IA006 – Exercícios de Fixação de Conceitos
EFC 1 – 2s2019

Parte 1 – Atividades teóricas

Exercício 1 – A tabela a seguir apresenta a distribuição conjunta de duas variáveis aleatórias binárias X e Y.

$\boldsymbol{X}$ / $\boldsymbol{Y}$	$\boldsymbol{Y}$ = **0**	$\boldsymbol{Y}$ = **1**
$\boldsymbol{X}$ = **0**	1/6	3/8
$\boldsymbol{X}$ = **1**	1/8	1/3

a) Obtenha $P(X)$ e $P(Y)$.
b) Calcule $P(X = 0 \mid Y = 0)$.
c) Calcule $E[X]$ e $E[Y]$.
d) As variáveis são independentes? Por quê?

Exercício 2 - A tabela a seguir apresenta a distribuição conjunta de duas variáveis aleatórias binárias X e Y.

$\boldsymbol{X}$ / $\boldsymbol{Y}$	$\boldsymbol{Y}$ = **0**	$\boldsymbol{Y}$ = **1**
$\boldsymbol{X}$ = **0**	0	1/4
$\boldsymbol{X}$ = **1**	3/8	3/8

a) Calcule $H(X)$, $H(Y)$ e $H(X,Y)$.
b) Calcule $H(X|Y)$ e $H(Y|X)$.
c) Calcule $I(X,Y)$.

Exercício 3 – Considere um problema de classificação com duas classes, C_1 e C_2. Há um único atributo envolvido (o atributo x).

a) Os dados da classe C_1 seguem uma densidade gaussiana de média -1 e variância 1. Por sua vez, os dados da classe C_2 seguem uma densidade gaussiana de média +1 e variância 1. Quais seriam as regiões de decisão do classificador de máxima verossimilhança (os valores de x para os quais se decide pela classe C_1 e os valores de x para os quais se decide pela classe C_2)?

b) Suponha agora que $P(C_1)$ = 0,7 e $P(C_2)$ = 0,3. Quais serão as regiões de decisão para o classificador MAP?

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 série de temperatura mínima diária referente à cidade de Melbourne, Austrália, no período de 1981 a 1990. As observações estão em graus Celsius e há 3650 amostras no total. Os dados são creditados ao *Australian Bureau of Meteorology*.

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 ao ultimo ano (1990) 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 $K = 30$).
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 de transformações não-lineares do vetor $\mathbf{x}(n)$. Em outras palavras, os atributos que efetivamente são linearmente combinados 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:

$$\mathbf{x}'_k(n) = \tanh\big(\mathbf{w}_k\mathbf{x}(n)\big),$$

para $k = 1, \ldots, T$. Os vetores $\mathbf{w}_k$ tem componentes gerados aleatoriamente conforme uma distribuição uniforme.

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

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 que o vetor original $\mathbf{x}(n)$ possui $K = 5$ valores atrasados da série.
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.
- **NÃO** será permitido o uso de pacotes prontos (*scikit-learn*, *toolboxes* do Matlab, etc.) na parte computacional. A ideia é que cada aluno desenvolva os próprios programas que implementem todas as etapas da atividade, incluindo a obtenção da solução ótima para regressão linear (sem e com regularização), a validação cruzada, etc.

Instruções para a entrega do trabalho

Cada aluno deve entregar por e-mail um **único documento**, em formato PDF, contendo as soluções dos exercícios teóricos e computacionais.

- ✓ **Título do e-mail:** Entrega_EFC1_IA006
- ✓ **Título do documento PDF:** meu_RA.pdf (e.g., 034021.pdf)
- ✓ Enviar para lboccato@dca.fee.unicamp.br, attux@dca.fee.unicamp.br