Inteligência Computacional

Regressão Linear Múltipla

Slides adaptados do material disponibilizado

pelo **Prof. Dr. Guilherme de Alencar Barreto (UFC)**

Regressão Múltipla

- Muitos problemas de regressão envolvem mais de uma variável regressora.
- Tais modelos são chamados de modelos de regressão múltipla.
- Em geral, a variável de saída ou resposta, y, pode ser relacionada a k variáveis de entrada.
- O modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \qquad (1)$$

é chamado às vezes de regressão linear múltipla com k variáveis de entrada.

Regressão Múltipla

- Os parâmetros j , $j = 0, 1, . . . , k$, são chamados de coeficientes de regressão.
- O modelo da Eq. (1) descreve um hiperplano no espaço k-dimensional das variáveis de entrada $\{x_j\}$.

Conceito Importante!

O parâmetro β_j representa a mudança esperada na resposta y por unidade de mudança em x_j, quando todas as demais variáveis independentes x_i $(i \neq j)$ são mantidas constantes.

Regressão Múltipla

- Modelos de regressão linear múltipla são usados, em geral, como funções aproximadoras ou interpoladoras.
- Ou seja, a verdadeira relação funcional entre y e x_1, x_2, ..., x_k é desconhecida, mas dentro de certos limites das variáveis de entrada o modelo de regressão linear é uma aproximação adequada.
- Modelos mais complexos que o da Eq. (1) também podem ser analisados pelas técnicas de regressão linear múltipla.

Regressão Múltipla

- Por exemplo, considere o modelo de regressão linear múltipla com três variáveis de entrada:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon, \qquad (2)$$

- Se fizermos $x_1 = x$, $x_2 = x^2$ e $x_3 = x^3$, então o modelo da Eq. (2) pode ser escrito como um modelo não-linear (no caso, polinomial cúbico) em uma variável de entrada:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon, \qquad (3)$$

Regressão Múltipla

- O método dos mínimos quadrados pode ser usado para estimar os coeficientes de regressão $\{\beta_j\}$, $j = 0, 1, ..., k$.
- Para isso, faremos as seguintes definições:
  1. x_{ij} é i-ésima observação da variável x_j .
  2. y_i é a i-ésima observação (medida) da variavel de saída.
- As seguintes suposições são também necessárias:
  1. Estão disponíveis $n > k$ observações (i.e., há mais equações do que incógnitas).
  2. O erro ou ruído no modelo (ε) tem média 0 e variância σ_ε^2.
  3. As observações $\{\varepsilon_i\}$ são não-correlacionadas.

Regressão Múltipla

- Feito isto, podemos escrever o modelo da Eq. (33) em termos das observações:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \qquad (4)$$

para $i = 1, 2, \ldots, n$.

- Isto equivale a ter o seguinte sistema com n equações e $k + 1$

incógnitas:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1, \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \varepsilon_2, \qquad (5) \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n,
\end{aligned}$$

Regressão Múltipla

Em forma matricial, o sistema de equações em (5) é escrito

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{6}$$

em que

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}_{n\times(k+1)},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}.$$

Regressão Múltipla

- Deseja-se encontrar o vetor de estimativas dos quadrados mínimos, $\hat{\beta}$, que minimize a seguinte função-custo:

$$J(\beta) = \|\varepsilon\|^2 = \varepsilon^T\varepsilon = \sum_{i=1}^{n} \varepsilon_i^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta). \qquad (7)$$

- A função-custo $J(\beta)$ pode ser entendida como uma função que busca encontrar o vetor de parâmetros $\hat{\beta}$ que produz o vetor ε de menor norma quadrática.
- A Eq. (7) pode ser decomposta em

$$\begin{aligned} J(\beta) &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \\ &= \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta \end{aligned} \qquad (8)$$

uma vez que $\beta^T\mathbf{X}^T\mathbf{y}=\mathbf{y}^T\mathbf{X}\beta$ resulta no mesmo escalar.

Regressão Múltipla

- As estimativas de quadrados mínimos devem satisfazer

$$\frac{\partial J(\beta)}{\partial \beta} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta = \mathbf{0}, \qquad (9)$$

em que **0** é um vetor de zeros.

- Simplificando a Eq. (9) resulta em:

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{y}. \qquad (10)$$

- A Eq. (10) define as equações normais dos quadrados mínimos da regressão linear múltipla.

