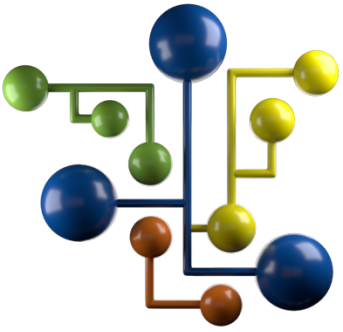

Data Science Academy

www.datascienceacademy.com.br

Deep Learning I

Mini-Projeto 2 - Detecção de Fraudes em Transações Financeiras com Deep Learning

Nosso objetivo neste mini-projeto é implementar um modelo de Deep Learning, a fim de classificar, ao mais alto grau possível de precisão, fraudes de cartão de crédito de um conjunto de dados reunido na Europa durante 2 dias no mês de setembro de 2013.

Um desafio ao analisar este dataset (e inerente a qualquer modelo de detecção de fraudes) é o enorme desequilíbrio no conjunto de dados: as fraudes representam apenas 0,172% das transações. Nesse caso, é muito pior ter falsos negativos do que falsos positivos em nossas previsões, pois falsos negativos significam que alguém cometeu fraude mas não fomos capazes de prever. Falsos positivos apenas causariam uma complicação para o cliente, sendo necessário uma verificação se a fraude ocorreu ou não.

Vamos analisar um dataset que contém transações realizadas com cartões de crédito em setembro de 2013 pelos detentores de cartão europeus. Este conjunto de dados apresenta as transações ocorridas em dois dias, onde temos 492 fraudes em 284.807 transações. O conjunto de dados está altamente desbalanceado, a classe positiva (fraudes) representa 0,172% de todas as transações.

O dataset contém apenas variáveis numéricas de entrada que são o resultado de uma redução de dimensionalidade com PCA (Principal Component Analysis). Infelizmente, devido a problemas de confidencialidade, o dataset não pode ser fornecido com as informações básicas originais.

Os atributos V1, V2, ... V28 são os componentes principais obtidos com PCA, os únicos atributos que não foram transformados com PCA são 'Time' e 'Amount'. O atributo 'Time' contém os segundos decorridos entre cada transação e a primeira transação no conjunto de dados. O atributo 'Amount' é a quantia da transação. O atributo 'Classe' é a variável de resposta sendo o valor 1 em caso de fraude e 0 caso contrário.

Dada a taxa de desbalanceamento das classes, recomendamos medir a precisão usando a Área sob a Curva de Recuperação de Precisão (AUPRC). A precisão da matriz de confusão não é significativa para a classificação desbalanceada.

O conjunto de dados foi coletado e analisado durante uma colaboração de pesquisa da Worldline e do Grupo de Aprendizado de Máquinas (http://mlg.ulb.ac.be) da ULB (Université Libre de Bruxelles) sobre Mineração de Big Data e Detecção de Fraude. Mais detalhes sobre projetos atuais e anteriores sobre tópicos relacionados estão disponíveis em http://mlg.ulb.ac.be/BruFence e http://mlg.ulb.ac.be/ARTML

Execute o mini-Projeto, analisando cada etapa executada! Este mini-projeto é para seu auto-estudo. Bom trabalho.