IA048 – Aprendizado de Máquina

Exercício de Fixação de Conceitos (EFC) 1 – Regressão Linear

Turma A – 2º semestre de 2022
Prof: Levy Boccato Email: lboccato@dca.fee.unicamp.br
Prof: Romis Attux Email: attux@dca.fee.unicamp.br

Introdução

Nesta atividade, vamos abordar uma instância do problema de regressão de grande interesse prático e com uma extensa literatura: a predição de séries temporais. A fim de se prever o valor futuro de uma série de medidas de uma determinada grandeza, um procedimento típico consiste em construir um modelo matemático de estimação baseado na hipótese de que os valores passados da própria série podem explicar o seu comportamento futuro.

Seja $x(n)$ o valor da série temporal no instante (discreto) n. Então, o modelo construído deve realizar um mapeamento do vetor de entradas $\mathbf{x}(n) \in \mathbb{R}^{K\times 1}$, que contém K amostras passadas considerando um horizonte de predição L, ou seja,

$$\mathbf{x}(n) = [x(n-L) \ \ldots \ x(n-L-K+1)]^T \,,$$

para uma saída $y(n)$, que representa uma estimativa do valor futuro da série $x(n)$ (que está L passos à frente do vetor de entrada $\mathbf{x}(n)$). Uma ilustração do processo de predição se encontra na Figura 1.

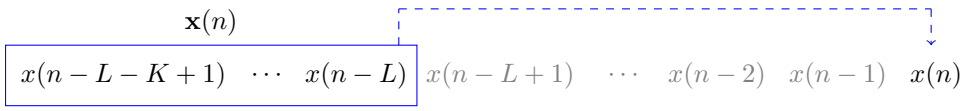

Figura 1: Predição L passos à frente da série temporal $x(n)$.

Neste exercício, vamos trabalhar com a famosa série de Mackey-Glass, a qual está associada a um sistema dinâmico contínuo, não-linear e caótico, definido pelas seguintes equações [Mackey & Glass, 1977]:

$$\frac{dP(t)}{dt} = \frac{\beta_0 \theta^n}{\theta^n + P(t-\tau)^n} - \gamma P(t) \tag{1}$$

$$\frac{dP(t)}{dt} = \frac{\beta_0 \theta^n P(t-\tau)}{\theta^n + P(t-\tau)^n} - \gamma P(t), \tag{2}$$

onde $P(t)$ denota a densidade de células brancas do sangue e β_0, θ, n, τ e γ são constantes reais relacionadas a certos parâmetros hormonais de um organismo, geralmente sendo determinadas experimentalmente.

Conforme demonstrado em [Mackey & Glass, 1977], a Equação (2) exibe comportamento caótico para valores mais elevados de τ. Além disso, os pesquisadores perceberam que regimes caóticos dessa equação estão atrelados a certos problemas fisiológicos do organismo. A Figura 2 exibe o trecho inicial da série de Mackey-Glass fornecida no arquivo '`mackeyglass.csv`'.

Primeira Parte

Inicialmente, vamos explorar um modelo linear para a previsão, tal que:

$$y(n) = \mathbf{w}^T \mathbf{x}(n) + w_0, \tag{3}$$

considerando que o horizonte de predição é $L = 7$.

Para o projeto do preditor linear, separe os dados disponíveis em dois conjuntos, um para treinamento e outro para teste. No caso, reserve as últimas 750 amostras em seu conjunto de teste (o que equivale a 15% dos dados disponíveis). Além disso, adote um esquema de validação cruzada para selecionar o melhor valor do hiperparâmetro K.

Faça a análise de desempenho do preditor linear ótimo, no sentido de quadrados mínimos irrestrito, considerando:

- A progressão do valor da raiz quadrada do erro quadrático médio (RMSE, do inglês *root mean squared error*), junto aos dados de validação, em função do número de entradas (K) do preditor (desde $K = 1$ a $K = 50$).

