MINERAÇÃO DE DADOS

Thiago Marzagão[1]

[1] marzagao.1@osu.edu

REGRESSÃO LINEAR MÚLTIPLA

regressão linear múltipla

- Aula passada: $\hat{y}_i = a + bx_i$
- Hoje: $\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} + ...$

regressão linear múltipla

- Aula passada: $\hat{y}_i = a + bx_i$ (regressão simples)
- Como encontrar b?
- $\min_{a,b} \sum_{i=1}^{N} e_i^2 = \min_{a,b} \sum_{i=1}^{N} (y_i - a - bx_i)^2$
- Hoje: $\hat{y}_i = a + b_1 x_{1i} + b_2 x_{2i} + ...$ (regressão múltipla)
- Como encontrar b_1, b_2, etc?
- $\min_{a,b_1,b_2,...} \sum_{i=1}^{N} e_i^2 = \min_{a,b_1,b_2,...} \sum_{i=1}^{N} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - ...)^2$

regressão linear múltipla

Em notação matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & ... \\ 1 & x_{12} & x_{22} & ... \\ ... & ... & ... & ... \\ 1 & x_{1N} & x_{2N} & ... \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_N \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ ... \\ e_N \end{bmatrix}$$

Erros:

$\boldsymbol{e = y - xb}$

Erros quadrados:

$\boldsymbol{e^2 = (y - xb)^2}$

$\boldsymbol{e'e = y'y - 2bx'y + b'x'xb}$

O que queremos minimizar:

$\boldsymbol{\min_b \sum e'e = \min_b \sum y'y - 2bx'y + b'x'xb}$

a mágica do cálculo matricial acontece...

regressão linear múltipla

- $\boldsymbol{b} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}$
- (Vale p/ regressão simples também!)

regressão linear múltipla

- $\boldsymbol{b} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}$
- O "menos um" é notação p/ matriz inversa.
- Se $\boldsymbol{AB} = \boldsymbol{BA} = \boldsymbol{I}$, então $\boldsymbol{B}$ é a matriz inversa de $\boldsymbol{A}$
- $\boldsymbol{I}$ é uma matriz identidade (1s na diagonal principal e 0s nas demais células)
- Apenas matrizes quadradas (n x n) têm matrizes inversas.

crime per capita nos bairros de Boston

crime	n. de cômodos	% $<$ 1940	distância	$ valor
0,00632	6,575	65,2	4,0900	24.000
0,02731	6,421	78,9	4,9671	21.600
0,02729	7,185	61,1	4,9671	34.700
0,03237	6,998	45,8	6,0622	33.400
0,06905	7,147	54,2	6,0622	36.200
0,02985	6,430	58,7	6,0622	28.700
0,08829	6,012	66,6	5,5605	22.900
0,14455	6,172	96,1	5,9505	27.100
...	...	...	...	...

fonte: http://archive.ics.uci.edu/ml/datasets/Housing

crimes per capita nos bairros de Boston

- Impossível plotar c/ mais de 2 variáveis.
- Mas idéia é a mesma da aula passada (sorvete vs temperatura), apenas com mais dimensões.
- (Desenhar no quadro.)
- Já a interpretação dos coeficientes muda um pouco.
- Regressão simples: y varia em b unidades quando x varia uma unidade.
- Regressão múltipla: y varia em b_1 unidades quando x_1 varia uma unidade e x_2, x_3, etc, são mantidos constantes.
- Essa é a beleza da regressão múltipla: ela nos permite calcular o coeficiente de uma variável *mantendo as demais variáveis constantes*.
- Em outras palavras, a regressão múltipla nos permite calcular o efeito *líquido* de x_1 sobre y, i.e., já descontados os efeitos de x_2, x_3, etc.
- Regressão múltipla != múltiplas regressões simples.

crimes per capita nos bairros de Boston

- x_1: n. de cômodos
- x_2: % $<$ 1940
- x_3: distância
- x_4: valor
- $\hat{y} = 7,317 + 1,420x_1 + 0,003x_2 - 1,217x_3 - 0,365x_4$
- Interpretação:
- P/ cada cômodo adicional, crime per capita aumenta 1,420 (mantendo as outras três variáveis constantes).
- P/ cada 1 ponto percentual a mais de imóveis $<$ 1940, crime per capita aumenta 0,003 (mantendo as outras três variáveis constantes).
- P/ cada 1 milha a mais de distância do centro, crime per capita diminui 1,217 (mantendo as outras três variáveis constantes).
- P/ cada US$ 1 mil dólares a mais de valor da propriedade, crime per capita diminui 0,365 (mantendo as outras três variáveis constantes).

