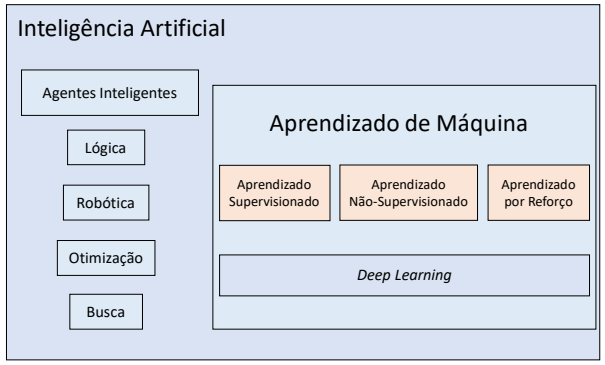
Universidade Federal do Pará

Programa de Pós-Graduação em Ciência da Computação



Aprendizado de Máquina

Prof. Dr. Bianchi Serique Meiguins
Prof. Dr. Tiago Araújo

Definição

- Área de estudo que estuda a habilidade de computadores aprenderem sem programação explícita.
- É dito que um programa de computador aprende com a experiência **E** em respeito a uma tarefa **T** baseado na medida de desempenho **P**, se o desempenho em **T**, medido por **P**, melhora com a experiência **E**.

Representação

Estudo com antibióticos

Medição do experimento

```
..., {
    "Bacteria":"Aerobacter aerogenes",
    "Penicillin":870,
    "Streptomycin":1,
    "Neomycin":1.6,
    "Gram_Staining":"negative",
    "Genus": "other"
},
{
    "Bacteria":"Bacillus anthracis",
    "Penicillin":0.001,
    "Streptomycin":0.01,
    "Neomycin":0.007,
    "Gram_Staining":"positive",
    "Genus": "other"
}, ...
```

Vetor de características

[[870, 1, 1.6, 0], [0]]
[[0.001, 0.01, 0.007, 0], [1]]

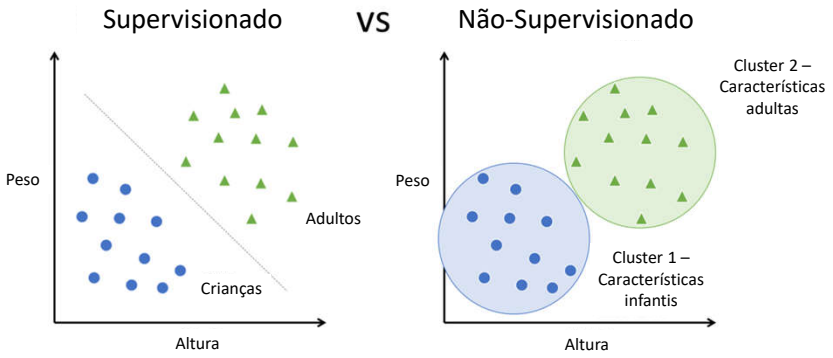
Tarefas

- Aprendizado Supervisionado
  - Entrada: Conjunto de dados com rótulo
  - Objetivo: Descobrir o rótulo de amostras não vistas
  - Tipos:
    - Classificação – Rótulo discreto
    - Regressão – Rótulo contínuo
- Aprendizado Não-Supervisionado
  - Entrada: Conjunto de dados sem resposta
  - Objetivo: Encontrar estrutura nos dados

Dados
- Normalização
- Transformações
- Validação
- Seleção de atributos

Treino
- Seleção de hiperparâmetros
- Seleção de métodos
- Teste e validação

Produção
- Infraestrutura
- Monitoramento
- Refinar modelo para escalar

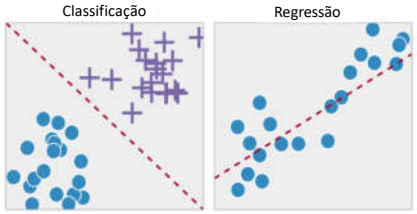


Aprendizado Supervisionado

- Dado um conjunto de dados $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ um algoritmo de aprendizado procura uma função $f(X) = Y$
- $X =$ vetores de características (x_i representa o i-ésimo vetor)
- $Y =$ rótulos (y_i representa o i-ésimo rótulo)
- Dois tipos, depende de Y
  - Regressão quando é Y contínuo
  - Classificação quando é Y categórico

Aprendizado Supervisionado

- Divide-se em dois problemas principais: classificação e regressão
- O aprendizado é feito de vários modos:
  - Sem treino
  - Mínimos Quadrados
  - Método do Gradiente
  - Entropia



K-Nearest Neighbor (KNN)

- Algoritmo simples, sem otimização
- Pontos novos procuram os vizinhos mais próximos, calculando distância
- As classes dos pontos mais próximos define a classe dos pontos novos

Regressão logística

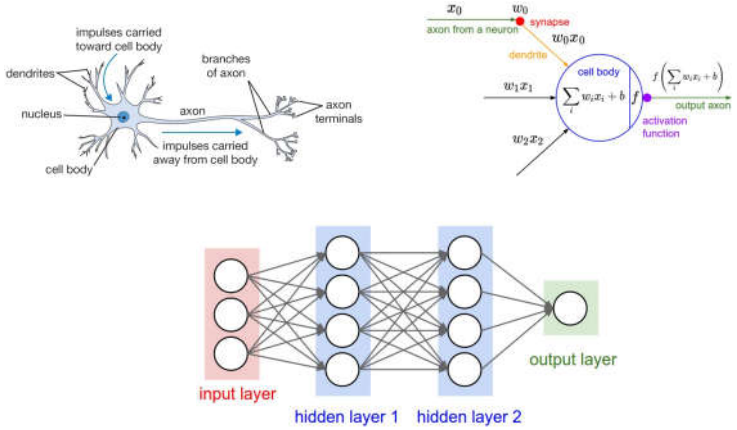
- Usa uma função para definir o relacionamento entre os dados
- Pode assumir muitas formas
- Como escolher a função $f(X)$? Função de custo

Árvores de decisão

- Usa um conjunto simples de decisões hierárquicas para tomar decisões
- Simples de entender, rápido de executar
- O conjunto de árvores (floresta) perde em interpretabilidade mas ganha em acurácia
- Aprendizado pode ser baseado em entropia, que cria nós baseado no ganho de informação

Rede Neural

- Vagamente inspirado por um modelo biológico
- Usa conexões de neurônios para calcular valores e escolher classes
- Altamente eficiente em tarefas de percepção
- O uso de Redes Neurais atende por outro nome: *Deep Learning*



Aprendizado não-supervisionado

- Dado um conjunto de dados $\{x_1, x_2, \ldots, x_n\}$ um algoritmo de aprendizado procura estruturas latentes
- Tarefa comum: clusterização
- Maior parte da informação disponível hoje não tem rótulo

Clusterização

- Mesmo sem rótulo, encontrar meios de agrupar informação tem muitas utilidades
- Exemplos:
  - Identificação de células cancerígenas
  - Detecção de fraudes e dispor linhas de crédito
  - Propaganda
  - Redes sociais
  - Análise de astros

K-Means

- Método simples de agrupamento, útil quando se sabe um pouco do domínio
- Parâmetro k define o número de clusters, que são definidos em razão de um centroide.
- Os pontos se agrupam nos clusters pela distância até os centroides, que são atualizados em cada etapa
- Muito sensível

DBSCAN

- Procura clusters pela densidade de pontos
- Parâmetros $\epsilon, minPoints$: distância e densidade esperada
- Os pontos se agrupam nos clusters pela distância até os centroides, que são atualizados em cada etapa
- Um pouco mais robusto, mas depende muito da densidade