CURSO SUPERIOR DE ADS

Machine Learning

Prof. Fernando Marlon Soares Figueiredo

Disciplina: Ciência de Dados e Bigdata

SECTRAS

ADS

CRONOGRAMA

- MACHINE LEARNING – CONCEITOS IMPORTANTES
- MACHINE LEARNING NA PRÁTICA
- DISCUSSÃO TRABALHO 1

1. O que é Aprendizado de Máquina?

- Aprendizado de máquina (Machine Learning) é a técnica de usar algoritmos para analisar dados, aprender com esses dados e, em seguida, fazer previsões ou tomar decisões com base no aprendizado.
- A ideia central é construir modelos que possam generalizar a partir de dados observados para novas entradas.

REVISÃO

1.Definição de Aprendizado de Máquina:

1. Técnica que usa algoritmos para aprender com dados e fazer previsões ou tomar decisões automáticas.

2.Tipos de Aprendizado:

1. **Supervisionado:** Dados rotulados, ex.: classificação (spam/não spam).
2. **Não Supervisionado:** Descobre padrões em dados sem rótulos, ex.: agrupamento.
3. **Reforço:** Aprende por tentativa e erro, ex.: controle de veículos autônomos.

3.Modelagem:

1. Construção de modelos matemáticos com dados de treino e validação com dados de teste, equilibrando complexidade e generalização.

4.Sobreajuste e Subajuste:

REVISÃO

1. **Sobreajuste:** Modelo muito complexo, ajusta ao ruído dos dados de treino.
2. **Subajuste:** Modelo muito simples, não captura a complexidade dos dados.

1.Compromisso entre Polarização e Variância:

1. **Polarização (bias):** Simplificação excessiva que causa erros sistemáticos.
2. **Variância:** Modelo se adapta demais aos dados, resultando em erros futuros.

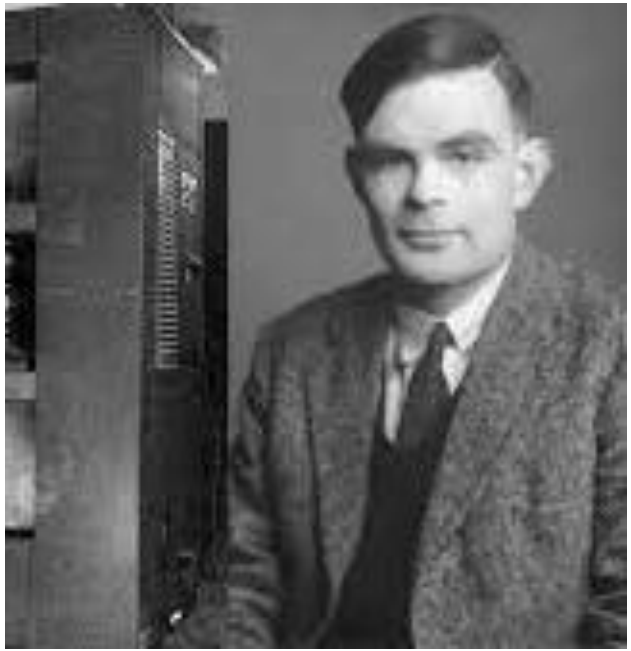
2.Algoritmos Comuns:

1. Regressão Linear, Árvores de Decisão, KNN, Redes Neurais, entre outros, usados para diversos problemas de predição e classificação.

- Esses seis tópicos cobrem o essencial sobre aprendizado de máquina de maneira resumida e direta.

Origens Machine Learning

- As origens do Machine Learning remetem à década de 1950, quando Alan Turing, reconhecido como pai da computação e um dos homem-chave da Segunda Guerra Mundial, desenvolveu testes para experimentar a capacidade das máquinas de “raciocinarem”.



O JOGO DA IMITAÇÃO

MACHINE LEARNING

- Baseado nos trabalhos de Turing, o engenheiro Arthur Samuel conseguiu, apenas dois anos depois, desenvolver o primeiro programa capaz de aprender: um jogo de damas em que o sistema aprimorava seu desempenho a cada partida que realizava.
- Foi o próprio Samuel que, em 1959, usou pela primeira vez o termo “Machine Learning”.

