Visão geral do projeto

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.

Detalhes do Projeto

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. Você pode obter este projeto inicial usando o git: `git clone https://github.com/udacity/ud120-projects.git`
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.
Adicionalmente, nós o alertamos para que você tome notas durante o projeto. Como parte do seu projeto a ser submetido, você irá responder uma série de perguntas (dadas na próxima página), para que nós possamos entender sua proposta em diferentes tocantes da análise. Seu processo de pensamento é, em diversos modos, mais importante que seu projeto final e nós precisamos analisar seu processo de pensamento ao ler suas respostas.

Identificar fraude no Email da Enron

Project Submission

Banque o detetive e coloque suas habilidades de aprendizado de máquina em uso através da construção de um algoritmo para identificar funcionários da Enron que possam ter cometido fraude. Sua base será um conjunto de dados financeiros e de e-mail público da Enron.

Submissão e uma visão geral de Avaliação

Sua submissão irá conter vários arquivos: o código/classificador que você criou e alguma documentação escrita do seu trabalho. Vamos avaliar o seu projeto de acordo com esta rubrica; apenas os projetos que satisfaçam **todos** os itens "corresponde às expectativas" irão passar. **Por favor, auto avalie seu projeto antes de enviar!** Se você não acha que o seu projeto satisfaz todos os critérios, o avaliador do projeto provavelmente também não achará.

Submissão

Pronto para enviar seu projeto? Volte para o seu Udacity Home, clique sobre o projeto e siga as instruções para enviar!

- Você pode enviar-nos um link do GitHub com os arquivos ou fazer upload de um diretório comprimido (arquivo zip).
- Dentro da pasta zip inclua um arquivo de texto com uma lista de sites, livros, fóruns, blogs, repositórios GitHub etc os quais você fez referência ou usou nesta submissão (Adicione N/A se você não usar esses recursos).

Pode nos levar até uma semana para avaliar o seu projeto mas na maioria dos casos é muito mais rápido. Você receberá um e-mail quando a sua submissão for revista.

Itens a serem incluídos na submissão:

Código/Classificador

Ao fazer o seu classificador, você criará três arquivos *pickle* (`my_dataset.pkl`,`my_classifier.pkl`, `my_feature_list.pkl`). O avaliador do projeto irá testar seus arquivos usando o script `tester.py`. Você é encorajado a usar esse script para avaliar se o seu desempenho é bom o suficiente.
Você também deve incluir o seu arquivo `poi_id.py` modificado para verificarmos qualquer problema com o funcionamento de seu código ou para verificar o que é relatado em suas respostas (veja o próximo parágrafo). Notavelmente, devemos ser capazes de executar o `poi_id.py` para gerar os três arquivos de *pickle* que refletem seu algoritmo final, sem a necessidade de modificar o script de qualquer forma. Se você tem códigos intermediários que você gostaria de fornecer como material complementar, é recomendado que você salve-os em arquivos separados do `poi_id.py`. Se você fizer isso, certifique-se de fornecer um arquivo leia-me que explique para que serve cada arquivo.

Documentação do seu trabalho

Documente o trabalho que você fez respondendo (com um parágrafo cada) às perguntas encontradas aqui. Você pode escrever suas respostas em um PDF, texto/arquivo de markdown, HTML ou formato similar. As respostas na documentação devem permitir que um revisor entenda e siga os passos que você tomou em seu projeto e verifique a sua compreensão dos métodos que você executou.

Arquivo de texto listando suas Referências

Inclua um arquivo com a lista de sites, livros, fóruns, blogs, repositórios github etc. os quais você fez referência ou utilizou nesta submissão (adicionar N/A se você não usar esses recursos). Por favor, leia com atenção a seguinte declaração e inclua-a em seu documento "Confirmo que esta submissão é o meu trabalho. Citei acima quaisquer referências e informações retiradas de sites, livros, fóruns, blogs, repositórios github, etc.

Boa sorte!