Regressão Linear Múltipla

Juliana Araújo

Universidade Federal de São João Del Rei, Dezembro 2019

1 Base de dados

A base de dados contém 1802 dados referentes aos preços de ações da empresa Petrobras, do ano de 2010 a 2017. Os dados são divididos nas seguintes categorias: data (date), valor de abertura da ação (open), maior valor alcançado (high), menor valor (low), preço de fechamento (close) e volume negociado na ação (volume).

2 Solução

Foi implementado o algoritmo para regressão linear múltipla mínimos quadrados, de modo que a base de dados foi dividida em treino (70%) e teste (30%). Para isso foi utilizada a classe train_test_split da biblioteca scikit-learn.

Posteriormente foi realizada a análise da correlação de cada categoria com o vetor de resultados (close), e foi decidido que a coluna de 'volume' seria descartada por estar pouco correlacionada, e também a categoria de data. A matriz G foi gerada com as demais variáveis regressoras (open, high, low) e foi adicionada a coluna contendo apenas '1' para executar o algoritmo dos mínimos quadrados.

Após realizar a operação $G^t * G$ e $G^t * y$, foram obtidos os coeficientes de regressão por meio da eliminação de gauss. Não foi utilizada a matriz inversa de $G^t * G$, tendo em vista que seus resultados não foram satisfatórios. Após obter o vetor de coeficientes, foi aplicado o algoritmo de regressão linear da biblioteca numpy a fim de conferir os resultados.

3 Análise do erro

A análise do erro foi realizada utilizando a biblioteca scikit-learn, onde é possível obter o erro quadrático médio, ou seja, a média da diferença entre o valor estimado e do parâmetro ao quadrado. O valor obtido está próximo de zero, sendo igual a 0.15843997228226042.

4 Limitações da solução

O resultado foi testado apenas para um conjunto de categorias, sem que houvesse a verificação de que se há outra configuração a obter um erro menor.