Aprendizagem de Máquina

Advanced Institute for Artificial Intelligence

https://advancedinstitute.ai

Aprendizagem de Máquina

Agenda

- Ciência dos Dados
- O que é aprendizagem de máquina
- Categorias de algoritmos de ML
- Exemplos de problemas de predição
- Problemas supervisionados
- Problemas não-supervisionados
- Elementos de um algoritmo de aprendizagem de máquina
- Organizando dados para desenvolver um modelo de aprendizagem de máquina
- Processo iterativo de desenvolvimento de algoritmos de aprendizagem de máquina

Ciência dos Dados

- Houve uma evolução drástica da infraestrutura para armanezagem e coleta de dados nos últimos 15 anos
  - Praticamente todas as instituições coletam dados sobre seus processos e clientes
- A análise de dados para se obter vantagens competitivas não é algo novo
  - A mudança para os dias de hoje está na inviabilidade da análise manual desses dados
- Os computadores modernos possibilitaram automatizar as possíveis combinações de dados
  - Permite realizar análise mais sofisticadas

Ciência dos Dados

- A formalização do processo (semi)automatizado de análise de dados tem recebido diferentes nomes ao longo dos anos
- Muitos dos métodos classificados como de Data Mining, também são classificados como de Aprendizado de Máquinas
- Há ainda interseções (ou sobreposições) com Business Intelligence, que muitas vezes inclui métodos de data mining/machine learning e outras técnicas da área de Banco de Dados
- Mais recentemente, o processo de análise de dados tem sido conhecido como Data Science

Aprendizagem de Máquina

O que é aprendizagem de máquina?

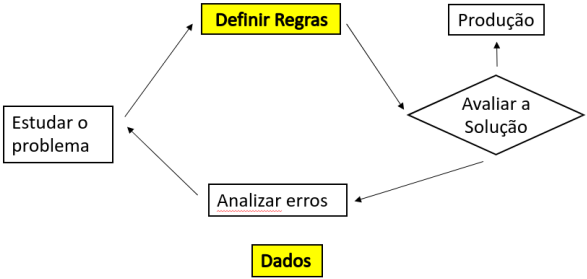
- Machine Learning é a ciência (arte) da programação de computadores para que eles possam aprender com os dados.
- Machine Learning é o campo de estudo que oferece aos computadores a capacidade de aprender sem serem explicitamente programados". Arthur Samuel, 1959
- Um algoritmo determinístico possui regras claras para retornar resultados de acordo com a entrada fornecida
- Se a entrada pode variar muito, esse conjunto de regras será muito grande, podendo tornar o tempo de execução inviável

Aprendizagem de Máquina

Sistemas baseados em regras:

- Natureza dinâmica dos problemas exige redefinição constante das regras
- Sistema de detecção de SPAM de email
  - Spammers identificam que as regras nao detectam números e trocam "Dois" por 2
  - Cada pequena mudança exigirá uma adaptação de regras

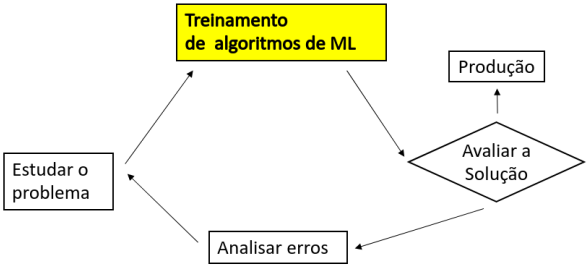
Aprendizagem de Máquina



Aprendizagem de Máquina

- Um programa tradicional necessitará de uma longa lista de regras
- Um filtro de spam baseado em aprendizagem de máquina é capaz de utilizar critérios diversos para fazer tal classificação
- Caracterização de um SPAM pode ser adaptada dinâmicamente de acordo com marcações atribuídas pelos usuários

Aprendizagem de Máquina

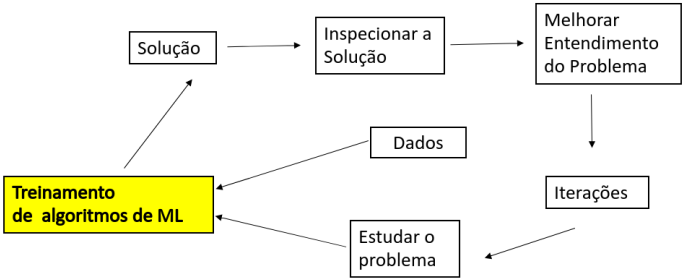


Aprendizagem de Máquina

Processo de utilização de aprendizagem de máquina

- A utilização de algoritmos de aprendizagem de máquina podem ser melhorados, a partir da análise dos resultados dos algoritmos
- A aplicação de técnicas de ML para avaliar grandes quantidades de dados pode ajudar a descobrir padrões que não eram aparentes.
- A utilização de aprendizagem de máquina pode ser entendida como um processo iterativo, em busca de soluções a partir dos dados, e otimização do uso dos dados e algoritmos
- Esse processo pode ser automatizado

