FORMAÇÃO COMPLEMENTAR

Mentor: Felipe Moreira de Assunção
Conteúdo para formação complementar de Machine Learning

INFORMAÇÕES DO MÓDULO

Descrição

Conheça os conceitos básicos sobre Machine Learning

Objetivos do Ensino

- Conhecer os principais conceitos de Machine Learning
- Conhecer o ecossistema de desenvolvimento para Machine Learning
- Principais etapas
- Principais bibliotecas
- Exemplo de classificação de imagem

Dica de ferramenta

Você pode utilizar o Google Colab, um editor de código totalmente web e com muitas facilidades para você codificar.

Conta com as principais bibliotecas para Machine Lerning, Análise e Visualização de Dados, sem a necessidade de instalação ou qualquer configuração avançada.

Além do mais, o Google Colab conta com possibilidade de utilizar uma GPU integrada que irá contribuir para que o seu programa seja executado mais rapidamente.

Dica para estudo

Para o melhor aproveitamento deste módulo, sugerimos como material complementar a leitura do Livro: “A. Hands-On Machine Learning with Scikit-Learn and TensorFlow”, principalmente o primeiro capítulo.

Para um primeiro exemplo, você pode acessar a sessão de Machine Learning do site https://www.w3schools.com/python/default.asp e testar os exemplos do site, para que você comece a entender algumas possibilidades na construção de um modelo.

MACHINE LEARNING

1. Introdução ao Machine Learning

Machine Learning é o cérebro que move a tecnologia de IA e as tecnologias de IA que fazem as ações. É o processo de aplicar processamento analítico para descobrir padrões escondidos ou tendências que são uteis para predições a partir de modelos matemáticos. A partir da construção desse modelo, você pode fazer previsões sobre dados futuros. Por exemplo, uma possível aplicação de um modelo de Machine Learning é prever a probabilidade de um cliente comprar um determinado produto com base no comportamento passado.

Alguns termos importantes:

- Feature: sinônimo de variável, colunas, atributos e campo
- Instância: sinônimo de linha, observação, dado, valor e caso
- Target ou Label: sinônimo de predito e variável dependente
- Dado: Sinônimo de preditor e variáveis preditivas

Quando usar o Machine Learning?

É importante lembrar que o ML não é uma solução para todos os tipos de problemas. Em alguns casos, é possível desenvolver soluções robustas sem o uso de técnicas de ML. Por exemplo, você não precisa de ML se pode determinar um valor de destino usando regras simples, cálculos ou etapas predeterminadas que podem ser programadas sem a necessidade de qualquer aprendizagem orientada por dados. Use Machine Learning nas seguintes situações:

- Você não pode codificar as regras: muitas tarefas humanas (como reconhecer se um e-mail é spam ou não) não podem ser adequadamente resolvidas usando uma solução simples (determinística) baseada em regras. Um grande número de fatores pode influenciar a resposta. Quando as regras dependem de muitos fatores e muitas regras se sobrepõem ou precisam ser ajustadas com muita precisão, rapidamente se torna difícil para um ser humano programar com precisão as regras. Você pode usar o ML para resolver esse problema com eficácia.
- Você não pode dimensionar: você talvez possa reconhecer manualmente algumas centenas de e-mails e decidir se são spam ou não. No entanto, essa tarefa se torna entediante quando se trata de milhões de e-mails. As soluções de ML são eficazes para lidar com problemas de grande escala.

Onde podemos aplicar Machine Learning?

