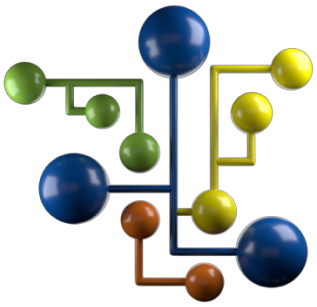

Data Science Academy

www.datascienceacademy.com.br

Big Data Real-Time Analytics com Python e Spark

O que é Pré-Processamento?

Os algoritmos de Machine Learning aprendem a partir dos dados. Portanto, é responsabilidade do Cientista de Dados alimentar os algoritmos com os dados de forma consistente, de acordo com o problema que se pretende resolver. Mesmo que seus dados estejam em bom estado, é preciso garantir que os dados estejam na mesma escala, no mesmo formato e que as variáveis mais significativas estejam em seu dataset.

A etapa de processamento de dados recebe muitas nomenclaturas: Data Munging, Data Wrangling, Data Preparation, Data Processing. Vamos chamar aqui de "Processo de Preparação dos Dados". Este processo é normalmente subdividido em 3 passos:

1- Seleção dos dados
2- Pré-Processamento dos Dados
3- Transformação dos Dados

Seleção dos Dados

Nesta etapa, selecionamos um subset dos dados com os quais iremos trabalhar. Mesmo na era do Big Data, nem todos os dados serão relevantes, de acordo com o problema a ser resolvido e devemos ter atenção para não selecionar mais dados do que o necessário para o processo de análise. Cada problema a ser resolvido, pode requerer diferentes conjuntos de dados e por isso a definição do problema é tão crítica dentro do processo de Big Data Analytics. Algumas perguntas podem ajudar no processo de definir que dados serão usados no processo de análise:

- Qual a extensão dos dados que temos disponíveis? Temos bancos de dados, Data Warehouses, sistemas departamentais, processos internos de ETL, dados em fitas de backup? Estes dados podem ser facilmente acessados e coletados? Existem questões de segurança na governança de dados (dados de cartão de crédito armazenados, por exemplo)? Qual o "timeframe" dos dados? Temos dados de 2, 3, 5 anos? Eles são relevantes? As respostas a estas questões vão ajudar a definir o alcance do seu processo de análise.

- Quais dados não estão disponíveis? No cadastro de clientes, temos os registros de endereço, como cidade e estado? Esta informação é relevante para seu processo? Se não temos os dados, como podemos obtê-los? Vamos preencher com um valor default?

- Que dados você não precisa para sua análise? Definir que dados serão excluídos é normalmente mais fácil do que definir que dados serão incluídos.

Pré-Processamento

Uma vez que os dados estejam selecionados, agora precisamos processá-los. O Pré-Processamento é a etapa de limpeza e transformação dos dados em um formato que possa ser utilizado ao longo do processo de análise. Normalmente 3 tarefas são executadas nesta etapa:

Formatação – os dados podem não estar no formato necessário para aplicar algoritmos de Machine Learning. Os dados podem estar em um formato específico ou proprietário e precisam ser formatados de acordo com as ferramentas de análise que você tem em mãos.

Limpeza – o processo de limpeza compreende remover dados duplicados, ajustar dados missing, remover caracteres especiais ou outras "sujeiras" presentes nos dados. Dados incompletos são um dos maiores problemas enfrentados nesta fase, pois a decisão de como isso será tratado vai impactar no resultado final da sua análise. Aqui também é preciso definir o que fazer com dados que tem caráter sensível ou sigiloso, de modo a não infringir leis ou regras.

Amostragem – trabalhar com todo o conjunto de dados pode ser uma tarefa demorada e desnecessária. A Estatística oferece técnicas de amostragem que podem ser aplicadas como forma de testar as primeiras versões dos seus modelos preditivos, antes de aplicar a versão final ao seu conjunto inteiro de dados.

Transformação

Esta é a etapa final do "Processo de Preparação dos Dados" e normalmente composta de 3 ações que precisam ser executadas:

Escala – após a etapa de pré-processamento, os dados podem estar em diferentes escalas e unidades. Diversos algoritmos de Machine Learning requerem que os dados estejam na mesma escala, normalmente com valores entre 0 e 1 para os menores e maiores valores em determinado atributo. Existem diversas técnicas de escala e isso precisa ser definido e a regra escolhida aplicada.

Decomposição – uma única variável no seu conjunto de dados pode conter informações complexas, que são representadas por um único código por exemplo. Neste caso, precisamos decompor a variável em 2 ou 3 variáveis, cada qual representando parte da informação contida na variável inicial. Isso pode fazer toda a diferença no processo de aprendizado de máquina. Um exemplo clássico é uma variável que armazena uma data completa com dia, mês, ano, hora, minuto e segundo. Esta é uma excelente variável candidata a ser decomposta em 2, 3, 4 ou 5 variáveis diferentes. Talvez apenas a hora seja relevante para sua análise. Faça a decomposição e descarte as informações irrelevantes.

Agregação – agregação é o processo inverso a decomposição. Pode ser necessário agrupar 2 ou 3 variáveis em uma única variável que seja mais significativa para o processo de análise. Por exemplo: um determinado sistema pode armazenar a data que cada usuário efetua login em um sistema. Podemos contar estas instâncias e gerar uma coluna chamada "count" com a quantidade de logins por hora.

Todo o "Processo de Preparação de Dados" é um trabalho complexo que envolve uma série de decisões que precisam ser tomadas pelo Cientista de Dados e sua equipe. Embora existem boas práticas, estas decisões vão envolver o problema a ser resolvido, a experiência dos profissionais, as ferramentas que estão sendo usadas e o objetivo final.

Python, através do módulo scikit-learn, oferece diversas ferramentas para as tarefas descritas acima e um pacote específico para pré-processamento, chamado *preprocessing*, que permite por exemplo ajustar a escala dos dados e que veremos mais a frente.

O Pré-Processamento é um trabalho iterativo. O resultado do seu trabalho nesta etapa pode ser visualizado apenas depois que você gerar seu modelo preditivo e neste caso, será necessário retornar e aplicar novas regras.

Obrigado

Equipe DSA