Trabalho 1: Regressão Linear com Múltiplas Variáveis
Reconhecimento de Padrões – 2021/1

Nesse trabalho sobre regressão linear múltipla, utilizaremos a base de dados hospital do matlab (disponível na pasta Bases de Dados como hospital.xls), com 4 variáveis de predição (sexo, idade, peso, fumante/não fumante), para predizer o valor da pressão máxima de um paciente. Essa base de dados contém 100 observações. Para tanto é desejável que:

1) Seja empregada validação cruzada com 5 pastas;
2) Sejam determinados os valores dos coeficientes de Pearson e o valores dos erros médio quadráticos para cada pasta (5 ao total);
3) Utilizando os valores obtidos em cada pasta determine o valor médio do coeficiente de Pearson e do erro médio quadrático.

Sugestão para elaboração de relatório:

1. Introdução
   Expor o problema a ser resolvido
2. Fundamentação Teórica
   Mostrar o regressor de múltiplas variáveis utilizando a pseudo-inversa
3. Metodologia
   Descrever como o experimento será feito utilizando as 5 pastas: quantos registros cada pasta terá, como os resultados serão avaliados, etc
4. Resultados
   Apresentar os 5 valores do coeficiente de Pearson e do erro médio quadrático, bem como o valor médio dos mesmos.
5. Conclusões
   Avaliar os resultados obtidos.

Fazer o upload no sistema de EAD até 05/04/2021

Apêndice: Método da PseudoInversa

Suponha que em um problema estejam disponíveis *p* variáveis para serem utilizadas na predição do valor de uma variável *y* de saída. Por exemplo, na determinação da pressão sanguínea de um paciente, podem ser utilizadas as variáveis; sexo, idade, peso e a condição de ser fumante ou não. Na regressão múltipla, a variável a ser predita é expressa através de uma equação linear:

$$y_k = x_{k1}w_1 + x_{k2}w_2 + \cdots + x_{kp}w_p \qquad (1)$$

$$y_k = \boldsymbol{w^T x_k}$$

Em que:

$\boldsymbol{x_k} - \ padrão\ _k$

$\mathbf{w}$ – vetor de pesos

$y_k - valor\ a\ ser\ predito$

Para *N* observações, temos o seguinte conjunto com *N* equações:

$$y_1 = x_{11}w_1 + x_{12}w_2 + \cdots + x_{1p}w_p + k$$
$$y_2 = x_{21}w_1 + x_{22}w_2 + \cdots + x_{2p}w_p y_2 + k$$
$$..........$$
$$..........$$
$$y_N = x_{N1}w_1 + x_{N2}w_2 + \cdots + x_{Np}w_p + k \qquad (2)$$

O Sistema linear de (2) é resumido a seguir:

$$\mathbf{y} = X\boldsymbol{w} \qquad (3)$$

E que:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} & 1 \\ x_{21} & x_{22} & \vdots & x_{2p} & 1 \\ \dots & \dots & & \dots & \\ x_{N1} & x_{N2} & \dots & x_{Np} & 1 \end{pmatrix} \qquad (4)$$

$$\boldsymbol{w} = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ w_p \\ k \end{pmatrix} \qquad (5)$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ y_N \end{pmatrix} \qquad (6)$$

Se o número de variáveis p for igual ao número de observações N, o sistema linear mostrado em (3) pode ser solucionado através da matriz inversa de *X*:

$$\boldsymbol{w} = X^{-1}.\boldsymbol{y} \qquad (7)$$

Normalmente, a dimensão de *p* é muito menor do que a dimensão de *N*, de tal forma que o sistema linear mostrado em (3) não pode ser resolvido por (7). Conforme demonstrado em sala de aula, para minimizar o erro médio quadrático mostrado em (8), a solução é dada pela equação (9):

$$E = \sum_{i=1}^{64} \|y_i - X_i.\boldsymbol{w}\|^2 \qquad (8)$$

$$\boldsymbol{w} = X^{+}.\boldsymbol{y} \qquad (9)$$

Em que:

$X^{+} - matriz\ pseudo\ inversa\ (PI) = (X^{T}.X)^{-1}X^{T}$ (10)

$X_i = (x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip} \quad 1)$ (11)

Com base na matriz PI o seguinte método de previsão da variável *y* pode ser proposto:

1. Define um conjunto de treinamento com *N* observações;
2. Para esse conjunto determine a matriz PI através de (10);
3. Determinar os pesos através de (9)
4. Utilizar a matriz PI para a predição através da equação $y = \boldsymbol{w^T x}$