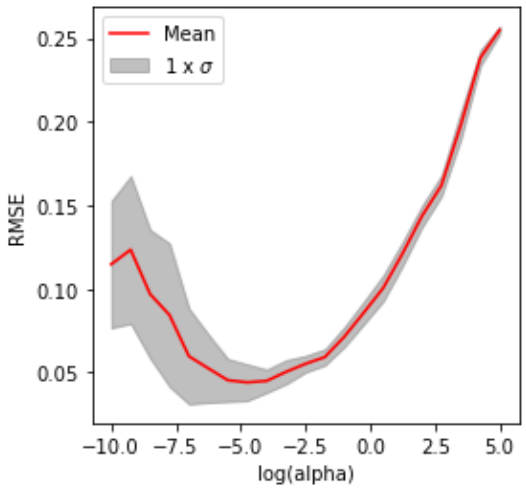

Figura 7 – Busca do hiper parâmetro α, utilizando como referência o menor valor de RMSE. O log aqui é na base 10. A parte em cinza no gráfico representa o desvio padrão resultante da validação cruzada.

Para $\alpha = 10^{-4}$, foi procurado o melhor valor do hiper parâmetro Vm como ilustrado na figura 8. Com esse comportamento, vemos que com poucos atributos V, o erro aumenta e para muitas amostras o erro cai, atingindo seu ponto mínimo em torno de $V = log\,(2.5) = 316$, porém com uma incerteza relevante.

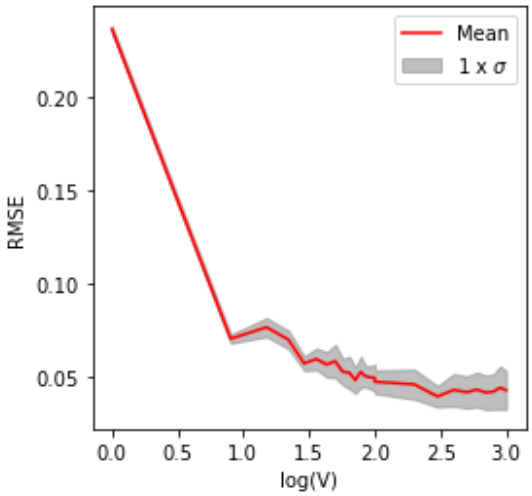

Figura 8 – Busca do hiper parâmetro V, utilizando o menor valor de RMSE

Para encontrar o melhor valor conjunto de α e V, foi feito um mapa de calor com o valor da média (Figura 9) e do desvio-padrão (Figura 10) do RMSE

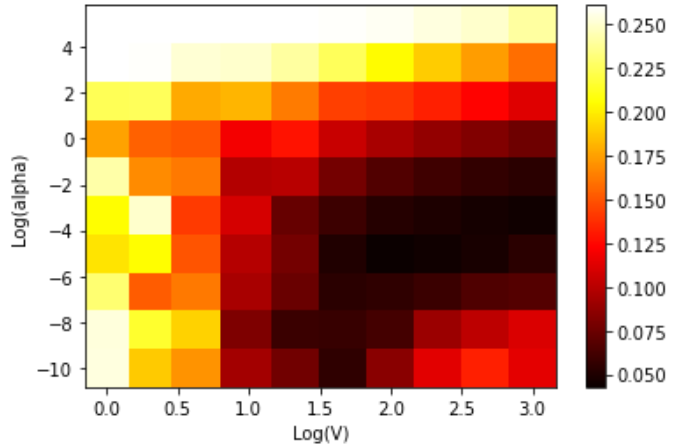

Figura 9 – Mapa de calor ilustrando a média do RMSE para cada hiper parâmetro V e α.

Dessa maneira, os melhores valores encontrados foram $V = 100$ e $\alpha = 10^{-5}$. Também apresentamos na figura 10 o desvio padrão do RMSE por completeza. Para baixos valores de α, e altos valores de V, o desvio padrão aumenta consideravelmente. Isso pode representar uma falta de robustez do modelo pra esses valores.