- Predição de vendas
- Segmentação de usuário
- Detecção de fraudes
- Predição de classes (texto, áudio, imagem etc)

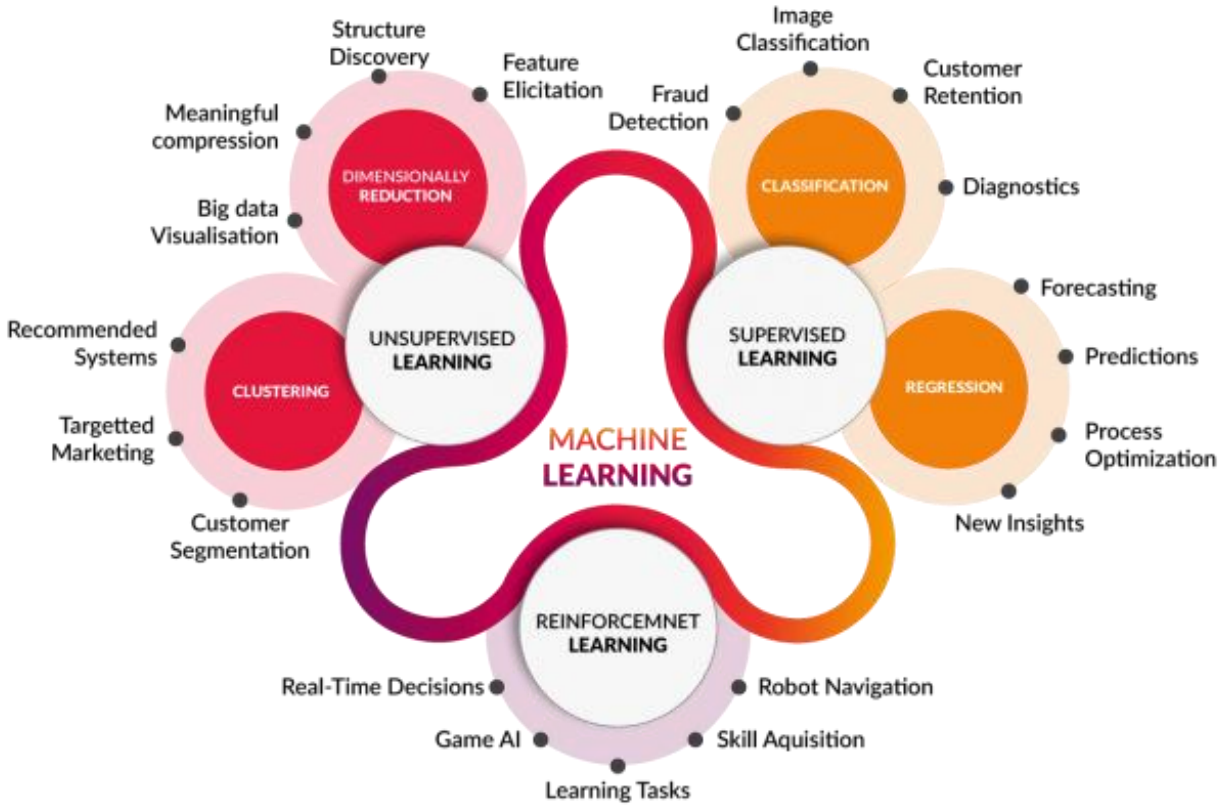
Quais são os tipos de problemas que podemos resolver com Machine Learning?

- Classificação multiclasse
- Classificação binária
- Regressão

Dentre os tipos de aprendizado e seus usos, temos:

- Aprendizado supervisionado: Faz predições de dados que possuem categorias ou classes
- Aprendizado não-supervisionado: Faz predições de dados que não possuem categorias ou classes

- Aprendizado por reforço: A aprendizagem por reforço é o treinamento de modelos de aprendizado de máquina para tomar uma sequência de decisões, muito aplicada em modelos de Inteligência Artificial.



2. **Etapas de criação de uma aplicação de Machine learning**

Formulação do problema:
Coleta de dados rotulados: Para coletar dados, podemos utilizar o Kaggle e o UCI Repository que possuem diversos datasets.
Analisar os dados: Análise exploratória e limpeza dos, dados, avaliar algum pré processamento etc.
Processamento de recursos ou Extração de características: Extração de características relevantes para utilização na etapa de treinamento do nosso modelo
Divisão dos dados em conjunto de treinamento e avaliação: Podemos usar amostragens randômicas para gerar amostras e depois dividir os dados entre conjunto de teste e conjunto de treino.