- ***“ Fiquei tão intrigado com esse problema geral de escrever um programa que parecesse exibir inteligência que ele ocuparia meus pensamentos durante quase todos os momentos livres durante todo o período em que trabalhei na IBM e, na verdade, por alguns anos além disso .”*** **— Arthur Samuel**

- No computador IBM 704, Arthur implementou o primeiro algoritmo de poda alfa-beta. Semelhante ao que vemos em algoritmos de Aprendizado por Reforço, ele implementou uma função de perda que calcularia a probabilidade de ganhar o jogo com base na posição atual.
- A função levou em conta vários fatores, como o número de peças de cada lado, o número de reis e a proximidade das peças de se tornarem reis.
- O IBM 704 realizou um bilhão de cálculos por dia ao calcular a órbita de um satélite artificial.

História: como surgiu o Machine Learning?

1. **Arthur Samuel e o Início do Aprendizado de Máquina:**
    1. Pioneiro no campo da aprendizagem de máquina.
    2. Desenvolveu o primeiro programa de aprendizado de máquina e cunhou o termo “Machine Learning”.
2. **Inverno da IA (décadas de 70 e 80):**
    1. Período de desinteresse e falta de investimentos em IA e aprendizado de máquina.
3. **Fim do Inverno da IA - Avanços:**
    1. **NETtalk (1985):** Rede Neural Artificial que aprendeu a pronunciar textos em inglês.

História: como surgiu o Machine Learning?

1. **Deep Blue (1997):** Algoritmo que derrotou Garry Kasparov no xadrez.

1.Elastic Compute Cloud (EC2) - Amazon (década de 2000):

1. Início da comercialização de máquinas virtuais, permitindo maior poder de processamento.

2.Avanços com Elastic Compute Cloud:

1. Possibilitou a implementação de modelos de ML mais complexos e estudos avançados de dados.

3.Impacto Atual:

1. Esses avanços permitiram a evolução contínua no campo de Machine Learning e IA, que vemos nos dias atuais.

DEEP BLUE

- **Deep Blue e o Xadrez:**
- **Deep Blue** foi um supercomputador desenvolvido pela IBM, projetado especificamente para jogar xadrez.
- Em **1997**, tornou-se famoso ao derrotar o campeão mundial de xadrez **Garry Kasparov**, marcando a primeira vez que uma máquina venceu um campeão mundial em uma competição oficial.
- O Deep Blue utilizava força bruta de processamento, analisando milhões de posições de xadrez por segundo.

Go - Um Novo Desafio para a IA

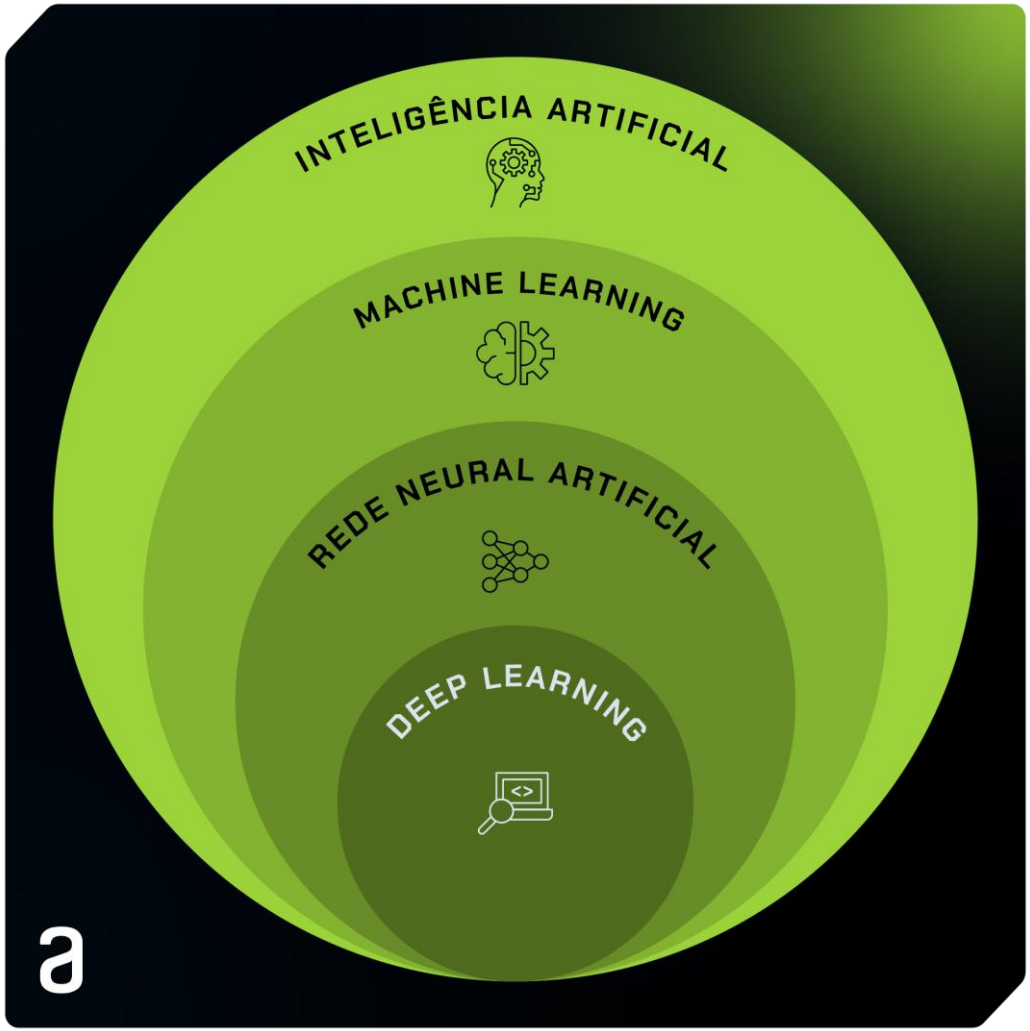
- O **Go** é um jogo de tabuleiro muito mais complexo que o xadrez, com um número de combinações possível muito maior.
- Durante anos, acreditava-se que o Go seria difícil para a IA, devido à sua complexidade estratégica e o grande número de movimentos possíveis.