Regressão Múltipla

- Note que a matriz $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ é quadrada (dim=$(k + 1) \times (k + 1)$).
- Para resolver as equações normais basta multiplicar ambos os lados da Eq. (10) pela inversa de $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.
- Assim, a estimativa de quadrados mínimos de β é dada por

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{11}$$

- Portanto, o modelo de regressão ajustado (preditor) é definido

como

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}. \tag{12}$$

- O vetor de erros de predição (resíduos) é denotado por

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}. \tag{13}$$

Regressão Múltipla

Regularização de Thikonov

- Muitas vezes a matriz $\mathbf{X}^T\mathbf{X}$ é quase singular, ou seja

$$\det(\mathbf{X}^T\mathbf{X}) \approx 0$$

- Isso certamente causará problemas numéricos durante a inversão desta matriz.
- Isto ocorre geralmente quando as variáveis de entrada são intercorrelacionadas.
- Quando essa intercorrelação é grande, dizemos que existe *multicolinearidade*, ou seja, as linhas da matriz $\mathbf{X}^T\mathbf{X}$ não são linearmente independentes.

Regressão Múltipla

Regularização de Thikonov

- Os efeitos nocivos da multicolinearidade podem ser minimizados reescrevendo a Eq. (43) como

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{46}$$

  em que
  - $0 \leq \lambda \ll 1$ é uma constante de valor pequeno.
  - $\mathbf{I}$ é uma matriz identidade de dimensão $(k+1) \times (k+1)$.
- A técnica da Eq. (46) é chamada de **regularização de Tikhonov**, enquanto a regressão que a utiliza é chamada de **regressão de cumeeira** (*ridge regression*).

Regressão Múltipla

- Medida de adequação do modelo na regressão linear múltipla

Coeficiente de Determinação na Regressão Múltipla

- O coeficiente de determinação R^2 também é usado na regressão múltipla como medida de adequação do modelo:

$$R^2 = 1 - \frac{SQ_E}{S_{yy}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (48)$$

em que $0 \leq R^2 \leq 1$.

- No entanto, um valor alta de R^2 não implica que o modelo seja bom!
- O acréscimo de uma variável ao modelo causará sempre, um aumento em R^2, independentemente de a variável adicional ser ou não significante (informativa).

Regressão Múltipla

- Medida de adequação do modelo na regressão linear múltipla

Coeficiente de Determinação Ajustado

- Alguns autores preferem usar o *coeficiente de determinação* R^2 *ajustado* (R^2_{aj}):

$$R^2_{aj} = 1 - \frac{SQ_E/(n-p)}{S_{yy}/(n-1)}, \tag{49}$$

  em que $p = k + 1$.
- O valor $S_{yy}/(n-1)$ será constante, independente do número de variáveis no modelo.
- O valor $SQ_E/(n-p)$ é a média quadrática para o erro, que mudará com o acréscimo (ou retirada) de variáveis ao modelo.
- Portanto, R^2_{aj} crescer'apenas se a adição de um novo termo reduzir significantemente a média quadrática dos erros.

Regressão Múltipla
(Regressão Polinomial)

- O modelo linear $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, é um modelo geral que pode ser usado para ajustar qualquer relação que seja linear nos parâmetros desconhecidos .
- Isso inclui a importante classe dos modelos de regressão polinomial. Por exemplo, vimos que o modelo polinomial cúbico em uma variável de entrada:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon,$$

é um tipo de modelo de regressão múltipla se fizermos $x_1 = x$, $x_2 = x^2$ e $x_3 = x^3$.

- Modelos de regressão polinomial são amplamente usados nos casos em que a relação entre a variável de saída e de entrada é curvilínea (i.e. não-linear).

Regressão Múltipla
(Regressão Polinomial)

- Em regressão polinomial, a matriz **X** do modelo linear **y** = **X**β+ε passa ser definida como

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix}_{n\times(k+1)}$$

em que x_i é a *i*-ésima observação da variável de entrada.

- A estimativa de quadrados mínimos é então calculada por meio da Eq. (11).
- Predições de novos valores podem ser feitas por meio da Eq. (12) e resíduos são calculados pela Eq. (13).

Regressão Múltipla
(Regressão Polinomial)

- Usando os dados do aerogerador ajustou-se o seguinte modelo polinomial de quarta ordem ($k = 4$):

$$\hat{y} = -0.391 + 10.37x - 5.00x^2 + 1.43x^3 - 0.068x^4$$

com $R^2 = 0.974$. A curva do modelo superposto ao gráfico de dispersão é mostrada abaixo.