Exemplo de divisão:

- 2/3 dos dados -> conjunto de treino (usado para treinar o modelo)
- 1/3 dos dados -> conjunto de teste (usado para medir se o modelo está performando bem)

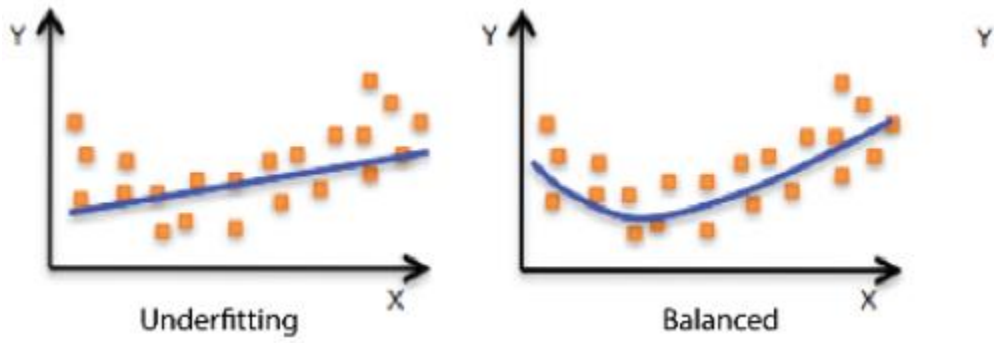
Treinamento do modelo:
Considerar modelos lineares e não lineares, algoritmo de aprendizagem, parâmetros de treinamento, taxa de aprendizado, construção do modelo, número de iterações, embaralhamento e divisão dos dados, mecanismos de regularização.

Avaliar a precisão do modelo: Através de métricas de avaliação como matriz de confusão, acurácia, precisão, recall (revocação), f1 score, especificidade, sensibilidade, ROC, dentre outros.

Aprimorar a precisão do modelo:
1. Colete dados: aumente o número de exemplos de treinamento
2. Processamento de recursos: adicione mais variáveis e um melhor processamento de recursos
3. Ajuste do parâmetro de modelo: considere valores alternativos nos parâmetros de treinamento usados pelo algoritmo de aprendizagem

Ajuste do modelo: underfitting e overfitting

Compreender o ajuste de modelo é importante para entender a causa raiz da precisão de modelo insatisfatória. Essa compreensão orientará você a tomar medidas corretivas. Podemos determinar se um modelo preditivo está fazendo o subajuste ou o sobreajuste dos dados de treinamento consultando o erro de previsão nos dados de treinamento e nos dados de avaliação.



Biblioteca Sci-kit Learn: É uma biblioteca Python de código aberto que implementa uma variedade de aprendizado de máquina, pré-processamento, validação cruzada e visualização algoritmos usando uma interface unificada.

Ideias para próximos projetos

Agora você já sabe o básico sobre como criar modelos de ML, você já pode explorar outros datasets! O Kaggle é uma excelente fonte de datasets e contém vários desafios para você

exercitar. Sugerimos dois tutoriais com alguns desafios que você pode fazer na idle de sua preferência, ou no Google Colab, por exemplo.

1. Criar um algoritmo capaz de ler uma tabela .csv com dados do NETFLIX e trabalhar todas as etapas de Machine Learning, de modo que você possa aprender a partir de tarefas comuns da Ciência de Dados (Netflix dataset)
2. Criar um algoritmo que consegue identificar a espécie correta de uma flor a partir de suas medidas (Iris dataset)

BIBLIOGRAFIA

GÉRON, A. Hands-On Machine Learning with Scikit-Learn and TensorFlow.

NIELSEN, M. Neural Networks and Deep Learning. Determination Press, 2015. California, EUA: O'Reilly Media, 2017.