AlphaGo e a Revolução no Go:

- Em **2016**, a IA chamada **AlphaGo**, desenvolvida pela DeepMind (empresa do Google), derrotou o campeão mundial de Go, **Lee Sedol**, em uma série histórica de partidas.
- Diferente do Deep Blue, que usava força bruta, o AlphaGo utilizava redes neurais e aprendizado profundo (deep learning), sendo capaz de aprender estratégias a partir de milhões de partidas anteriores.

Impacto do AlphaGo:

- A vitória do AlphaGo foi um marco no campo da inteligência artificial, demonstrando o poder de algoritmos de aprendizado de máquina avançados.
- Este evento consolidou o uso de técnicas modernas de aprendizado de máquina para resolver problemas que anteriormente eram considerados muito complexos para computadores.


INTELIGÊNCIA ARTIFICIAL
MACHINE LEARNING
REDE NEURAL ARTIFICIAL
DEEP LEARNING
a

GOFAI

- Embora o nome "Inteligência Artificial" sugira que sempre terá um algoritmo de Machine Learning, nesse campo também há aplicações que não utilizam ML.

- Um subconjunto em questão é chamado de GOFAI (Good Old-Fashioned Artificial Intelligence), ou "boa e velha IA".

- Nele os sistemas de IA se baseiam em regras e lógicas programadas manualmente por pessoas desenvolvedoras, sem depender da capacidade de aprender com dados.

- Um exemplo clássico de GOFAI são os chatbots simples que respondem a perguntas fornecendo respostas pré-programadas, muito utilizados em serviços de atendimento ao cliente.

Deep Learning, Redes Neurais e Machine Learning

- **Deep Learning, Redes Neurais e Machine Learning**
- Machine Learning, Deep Learning e Redes Neurais são conceitos que se relacionam entre si, cada um desempenhando um papel específico no campo da Inteligência Artificial.
- O Machine Learning é o campo mais amplo entre os três, representando um **conjunto de algoritmos** que permitem a um sistema aprender padrões a partir de dados e tomar decisões com base nesse aprendizado.
- Existem uma diversidade de algoritmos de ML e alguns deles estão dentro do campo do Deep Learning.

Redes Neurais

- As Redes Neurais se caracterizam por serem uma técnica de aprendizagem dentro do domínio mais amplo do Machine Learning.
- Ela é uma estrutura computacional composta por nós computacionais (os “neurônios”), que se organizam em camadas. Essas camadas incluem uma camada de entrada, uma camada de saída e, potencialmente, camadas ocultas.
- Quando a rede neural tem mais camadas além da entrada e da saída (ou seja, mais de três ou mais camadas), é classificada como "profunda". Nesse caso, é um exemplo de Aprendizado Profundo, ou **Deep Learning**.

Redes Neurais

- **Redes Neurais**
- Originadas de simples Perceptron, as Redes Neurais Artificiais são uma arquitetura fundamental que serve como base para redes neurais mais avançadas. Mesmo existindo Redes Neurais Simples, são as Redes Neurais focadas em **Aprendizado Profundo** que carregam a fama das Redes Neurais. Vamos conhecer as mais notáveis:
- **Redes Neurais Feedforward (FF)**
- As Redes Neurais Feedforward são uma das formas mais antigas de redes neurais. Nesse tipo de arquitetura, os dados fluem em **uma única direção**, passando por diferentes camadas de neurônios artificiais até que a saída desejada seja alcançada.