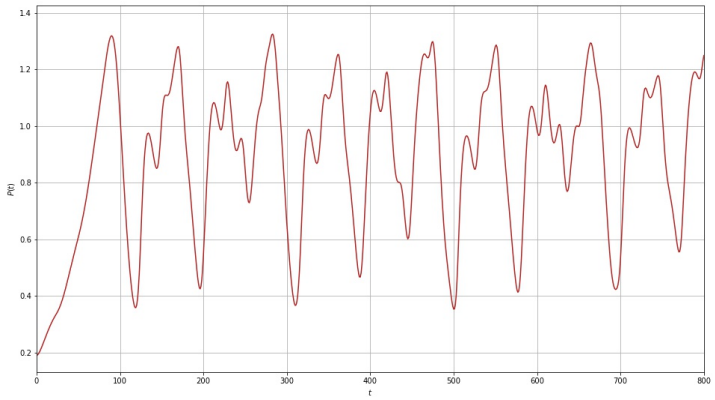

Figura 2: Trecho da série temporal associada ao sistema de Mackey-Glass ($t = 0$ a $t = 800$). Os parâmetros utilizados na simulação numérica do sistema dinâmico foram: $n = 10$, $\gamma = 0{,}1$, $\beta_0 = 0{,}2$, $\theta = 1$, $\tau = 22$, $dt = 1$ e $P(0^-) = 0{,}1$.

- O gráfico com as amostras de teste da série temporal e com as respectivas estimativas geradas pela melhor versão do preditor (i.e., usando o valor de K que levou ao mínimo erro de validação).

Observação: Neste exercício, não é necessário utilizar regularização, nem efetuar normalizações nos dados.

Segunda Parte

Agora, vamos explorar um modelo de predição linear nos parâmetros que utiliza como entrada valores obtidos a partir de transformações não-lineares do vetor $\mathbf{x}(n)$. Em outras palavras, os atributos que efetivamente são combinados linearmente na predição resultam de mapeamentos não-lineares dos atrasos da série presentes no vetor original $\mathbf{x}(n)$. No caso, vamos gerar V atributos transformados da seguinte forma:

$$x'_k(n) = \tanh\left(\mathbf{w}_k^T \mathbf{x}(n)\right), \tag{4}$$

para $k = 1, \ldots, V$, $n = 0, \ldots, N-1$, onde N denota a quantidade de amostras do conjunto de dados. Os vetores $\mathbf{w}_k$ tem seus elementos gerados aleatoriamente de acordo com uma distribuição uniforme.

Utilizando o mesmo esquema de validação cruzada, juntamente com a técnica *ridge regression* para a regularização do modelo[1]:

- Apresente o gráfico com os valores de RMSE do preditor em função do número de atributos (V) utilizados, desde $V = 1$ a $V = 100$. Neste caso, considere $K = 10$ (número de atrasos presentes no vetor $\mathbf{x}(n)$).
- Apresente o melhor valor do parâmetro de regularização obtido para cada valor de V.
- Por fim, aplique o modelo com os melhores valores de λ (regularização) e de V aos dados de teste. Meça o desempenho em termos de RMSE e mostre o gráfico com as amostras de teste da série temporal e as respectivas estimativas geradas pela melhor versão do preditor.

Curiosidade: a estrutura explorada neste exercício corresponde, na realidade, a uma rede neural conhecida como *extreme learning machine* (ELM) [Huang, Zhu & Siew, 2006].

Observações: sejam criteriosos na escolha de todos os parâmetros e justifiquem todas as opções relevantes feitas. Além disso, analisem e comentem todos os resultados obtidos.

[1] Neste exercício, é preciso levar em consideração a escala dos valores da série ao se pensar no intervalo admissível para os coeficientes aleatórios das projeções. Também é possível tratar esta questão através de normalizações. Contudo, os valores de RMSE e a exibição da série de teste estimada devem estar no domínio original do problema.

Referências

[Huang, Zhu & Siew, 2006] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, *Extreme learning machine: theory and applications*. Neurocomputing, vol. 70, pp. 489–501, 2006.

[Mackey & Glass, 1977] M. C. Mackey and L. Glass, *Oscillation and Chaos in Physiological Control Systems*, Science, vol. 197, no. 4300, pp. 287-289, 1977.