Projeto de Machine Learning

Identificar Fraude nos E-Mails da Enron

1. Visão Geral do Projeto

Neste projeto, você irá atuar como um detetive, e usar suas habilidades em aprendizagem de máquina para criar um algoritmo que identifique os funcionários da Enron que podem ter cometido fraude baseando-se no conjunto de dados público entitulado "Enron financial and email".

Por que este projeto?

Este projeto vai ensinar você a como fazer um processo de investigação de dados do início ao fim, sempre pensando como um especialista em aprendizagem de máquina.

Este projeto vai te ensinar como extrair e identificar atributos úteis que representem seus dados, alguns dos algoritmos mais utilizados atualmente na área, e como avaliar a performance dos seus algoritmos.

O que irei aprender?

Até o final do projeto, você será capaz de:

- Lidar com um conjunto de dados real e suas imperfeições
- Validar resultados de aprendizagem de máquina usando dados de teste
- Avaliar resultados de aprendizagem de máquina usando métricas quantitativas
- Criar, selecionar e transformar atributos
- Comparar a performance de algoritmos de aprendizagem de máquina
- Otimizar algoritmos de aprendizagem de máquina para obter máxima performance
- Comunicar seus resultados de aprendizagem de máquina de forma clara

Por que isso é importante para minha carreira?

Aprendizagem de Máquina é um ticket de primeira classe para uma das carreiras mais excitantes na área de análise de dados atualmente.

Como fontes de dados se proliferam juntamente do poder de processamento dos computadores atuais, analisar os dados diretamente é uma das formas mais interessantes de obter conhecimento novo e realizar previsões.

Aprendizagem de Máquina junta o poder da ciência da computação e da estatística para atingir esse poder preditivo.

2. Como completar este projeto

Uma nota antes de você começar: os mini-projetos do curso foram projetados para possuir muitos dados, dar resultados intuitivos, e até onde esperamos, se comportar bem. Este projeto será significativamente mais complicado pois estamos lidando com dados reais, que podem ser confusos e não possuem tantos dados quanto nós esperamos ter para trabalhar com aprendizagem de máquina. Não se sinta mal, dados imperfeitos são a realidade que analistas de dados enfrentam todos os dias! Se você encontrar algo que nunca viu antes, volte um pouco e pense em como contornar a situação. Você consegue!

Visão Geral do Projeto

Em 2000, Enron era uma das maiores empresas dos Estados Unidos. Já em 2002, ela colapsou e quebrou devido a uma fraude que envolveu grande parte da corporação. Resultando em uma investigação federal, muitos dados que são normalmente confidenciais, se tornaram públicos, incluindo dezenas de milhares de e-mails e detalhes financeiros para os executivos dos mais altos níveis da empresa. Neste projeto, você irá bancar o detetive, e colocar suas habilidades na construção de um modelo preditivo que visará determinar se um funcionário é ou não um funcionário de interesse (POI). Um funcionário de interesse é um funcionário que participou do escândalo da empresa Enron. Para te auxiliar neste trabalho de detetive, nós combinamos os dados financeiros e sobre e-mails dos funcionários investigados neste caso de fraude, o que significa que eles foram indiciados, fecharam acordos com o governo, ou testemunharam em troca de imunidade no processo.

Recursos Necessários

Você deve possuir o python e o sklearn rodando no seu computador, assim como um código inicial (que contém scripts python e o conjunto de dados Enron) que você fez download como parte do primeiro mini-projeto no curso de Introdução a Aprendizagem de Máquina.

O código inicial pode ser encontrado no diretório final_project o código que você fez download. Alguns arquivos relevantes são:

- **poi_id.py**: Código inicial do identificar de pessoas de interesse (POI, do inglês Person of Interest). É neste arquivo que você escreverá sua análise. Você também enviará uma versão deste arquivo para que o avaliador verifique seu algoritmo e resultados.

- **final_project_dataset.pkl**: O conjunto de dados para o projeto. Veja mais detalhes abaixo.
- **tester.py**: ao enviar sua análise para avaliação para o Udacity, você enviará o algoritmo, conjunto de dados, e a lista de atributos que você utilizou (criados automaticamente pelo arquivo poi_id.py). O avaliador usará este código para testar seus resultados, para garantir que a performance é similar a obtida no seu relatório. Você não precisa usar modificar este código, mas nós o tornamos transparente para os alunos para que eles testem seus algoritmos e futura referência.
- **emails_by_address**: este diretório contém diversos arquivos de texto, cada um contendo todas as mensagens de ou para um endereço de email específico. Estes dados estão aqui para referência, ou caso você deseje criar atributos mais complexos baseando-se nos detalhes dos emails. Você não precisa processar estes dados para completar este projeto.

Etapas para o Sucesso

Nós vamos te fornecer um código inicial que carrega os dados, seleciona os atributos de sua escolha, os coloca em um vetor numpy, que é a forma de entrada mais utilizada pelas funções do sklearn. Seu trabalho é de usar engenharia sobre os atributos, escolher e otimizar um algoritmo e testar seu modelo preditivo. Grande parte dos mini-projetos foram desenvolvidos com este projeto final em mente, então lembre-se deles para trabalhar com aquilo que você já usou e fez anteriormente.

Como etapa de pré-processamento deste projeto, nós combinamos os dados da base "Enron email and financial" em um dicionário, onde cada par chave-valor corresponde a uma pessoa. A chave do dicionário é o nome da pessoa, e o valor é outro dicionário, que contém o nome de todos os atributos e seus valores para aquela pessoa. Os atributos nos dados possuem basicamente três tipos: atributos financeiros, de email e rótulos POI (pessoa de interesse).

atributos financeiros: ['salary', 'deferral_payments', 'total_payments', 'loan_advances', 'bonus', 'restricted_stock_deferred', 'deferred_income', 'total_stock_value', 'expenses', 'exercised_stock_options', 'other', 'long_term_incentive', 'restricted_stock', 'director_fees'] (todos em dólares americanos (USD)).

atributos de email: ['to_messages', 'email_address', 'from_poi_to_this_person', 'from_messages', 'from_this_person_to_poi', 'shared_receipt_with_poi'] (as unidades aqui são geralmente em número de emails; a exceção notável aqui é o atributo ‘email_address’, que é uma string).

rótulo POI: [‘poi’] (atributo objetivo lógico (booleano), representado como um inteiro).

Nós o encorajamos a criar, transformar e re-escalar novos atributos a partir dos originais. Se você fizer isso, você deverá armazenar os novos atributos na estrutura **my_dataset**, e se você utilizar estes atributos no seu modelo final, não esqueça de adicioná-los também a lista chamada **my_feature_list**, para que o avaliador seja capaz de acessá-la durante os testes. Para um exemplo de criação de novos atributos, veja a aula sobre Seleção de Atributos.