- Cada camada processa as informações e as transfere para a próxima camada sem formar ciclos de realimentação, tornando-as eficazes em tarefas como **classificação e regressão**, onde os dados podem ser representados como entradas independentes e a ordem das entradas não possui um significado especial, como dados não sequenciais.
- **Redes Neurais Recorrentes (RNN)**
- As Redes Neurais Recorrentes são projetadas para **lidar com dados sequenciais**, como séries temporais. Elas introduzem a ideia de "memória" nas redes neurais, onde as saídas anteriores influenciam as saídas futuras. No entanto, as RNNs enfrentam desafios de retenção de informações a longo prazo.
- **Memória de Longo/Curto Prazo (LSTM)**
- As LSTMs são uma forma avançada de RNN que aborda o problema de curto alcance de retenção de informações. Elas são capazes de **"lembrar" eventos importantes de camadas anteriores**, permitindo a manipulação eficaz de dados sequenciais mais extensos.

- Essa capacidade de retenção de informações de longo prazo faz com que as LSTMs se destaquem em **tarefas complexas, como tradução automática e geração de texto buscando prever a próxima letra**.
- **Redes Neurais Convolucionais (CNN)**
- As Redes Neurais Convolucionais são amplamente empregadas em tarefas de Visão Computacional.
- Com uma arquitetura composta por camadas convolucionais e de pooling, as CNNs são eficientes em **extrair padrões espaciais em dados, sendo ideais para o processamento de imagens**.

- Elas aplicam filtros para identificar características relevantes antes de passar para camadas totalmente conectadas para tomadas de decisão.
- **Redes Adversárias Generativas (GAN)**
- As Redes Adversárias Generativas levam uma abordagem que envolve duas redes neurais, **uma geradora e uma discriminadora, competindo entre si**. A geradora cria dados que tentam ser idênticos aos dados reais, enquanto a discriminadora tenta diferenciá-los.
- Os dados criados e discriminados podem ser **imagens, áudios, vídeos e até textos**, tudo depende do objetivo da GAN. Essa competição resulta em um aprendizado mais refinado, sendo utilizado em aplicações como geração de imagens e tradução de estilo.

Quadro de Vantagens e Desvantagens do Machine Learning

Vantagens	Desvantagens
ML pode analisar grandes volumes de dados e identificar padrões complexos.	Pode amplificar preconceitos se os dados de treino forem enviesados.
Automatiza tarefas repetitivas, aumentando a eficiência operacional.	Modelos complexos podem ser difíceis de interpretar, dificultando a transparência.
Algoritmos podem ser adaptados com novos dados, tornando-se flexíveis.	A qualidade dos dados afeta diretamente a precisão das previsões.
Eficaz em reconhecer padrões e prever tendências futuras.	Preocupações sobre privacidade com o uso de grandes volumes de dados pessoais.

APRENDIZADO SUPERVISIONADO

- Como seres humanos, classificamos animais como porcos ou cachorros com base em suas características.
- Temos, então, um problema de classificação com duas categorias, ou classes: porco e cachorro.
- Aprendemos a fazer essa distinção porque, desde cedo, alguém nos ensinou: "Isso é um porquinho, isso é um cachorro". Esse aprendizado supervisionado nos ajudou a diferenciar entre as duas classes.

MACHINE LEARNING NA PRÁTICA

- **Classificação binária: duas classes**
- classificar porcos e cachorros é um exemplo de **classificação binária**, onde temos dois valores possíveis, como 0 ou 1. Da mesma forma, podemos classificar e-mails como "spam" ou "não spam", que também é um problema de classificação binária.

- Se o primeiro porco tem pelo longo, marcamos 1 para indicar "sim". Se o segundo porco não tem pelo longo, marcamos 0 para indicar "não". Fazemos o mesmo para o terceiro porco e para os cachorros. Assim, usamos 0 ou 1 para indicar se uma característica está presente ou não em cada animal. Além disso, temos uma coluna que nos diz se o animal é um cão ou um porco.

	PELO LONGO?	0 CÃO 1 PORCO
	1	1
	0	1
	0	1
	0	0
	1	0
	0	0

	PELO LONGO?	PERNA CURTA?	O CÃO 1 PORCO
	1	1	0
	0	1	1
	0	1	0
	0	1	1
	1	1	1
	0	0	1

	PELO LONGO?	PERNA CURTA?	AU AU?	0 CÃO 1 PORCO
	1	1	0	1
	0	1	1	1
	0	1	0	1
	0	1	1	0
	1	1	1	0
	0	0	1	0

CLASSIFICAÇÃO EM APRENDIZADO SUPERVISIONADO

- Portanto, para cada animal, temos três características: pelo longo, perna curta e se faz au-au. E temos a classificação desses animais, pois estamos trabalhando com problemas de classificação em aprendizado supervisionado.