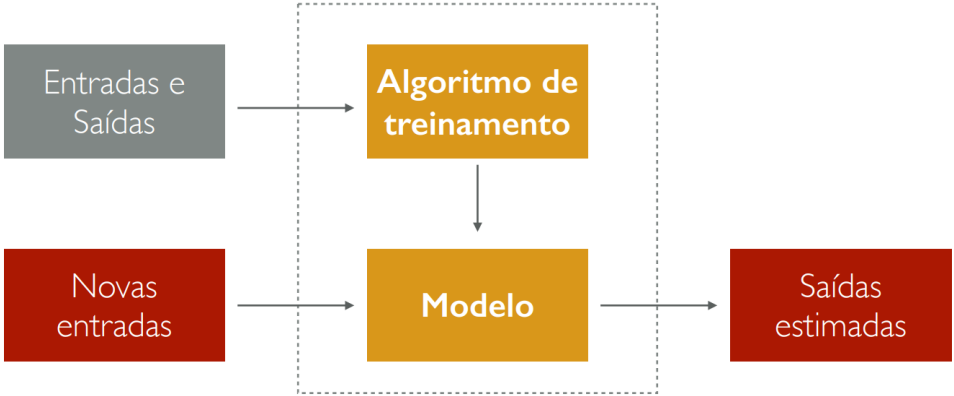
Aprendizagem de Máquina



Aprendizagem de Máquina

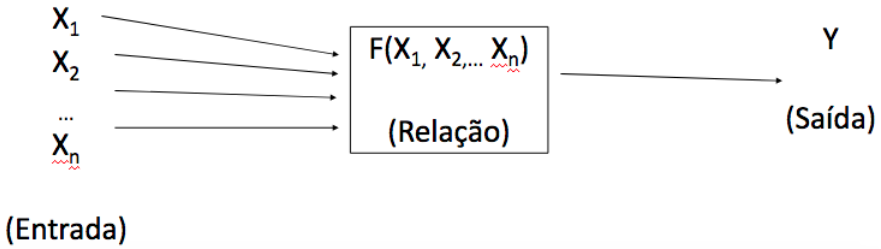
- Fundamentalmente, o aprendizado de máquina envolve a construção de modelos matemáticos para ajudar entender dados
- Ajustes de parâmetros nos modelos permite que os modelos sejam adaptados aos dados observados
- Desta forma, tais modelo podem ser usados para prever e entender aspectos de dados desconhecidos

Aprendizagem de Máquina



Aprendizagem de Máquina

- Aprendizagem Estatística:
  - Conjunto de abordagens para estimar encontrar funções preditivas a partir de dados
  - Técnicas para avaliar as estimativas obtidas;



Aprendizagem de Máquina

Porque estimar a função F?

- Predição: estimar o valor de uma variável de saída Y a partir de uma ou mais valores de variáveis de entrada X
- Inferência: entender a relação entre cada variável X e a variável Y

Aprendizagem de Máquina

Como estimar a função F?

- O processo estatístico é iniciado a partir de um conjunto de eventos conhecidos (Base de treino)
- Cada evento possui um ou mais valores de variáveis preditoras X: X1, X2, ..., Xn e um valor de saída Y
- Avaliação de desempenho da função F
- Distância entre o valor predito e o valor observado
- Processos estatísticos para avaliação da acurácia do modelo

Aprendizagem de Máquina

Categorias elementares de Algoritmos de Aprendizagem de Máquina

- Supervisionada
  - Classificação
  - Regressão
- Não-Supervisionada
  - Agrupamento
  - Redução de Dimensionalidade
- Semi-Supervisionada

Aprendizagem de Máquina

Aprendizagem Supervisionada

- Envolve modelar a relação entre medidas características dos dados e algum rótulo associado aos dados
- O modelo determinado pode ser usado para aplicar rótulos a novos dados
- Tipos de algoritmos supervisionados
  - Classificação: rótulos são categorias discretas
  - Exemplo filtro de spam: Emails são marcados como spam ou não-spam. Modelo classifica novos emails
  - Regressão: rótulos são quantidades contínuas
  - Exemplo: previsão do preço de um carro considerando um conjunto de variáveis preditoras (quilometragem, idade, marca)

Aprendizagem de Máquina

No aprendizado não supervisionado os dados de treinamento não são rotulados. O sistema tenta aprender sem referência ou dados anotados.

- Com base em dados sobre os visitantes de um site. Executar um algoritmo para tentar detectar grupos de visitantes semelhantes.
- Em nenhum momento você diz ao algoritmo a qual grupo um visitante pertence: ele encontra essas conexões sem ajuda.

Aprendizagem de Máquina

Para Qualquer problema a ser investigando como aprendizagem de máquina temos algums características comuns

- Amostras (*Samples*): linhas na base de dados
- Características (*Features*): colunas na base de dados
- Matriz de Características: Combinação de linhas e características
- Matriz alvo: coluna que se deseja predizer

Aprendizagem de Máquina

- Algoritmos de aprendizagem de máquina normalmento necessitam de uma grande quantidade de dados para aprensentar uma solução satisfatória
- Dados precisam ser representativos em relação ao problema que está sendo investigado
- Considerar a influência das categorias em relação a base completa
- Qualidade dos Dados:
  - Considerar detectar e se possível eliminar Outliers e Ruídos
  - Descartar dados redundantes
    - São desnecessários quando colocados no contexto de outro atributo
    - E.g., Classe social e renda mensal
  - Descartar dados irrelevantes
    - Não têm relação com o atributo-alvo
    - E.g., CPF e doença

Aprendizagem de Máquina

Projeto iterativo de aprendizagem de máquina:

- ☐ Definir o problema que se deseja atacar com um modelo preditivo
- ☐ Organizar os dados de acordo com o problema definido
- ☐ Definir uma métrica de avaliação
- ☐ Separar os dados em treino e teste de acordo coma métrica
- ☐ Inspecionar a solução
- ☐ Propor melhorias no modelo ou organização dos dados

O processo de organização de dados de acordo com o modelo definido envolve as seguintes atividades:

- Trocar dados categóricos ou ordinais por números
- Alterar a escala dos dados
- Eliminar valores faltantes ou substituir por outro valor
- Separar variáveis preditoras e variáveis alvo
- Dividr a base em treino e teste