Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

DCC057 Mineração de Dados

Trabalho Prático 4 — Relatório de Preparação dos Dados

Alexander Thomas Mol Holmquist

04 de abril de 2021

1 Objetivo

Este relatório apresentará questões encontradas no desenvolvimento do projeto, no que se relaciona à preparação dos dados para modelagem.

2 Desenvolvimento

Os principais passos tomados são dados abaixo. pricesDf se refere à tabela obtida do arquivo ”prices.csv”, e já pre-processada de acordo com os procedimentos da fase 1.

1. Substituição das colunas ”open”e ”close”no banco de dados por duas novas colunas: a média e a diferença entre essas colunas. Isso foi feito para aumentar a coerência do modelo. Ou seja, fazê-lo focar no preço daquele dia e na diferença entre o preço inicial e o preço final, ao invés de simplesmente fornecê-los o preço inicial e o preço final, o que poderia causar confusão.

2. Foi decidido transformar a coluna ”symbol”da seguinte forma. A média dos valores para ”Total Assets”no arquivo ”fundamental.csv”, relacionados com cada empresa, substituiu o seu respectivo símbolo na coluna ”symbol”. Esse identificador é numérico e comparável, fator necessário para que entre no modelo. A nova coluna foi chamada de ”totAssets”.

3. Foi decidido substituir as colunas ”totAssets”e ”volume”com uma transformação logarítimica. Esta decisão foi tomada por conta da cauda pesada apresentada pela distribuição destas duas colunas, além de uma assimetria positiva.

3 Decisões Importantes

1. Foi decidido considerar tão somente o modelo de regressão linear. Inicialmente, tinha-se proposto os seguintes: regressão linear, redes neuronais, regressão logística, árvore de decisão. Regressão logística foi rejeitado por ser um modelo mais voltado para classificação. Redes neuronais foi rejeitado pela complexidade de implementação, e por conta da quantidade de dados na tabela principal. Árvore de decisão foi rejeitado ao comparar com regressão linear, pois é mais lenta computacionalmente, e não buscamos, no presente projeto, interpretar os atributos.

2. Os dados serão normalizados de acordo com a z-pontuação.

4 Código fonte produzido

Toda a parte de pre-processamento dos dados para a aplicação no modelo de regressão linear, inclusive a z-normaliação, foi condensada em uma função que tomou o nome de “fetchFiltPricesDf”. Também foi provida uma função que automaticamente divide uma tabela de dados em conjuntos de treinamento e teste, chamada “splitTrainTest”.

Todo o código fonte produzido nesta fase, e que pode ser utilizado em fases posteriores, pode ser encontrado no diretório “src/preprocessing”.