IA048 – Aprendizado de Máquina

Atividade 1 – Regressão Linear

Turma A – 1º semestre de 2024
Prof: Levy Boccato Email: lboccato@dca.fee.unicamp.br
Prof: Romis Attux Email: attux@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 [Box et al., 2015]. A fim de se prever o valor futuro de uma série relacionada a determinada informação (e.g., preço, temperatura etc.), 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 preditor 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 corresponde a 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.

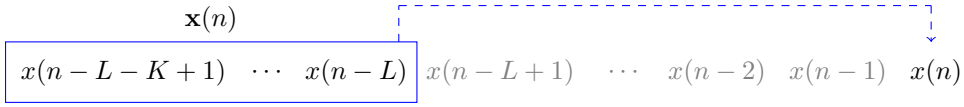

Figura 1: Predição L passos à frente da série temporal $x(n)$.

Neste exercício, vamos trabalhar com a base de dados U.S. Airline Traffic Data, a qual contém informações referentes ao tráfego aéreo mensal norte-americano no período de 2003 a 2023, disponibilizadas pelo *U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics*. Em particular, vamos explorar a série temporal do número total de vôos (domésticos e internacionais).

A análise desse tipo de série temporal pode ser útil para o planejamento das companhias aéreas e setores administrativos de aeroportos, para a elaboração de estratégias de preços e outras decisões de negócios.

Descrição da Atividade

Vamos explorar um modelo linear para a previsão, tal que:

$$y(n) = \mathbf{w}^T\mathbf{x}(n) + w_0, \tag{1}$$

considerando que o horizonte de predição é $L = 1$.

a) Exiba o gráfico da série temporal completa. Numa inspeção visual simples, é possível reconhecer ao menos três faixas distintas de comportamento aproximadamente "regular" na série: (i) Jan/2003 a Ago/2008; (ii) Set/2008 a Dez/2019; (ii) Jan/2020 a Set/2023. Discuta possíveis razões históricas / econômicas para essas transições de comportamento.

b) Divida a série em dois conjuntos: (i) treinamento e validação, com amostras de 2003 a 2019; (ii) teste, com amostras de 2020 a 2023. Faça a análise de desempenho do preditor linear ótimo, no sentido de quadrados mínimos irrestrito, considerando:

   b1) 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 $K = 24$). Apresente o gráfico obtido e busque tecer conjecturas sobre os motivos subjacentes a seu comportamento.

b2) O gráfico com as amostras de teste da série temporal e as respectivas estimativas geradas pela melhor versão do preditor (i.e., usando o valor de K que levou ao mínimo erro de validação). Obtenha, também, o RMSE e o erro percentual absoluto médio (MAPE, do inglês *mean absolute percentage error*) para o conjunto de teste.

b3) O gráfico com as amostras apenas dos dois últimos anos (2022 e 2023) e as estimativas geradas pelo melhor preditor, além dos respectivos valores de RMSE e MAPE.

c) Repita o procedimento detalhado nos itens b1) e b2), mas adotando a seguinte divisão dos dados: (*i*) treinamento – amostras de 2003 a 2019; (*ii*) validação – amostras de 2020 e 2021; (*iii*) teste, com amostras de 2022 e 2023. Discuta os resultados obtidos e faça uma comparação com o cenário anterior (especialmente com o que foi obtido no item b3).

Obs.: O nível de desempenho do preditor ótimo pode depender de fatores como pré-processamento e normalização. Cabe a cada grupo analisar a pertinência de lançar mão dessas possibilidades.

Referências

[Box et al., 2015] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung, *Time Series Analysis: Forecasting and Control*, Wiley, 5ª ed., 2015.