VOZES DIVERSAS
DIFERENTES SABERES

CONHECIMENTO FORMAÇÃO INOVAÇÃO
Salão UFRGS 2018

SALÃO DE INICIACAO CIENTIFICA
XXX SIC

15 A 19 OUTUBRO
CAMPUS DO VALE

UFRGS
PROPESQ

Análise de Algoritmos de Regressão Linear para Aplicação voltada à Análise Empírica da Qualidade de Rochas Reservatório

Júlia Eidelwein *{jeidelwein@inf.ufrgs.br}*

Introdução

A modelagem empírica gera um modelo matemático para descrever as relações de causa e efeito entre as variáveis de entrada e de saída de um conjunto de dados.

A predição de valores busca ajustar uma função matemática para explicar o comportamento dos dados sobre os quais ela é aplicada. A função é utilizada para prever o valor de saída para quaisquer novos dados oferecidos na entrada.

Objetivo

Identificar o algoritmo de regressão linear multivariada apropriado para a estimação da qualidade de um reservatório (capacidade de armazenamento e liberação de hidrocarbonetos), encontrando as feições petrográficas que explicam os valores de porosidade das rochas.

Algoritmos analisados

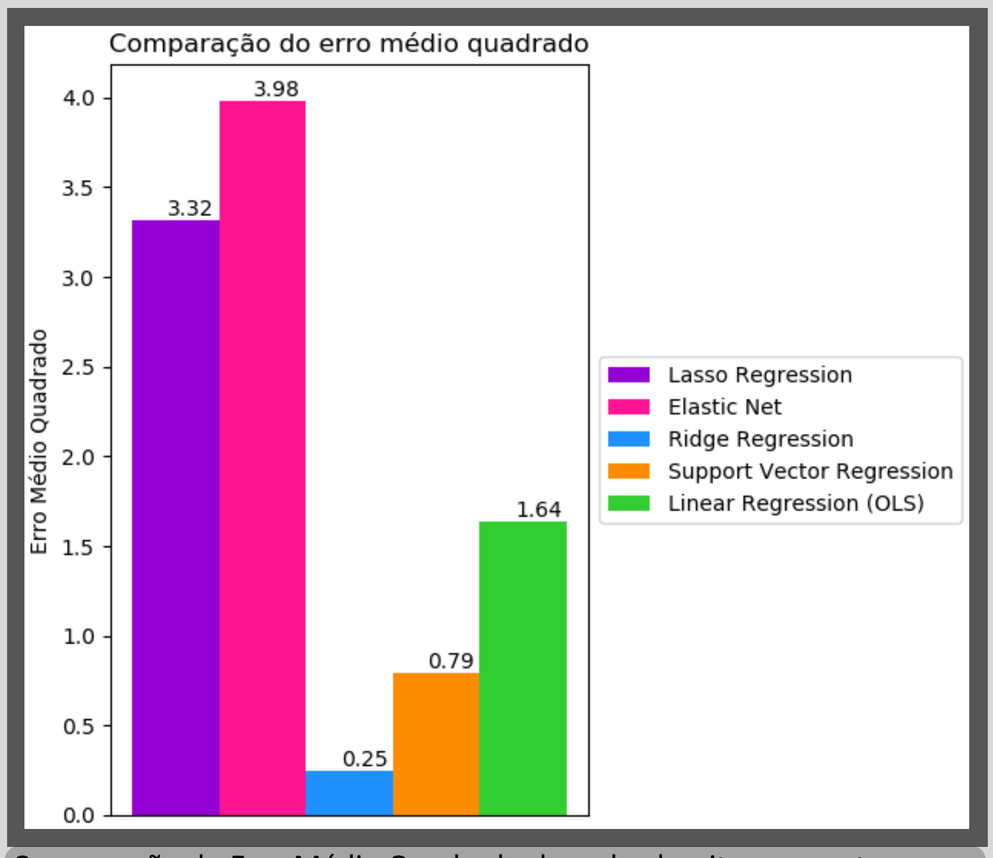
Foram analisados, no total, cinco algoritmos de regressão, todos buscando minimizar suas respectivas funções de erro: Regressão Linear pelo Método dos Mínimos Quadrados Ordinários, *Ridge Regression*, *Support Vector Regression*, *Lasso Regression* e *Elastic Net*. Sendo que estes dois últimos também procuram reduzir o número de features utilizadas para a predição.

A comparação entre os algoritmos foi feita por meio da métrica do Erro Médio Quadrado.

Na tabela abaixo:

- $w = (w_1, w_2, \dots, w_p)$ é o vetor que contém os p coeficientes do modelo
- X é o conjunto de p-1 feições explicativas
- y é o valor esperado da variável explicada
- α é o parâmetro de complexidade que penaliza o tamanho dos coeficientes
- $n_{samples}$ é o número de instâncias de treinamento
- ρ é o parâmetro de mistura, variando de 0 a 1,

onde 1 resulta na penalidade ℓ_1 e 0 resulta na penalidade ℓ_2



Comparação do Erro Médio Quadrado de cada algoritmo: quanto menor o valor, melhor o ajuste do modelo aos dados.

Conclusão

Como esperado, os modelos que penalizam uma grande quantidade de features utilizadas (*Lasso* e *Elastic Net*) possuem um erro médio quadrado mais elevado. Todavia, sua utilização não deve ser prontamente descartada devido à sua característica de redução de features: apenas as variáveis mais importantes são mantidas pelo modelo, propiciando uma primeira análise ágil.

Pela perspectiva que busca unicamente encontrar o modelo com melhor ajuste, o algoritmo mais propício para análise empírica da qualidade de reservatórios é o *Ridge Regression*, permitindo maior precisão na estimação dos coeficientes explicativos.

Algoritmo	Função Objetivo	Penalidade	Descrição
Método dos Mínimos Quadrados	$\min_w \|Xw - y\|_2^2$		Minimiza o quadrado dos erros.
Ridge Regression	$\min_w \|Xw - y\|_2^2 + \alpha\|w\|_2^2$	ℓ_2	Minimiza o quadrado dos erros, penaliza coeficientes com valores altos.
Lasso Regression	$\min_w \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + \alpha\|w\|_1$	ℓ_1	Minimiza o quadrados dos erros, penaliza o acréscimo de features ao modelo.
Elastic Net	$\min_w \frac{1}{2n_{samples}} \|Xw - y\|_2^2 + \alpha\rho\|w\|_1 + \frac{\alpha(1-\rho)}{2}\|w\|_2^2$	ℓ_1 e ℓ_2	Minimiza o quadrado dos erros, misturando as penalidades sobre quantidade de features e tamanho dos coeficientes.

Comparação da função objetivo de cada algoritmo*

**Support Vector Regression* não está incluído por ser modelado como a minimização de um dicionário primal

Instituto de Informática - UFRGS
Av. Bento Gonçalves, 9500.
CEP:91501-97 Porto Alegre - RS

Grupo Banco de Dados Inteligentes
http://www.inf.ufrgs.br/bdi

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico