Atividade I - Introdução ao Reconhecimento de Padrões

William Hitoshi Sumida

Abril 2020

Desde reconhecer situações de perigo, aprender a dominar o fogo e intuitivamente prever o valor de ações na bolsa de valores, os seres humanos foram capazes de identificar padrões e transformar observações em tomada de decisão. Com o intuito de replicar a inteligência humana ou animal em máquinas, surgiu na ciência da computação, um campo de estudo chamado Inteligência Artificial.

Recentemente se estabeleceu um cenário vantajoso para o uso de algoritmos de inteligência artificial, por exemplo, quando é necessária uma entidade inteligente (humano ou computador) para analisar um volume grande de dados. Estamos diante desta situação devido à evolução do poder de processamento dos computadores e por ser humanamente impossível de explorar tal volume de dados de forma ágil.

Um conjunto de processos capaz de identificar padrões em grande volume de dados é chamado de Mineração de Dados ou Data Mining, a qual consiste em aquisição, pré-processamento, visualização, reconhecimento de padrões e análise dos dados. Para o reconhecimento de padrões, a mineração de dados muitas vezes utiliza algoritmos de aprendizado de máquina, os quais foram criados no ramo de inteligência artificial.

Normalmente os termos mineração de dados, reconhecimento de padrões, aprendizado de máquina e inteligência artificial geram muita confusão em suas definições, pois resolvem problemas de formas similares. Em palavras simples a definição de cada uma é, inteligência artificial é o ramo de estudo da ciência da computação com objetivo de replicar inteligência humana ou animal em máquinas. Aprendizado de máquina ou machine learning é um ramo da inteligência artificial que estuda algoritmos com a capacidade criar modelos matemáticos a partir de um conjunto de treinamento. Reconhecimento de padrões surgiu na engenharia e utiliza métodos estatísticos (modelagem preditiva) ou aprendizado de máquina para, como diz o nome, reconhecer padrões (reconhecimento de voz, digital, letras e números escritos em um papel, etc). Já a mineração de dados tem como objetivo gerar conhecimento a partir de um conjunto de dados, é um processo que engloba desde a aquisição de dados até a interpretação dos mesmos, podendo ou não utilizar o reconhecimento de padrões.

Antes de criar qualquer modelo, precisamos nos preocupar com o conjunto

de dados que será analisado. Os dados podem ser classificados como categóricos ou quantitativos. Dados categóricos podem ser nominais (nome, gênero, clima) ou ordinais (sentimento, intervalo) já os dados quantitativos podem ser discretos (números inteiros) ou contínuos (números reais). No caso de análise de imagem, voz, ou sinal, é necessário encontrar a melhor forma de representá-las por meio de variáveis categóricas ou quantitativas. Por exemplo, extrair os canais RGB de uma imagem.

A próxima etapa normalmente é a seleção de atributos para que o modelo tenha alta precisão, acurácia e boa performance. Uma das formas de selecionar atributos, é fazer uma regressão linear com todos os atributos e identificar pares com coeficiente de correlação alto. Se dois atributos são linearmente correlacionados, provavelmente possuem a mesma relação com o atributo a ser previsto, portanto podemos remover um dos atributos.

Um dos problemas que prejudicam o modelo, é a maldição da dimensionalidade, a qual se refere à grande quantidade de atributos. A cada atributo adicionado em um dataset, cria-se a necessidade de obter uma quantidade exponencial de amostras para o modelo, pois foram criados espaços vazios no dataset que o algoritmo não conseguirá mapear, causando uma redução na precisão e acurácia. A maldição da dimensionalidade também afeta a performance do algoritmo, pois o mesmo terá que analisar um volume maior de dados.