Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Tese de Doutorado

Área: Computação Inteligente

Título: Uma abordagem para solucionar problemas de multicolinearidade em modelos de regressão semi-supervisionada com *ensemble*.

Orientadora – Roberta Andrade de A. Fagundes (roberta.fagundes@upe.br)
Co-orientadora - Renata Maria Cardoso Rodrigues de Souza (rmcrs@cin.ufpe.br)

Descrição

Análise de Regressão, segundo Montgomery et al. [1], é uma ferramenta estatística que tem como objetivo realizar a estimação de valores (y) e explicar o relacionamento entre variáveis através de modelos matemáticos. Essas variáveis são dividas em dois grupos: variáveis independentes (x); e variáveis dependentes (y). Uma das principais dificuldades encontradas por pesquisadores em análise de regressão é conseguir trabalhar com conjuntos de dados que possuem sérios problemas de multicolinearidade [2] e/ou de alta dimensionalidade, em que o número de co-variáveis é maior do que o número de observações. Há multicolinearidade em um modelo de regressão ocorre quando duas ou mais variáveis independentes (x) são fortemente relacionadas linearmente entre si. Estas situações podem causar resultados instáveis nos métodos de regressão por mínimos quadrados ordinários, além do aumento da variância dos coeficientes estimados. Assim, modelos de regressão Ridge e LASSO resolvem o problema de multicolinearidade acrescentando um pequeno vício ao estimador dos coeficientes de regressão encontrado por mínimos quadrados ordinários, afastando o sistema da singularidade. Tais modelos, apesar de viciarem o estimador, possibilitam encontrar resultados com menores erros quadráticos médios, à medida que diminui consideravelmente a variância dos estimadores. As principais diferenças entre a regressão Ridge e o LASSO estão na penalização utilizada na expressão da soma de quadrados dos erros que deve ser minimizada para estimar os coeficientes de regressão. Portanto, a regressão ridge levar as estimativas de alguns coeficientes à zero, enquanto que o LASSO faz com que as estimativas de alguns coeficientes sejam exatamente iguais a zero, tornando o segundo modelo mais interpretável, com menor número de co-variáveis. Essa proposta tem por objetivo solucionar o problema de multicolinearidade através da utilização de modelos de regressão Ridge e LASSO para diferentes cenários (reais e simulados) definidos por número de co-variáveis, tamanho de amostra e quantidade e intensidade de coeficientes (efeitos) significativos.

Referências Bibliográficas

[1] NORVIG P. e RUSSELL S**. Inteligência Artificial,** 3ª Edição, 2013**.**
[2] Douglas C Montgomery, Elizabeth A Peck, and G Georey Vining. **Introducion to linear regression analysis**, volume 821. John Wiley & Sons, 2012.
[3] Duzan, H.; Shariff, N. S. B. M. **Ridge Regression for Solving the Multicollinearity Problem: Review of Methods and Models,** Journal of Applied Sciences 15 (3): 392-404, 2015.