IA006 – Exercícios de Fixação de Conceitos
EFC 1 – 1s2019

Parte 1 – Atividades teóricas

Exercício 1. Dado que $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{4}$ e $P(AB) = \frac{1}{6}$, determine:

a) $P(A^C)$
b) $P(A^C \cup B)$
c) $P(A \cup B^C)$
d) $P(AB^C)$
e) $P(A^C \cup B^C)$

Exercício 2. Considere uma variável aleatória contínua com função densidade de probabilidade uniforme entre 0 e 2.

a) Deduza a fórmula da função cumulativa.
b) Calcule $E\{X\}$, $E\{X^2\}$ e $E\{X^3\}$.

Exercício 3. Sejam duas variáveis aleatórias X_1 e X_2 com funções de massa de probabilidade $P_1(\cdot)$ e $P_2(\cdot)$, definidas sobre o mesmo domínio de valores (X_1,X_2 = {0, 1, 2, 3}):

X_1/X_2	0	1	2	3
$P_1(\cdot)$	0,1	0,2	0,3	0,4
$P_2(\cdot)$	0,25	0,25	0,25	0,25

a) Apenas inspecionando as probabilidades, aponte qual variável tem maior entropia (não faça cálculos). Explique seu raciocínio.
b) Calcule $H(X_1)$ e $H(X_2)$.
c) Calcule $D(P_1||P_2)$ e $D(P_2||P_1)$.

Exercício 4. Considere que disponhamos de um conjunto de N amostras x_i, $i = 1, \dots, N$, obtidas independentemente. Essas amostras são geradas segundo uma densidade gaussiana com média μ e variância σ^2. Deseja-se obter um estimador para a média μ através da abordagem de máxima verossimilhança.

a) Apresente a expressão da verossimilhança (*likelihood*) $p(\mathbf{x}; \mu)$ ou $p(\mathbf{x}|\mu)$ considerando o caso escalar (uma amostra).
b) Apresente agora a verossimilhança para o caso de N amostras independentes.
c) Obtenha o estimador de μ que maximiza o logaritmo da verossimilhança obtida em b). Comente.

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.

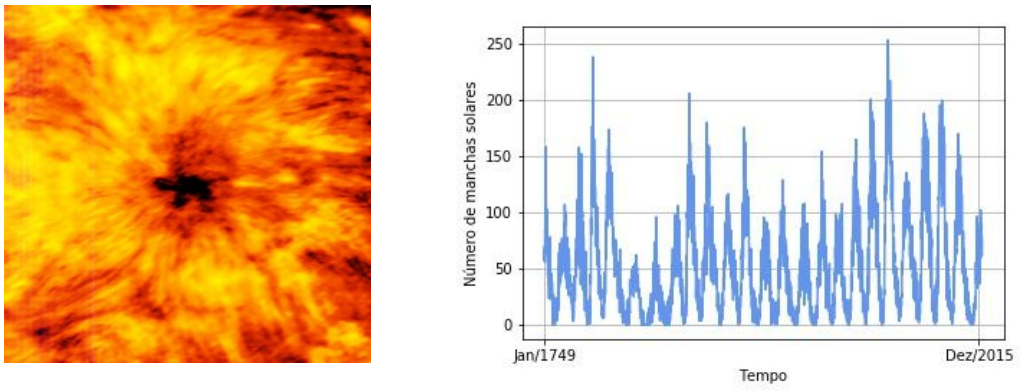
Seja $x(n)$ o valor da série temporal no instante (discreto) n. Então, o modelo construído deve realizar um mapeamento do vetor de entradas $\mathbf{x}(n) \in \mathbb{R}^{K\times 1}$, o qual é formado por um subconjunto de K amostras passadas, *i.e.*,

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

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 (*sunspot*). No caso, dispomos das leituras mensais desde 1749 a 2014, totalizando 3458 amostras. Além disso, o modelo que será empregado na previsão corresponde a um preditor linear, tal que:

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



(a) (b)

Figura. 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 2014. Dados obtidos a partir de:
http://www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries/Data/sunspot.long.data

(a) Tendo acesso à série de medidas mensais de manchas solares (arquivo sunspot.txt), realize o projeto de um preditor linear. Para isto, separe os dados disponíveis em dois conjuntos – um para treinamento e outro para teste. No caso, reserve as amostras referentes aos cinco (5) anos mais recentes (2010 a 2014) em seu conjunto de teste. Utilize $K = 20$ variáveis de entradas, as quais correspondem a 20 atrasos consecutivos, na ordem do mais recente ao mais antigo (i.e., a entrada 1 corresponde ao valor da série no mês anterior).

Analise o desempenho do preditor linear ótimo, no sentido de quadrados mínimos irrestrito, considerando:

1. A raiz quadrada do erro quadrático médio (RMSE, do inglês *root mean squared error*) junto aos dados de treinamento e de teste.

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

2. O gráfico com as amostras de teste da série temporal e com as respectivas estimativas geradas pelo preditor.

(b) Agora, vamos considerar a aplicação da estratégia de seleção de variáveis denominada *wrapper*, seguindo a abordagem *backward elimination*. Ademais, a fim de medirmos de forma mais robusta o desempenho de cada subconjunto de variáveis de entrada, vamos utilizar um esquema de validação cruzada do tipo k*-fold*. Por fim, vamos também considerar a técnica *ridge regression* para a regularização do modelo.

Partindo do conjunto completo de entrada com $K = 20$ atrasos candidatos, mostre a progressão do RMSE mínimo em função do número de variáveis retiradas pelo *wrapper*. Indique qual o conjunto "ótimo" de atrasos identificado pelo *wrapper*.

Apresente também o melhor valor do parâmetro de regularização obtido para cada conjunto de entradas descoberto pelo *wrapper*.

(c) Por fim, vamos aplicar uma metodologia do tipo filtro para selecionar o mesmo número "ótimo" de variáveis de entrada segundo algum critério (*e.g.*, correlação de Pearson). Utilize novamente um esquema de validação cruzada do tipo k *-fold* e meça o desempenho médio do preditor linear com regularização neste caso. Indique o melhor valor obtido para o parâmetro de regularização. Comente os resultados obtidos.