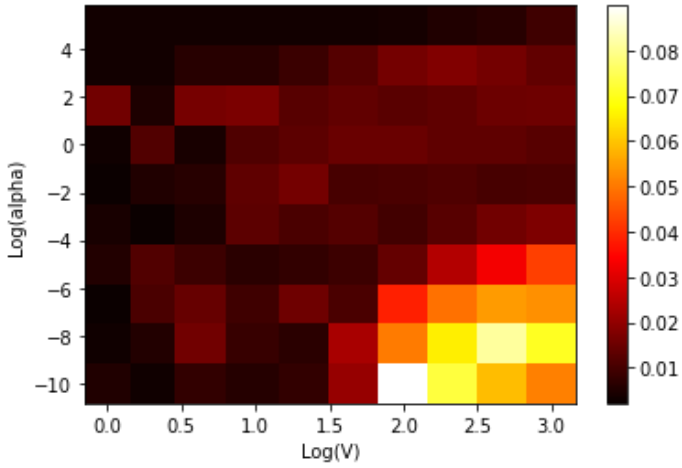

Figura 10 – Mapa de calor ilustrando o desvio-padrão do RMSE para cada hiper parâmetro V e α.

Na figura 11, mostramos os melhores valores de α em função de V. Note que os melhores valores de α aumentam com V, o que é compreensível já que quanto maior a dimensão dos vetores de peso, maior será a possibilidade de sobre-ajuste, e consequentemente, exigindo uma maior regularização do modelo.

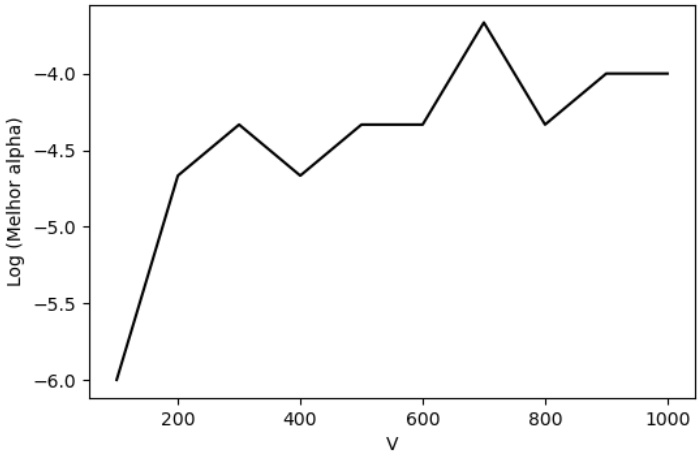

Figura 11 – Gráfico dos melhores valores de α em função de V.

A figura 12, abaixo, ilustra a predição com os melhores valores dos hiper parâmetros ($V = 100$ e $\alpha = 10^{-5}$), atingindo um RMSE = 0.036. Note que esse valor é ligeiramente melhor do que aquele encontrado na primeira parte do exercício.

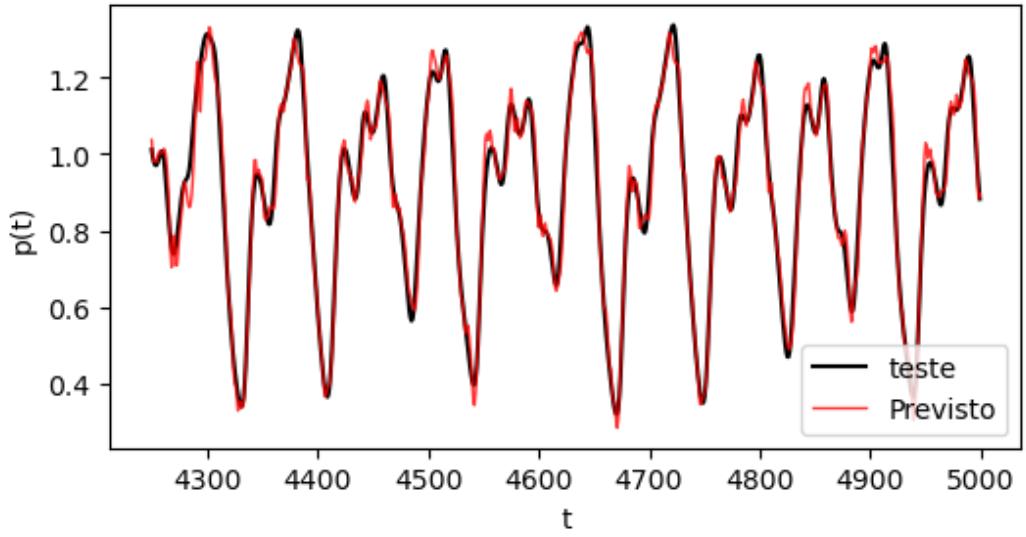

Figura 12 - Predição da série temporal com os melhores valores encontrados de V e α.