crimes per capita nos bairros de Boston

- Dataset está disponível no site da disciplina (http://thiagomarzagao.com/teaching/EPA109738).
- P/ reproduzir os resultados acima:

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression

data = pd.read_csv('aula5dataset.csv')
reg = LinearRegression()
x = data[['RM', 'AGE', 'DIS', 'MEDV']]
y = np.array(data['CRIM']).reshape(len(data), 1)
reg.fit(x, y)
reg.intercept_
reg.coef_
```

mineração de dados vs econometria

- Econometria: foco é na análise dos coeficientes.
- Além de estimar os coeficientes computam-se intervalos de confiança p/ eles.
- Perguntas centrais em econometria: x_1 e y realmente estão relacionados? qual a magnitude dessa relação?
- Mineração de dados: foco é na predição de y.
- Pergunta central em mineração de dados: dados x_1, x_2, etc, qual o valor esperado de $\hat{y}$?
- Coeficientes? Efeito de x_1 sobre $\hat{y}$?

mineração de dados

- Dados de treinamento vs dados de teste.
- (Dar exemplos.)
- Lógica:
- ... estimar modelo usando dados de treinamento
- ... testar modelo usando dados de teste
- Fundamental: dados de teste NÃO devem ser usados p/ *estimar* o modelo, apenas p/ *testar* o modelo.
- Objetivo é testar o modelo com dados ainda não vistos.

competição # 1

- Dataset de treino e documentação: no site da disciplina.
- Objetivo: prever o valor da variável 'ViolentCrimesPerPop' (crimes violentos por 100 mil habitantes) c/ menor soma dos erros quadrados possível.
- O que você vai entregar: código Python completo: desde *import ...* até *reg.fit()*
- A seu critério:
- ... quais variáveis independentes usar (menos 'ViolentCrimesPerPop', naturalmente)
- ... qual a forma funcional de cada uma dessas variáveis (exemplo: educação: anos de educação ou maior grau obtido?)
- ... normalizar os dados ou não
- Seu modelo será avaliado contra um dataset de teste (N = 500) que só será divulgado depois.

competição # 1 (cont.)

- "Hein?? Mas como eu sei se meu modelo está bom ou ruim??"
- Resposta: *reg.score(x, y)*
- Isso te dá R^2, que é o % da variância de y explicada pelo modelo.
- $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$
- Varia entre 0 (modelo não explica nada) e 1 (modelo explica tudo).
- *Em geral* quanto maior o R^2 melhor o modelo. Mas...
- ... R^2 sempre aumenta quando incluímos mais variáveis independentes no dataset de treino, mesmo que essas variáveis piorem o R^2 do modelo no dataset de teste
- ... o nome disso é overfitting: nós estamos forçando o modelo a responder a cada idiossincrasia do dataset de treino, em possível prejuízo do desempenho do modelo no dataset de teste
- ... nós veremos overfitting em mais detalhes quando chegarmos em classificação (árvores de decisão e SVM)

o que *não* vamos ver

- Inferência (b é estatisticamente diferente de zero?, etc; vide slide acima).
- ... inferência nos permite calcular intervalos de confiança p/ $\hat{y}$, o que é muito útil em mineração de dados; mas não há tempo p/ cobrirmos isso e vocês precisariam saber um bocado de estatística
- Relações não aditivas (ex.: $\hat{y} = a + b_1 x_1 + b_2 x_1^2$).
- Séries temporais (ex.: $\hat{y}_t = a + b_1 y_{t-1} + b_2 y_{t-2}$).
- Dados em painel (ex.: $\hat{y}_{it} = a + b_1 x_{1it} + b_2 x_{2it}$).
- Causalidade (x causa y? y causa x? z causa x e y?).
- Uma infinidade de outros tópicos!
- P/ quem quiser aprofundar em regressão:
- ... comecem com o Gujarati (Econometria Básica); excelente p/ self-learning
- ... depois partam p/ o Greene (Econometrics); é um tratamento menos didático mas mais avançado