IA048 – Aprendizado de Máquina
Exercícios de Fixação de Conceitos (EFC) 1 – 2s2020

Parte 1 – Atividades teóricas

Exercício 1 – Considere duas variáveis aleatórias binárias X e Y, com valores possíveis iguais a 0 e 1. A distribuição conjunta dessas variáveis, $P(X,Y)$, é apresentada na tabela a seguir.

	Y = 0	**Y = 1**
X = 0	0,5	0,05
X = 1	0,3	0,15

a) Obtenha $P(X)$ e $P(Y)$.
b) Calcule $P(X = 1|Y = 1)$.
c) As variáveis são descorrelacionadas? Por quê?
d) As variáveis são estatisticamente independentes? Por quê?
e) Calcule $H(X,Y)$, $H(X)$, $H(Y)$, $H(X|Y)$ e $H(Y|X)$.
f) Calcule $I(X,Y)$.

Parte 2 – Atividade computacional

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.

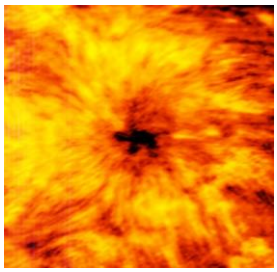
Seja $x(n)$ o valor da série temporal no instante (discreto) n. Então, o modelo construído deve realizar um mapeamento do vetor de entradas $\mathbf{x}(n) \in \mathbb{R}^{K\times 1}$, o qual é formado por um subconjunto de K amostras passadas, *i.e.*,

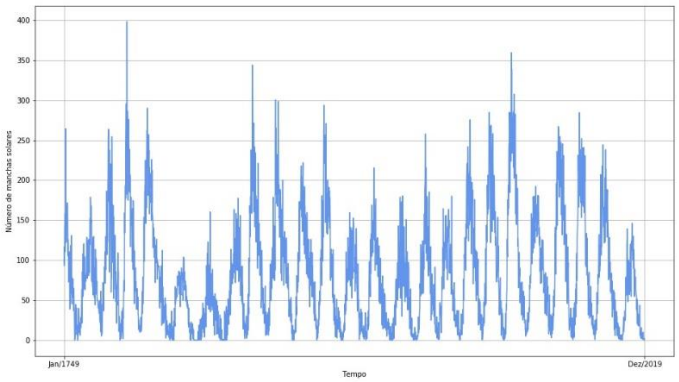
$$\mathbf{x}(n) = [x(n-1) \ \ldots \ x(n-K)]^{\mathrm{T}},$$

para uma saída $\hat{y}(n)$, que representa uma estimativa do valor futuro da série $x(n)$[*].

Neste exercício, vamos trabalhar com a famosa série histórica de medidas do número de manchas solares (*sunspots*). No caso, dispomos das leituras mensais desde 1749 a 2019, totalizando 3252 amostras. A Figura 1 exibe um registro de mancha solar juntamente com o gráfico da série completa.

[*] Esta modelagem está pressupondo o caso em que desejamos prever o valor da série um passo à frente.





(a) (b)

Figura 1. Em (a), temos um exemplo de uma mancha solar observada pelo Atacama Large Millimeter / Submillimeter Array (ALMA). Em (b), os valores mensais do número de manchas solares desde 1749 a 2019.

Exercício 1

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

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

Para o projeto do preditor linear, separe os dados disponíveis em dois conjuntos, um para treinamento e outro para teste. No caso, reserve as amostras referentes aos últimos dez anos (2010-2019) em seu conjunto de teste. Além disso, utilize um esquema de validação cruzada do tipo k*-fold* para selecionar o melhor valor do hiperparâmetro K.

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

1. A progressão do valor médio 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 $K = 24$).
2. 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.

Exercício 2

Agora, vamos explorar um modelo de predição linear 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 T atributos transformados da seguinte forma:

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

para $k = 1, \ldots, T$, $n = 1, \ldots, N$. Os vetores $\mathbf{w}_k$ tem seus elementos gerados aleatoriamente de acordo com uma distribuição uniforme.

Curiosidade: a estrutura explorada neste exercício corresponde, na realidade, a uma rede neural conhecida como *extreme learning machine* (ELM).

- Huang, G.-B., Zhu, Q.-Y., e Siew, C.-K. (2006). *Extreme learning machine: theory and applications*. Neurocomputing, 70, 489–501.

Utilizando um esquema de validação cruzada do tipo *k-fold*, juntamente com a técnica *ridge regression* para a regularização do modelo:

a) Apresente o gráfico com a média dos valores de RMSE do preditor em função do número de atributos (T) utilizados, desde $T = 1$ a $T = 100$. Neste caso, considere $K = 8$ (número de atrasos presentes no vetor $\mathbf{x}(n)$).
b) Apresente o melhor valor do parâmetro de regularização obtido para cada valor de T.
c) Por fim, aplique o modelo com os melhores valores de λ (regularização) e de T 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.

Observação: 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 ser referentes ao domínio original do problema.

Considerações gerais:

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