- **2. Definindo Dados de Treinamento**
- No notebook, criaremos variáveis que representam animais, porcos e cachorros, com três características: pelo longo, perna curta e faz "au, au".

```
porco1 = [0, 1, 0]
porco2 = [0, 1, 1]
porco3 = [1, 1, 0]
cachorro1 = [0, 1, 1]
cachorro2 = [1, 0, 1]
cachorro3 = [1, 1, 1]
dados = [porco1, porco2, porco3, cachorro1, cachorro2, cachorro3]
classes = [1, 1, 1, 0, 0, 0]
```

LinearSVC

- **1. O que é LinearSVC?**
- O **LinearSVC** é um algoritmo da família de **Máquinas de Vetores de Suporte (SVM)**, usado para problemas de classificação. Ele utiliza uma função linear para separar os dados em diferentes classes, no nosso caso, porcos e cachorros.
- **2. Coleta de Dados e Características**
- Inicialmente, coletamos dados de porcos e cachorros com diferentes características (ex.: pelo longo, perna curta, etc.).

LinearSVC

- Cada exemplo de dado contém um conjunto de características e um rótulo que indica a classe a que pertence (1 para porco, 0 para cachorro).
- **3. Treinamento e Hiperplano**
- Durante o **treinamento**, o LinearSVC tenta encontrar o **hiperplano** ideal que melhor separa as duas classes.
- O **hiperplano** é uma linha no caso de duas características, ou um plano em espaços com mais dimensões. O objetivo é encontrar a posição que maximiza a distância entre os pontos mais próximos de cada classe, chamados de **vetores de suporte**.

LinearSVC

- **4. Classificação de Novos Dados**
- Após o treinamento, o modelo pode classificar novos exemplos.
- Se o novo dado estiver de um lado do hiperplano, será classificado como **porco**, e se estiver do outro lado, será classificado como **cachorro**.
- **5. Vantagens do LinearSVC**
- **Simples e eficiente**: Funciona bem para dados lineares, ou seja, onde as classes podem ser separadas por uma linha reta.
- **Escalabilidade**: Pode ser utilizado em grandes conjuntos de dados com muitas características.

LinearSVC

- **6. Limitações**
- **Dados não lineares**: O LinearSVC pode não funcionar bem em dados mais complexos que não possam ser separados linearmente. Nesses casos, é possível usar outras variações do SVM que utilizam **funções kernel** para transformar os dados e encontrar uma separação não linear.

ATIVIDADE

- Leitura e resumo do seguinte artigo:

- https://medium.com/liga-mackenzie-de-ia-ci%C3%AAncia-de-dados/svm-ou-support-vector-machine-7efcabdcc7be

SEMINÁRIO

- **Objetivo:**
- Os alunos deverão utilizar o algoritmo **SVM (Support Vector Machine)** para resolver problemas de classificação com base em grandes conjuntos de dados. O seminário consistirá em treinar a máquina utilizando diferentes datasets, analisar os resultados, e compilar as conclusões em uma apresentação formal.
- **Tarefas a serem realizadas:**

1.Escolha de Dataset:

1. Cada grupo deverá selecionar um dos datasets disponíveis, realizar a limpeza dos dados (se necessário) e utilizá-los para treinar um modelo de classificação utilizando SVM.

2.Treinamento do Modelo:

1. Aplicar o algoritmo SVM nos dados de treino e validar o desempenho nos dados de teste.
2. Utilizar bibliotecas como **scikit-learn** no ambiente **Google Colab** ou local.

3.Análise dos Resultados:

SEMINÁRIO

1. Avaliar métricas de desempenho, como **acurácia**, **precisão**, **recall**, e **F1-score**.
2. Verificar a influência de hiperparâmetros, como **kernel** (linear, polinomial, RBF) e **C** (parâmetro de regularização).

1.Compilação dos Resultados:

1. Cada grupo deverá preparar uma apresentação com os seguintes itens:
    1. Descrição do dataset e do problema abordado.
    2. Método de pré-processamento utilizado.
    3. Processo de treinamento e análise das métricas de desempenho.
    4. Discussão sobre os resultados obtidos e possíveis melhorias no modelo.
    5. Conclusão sobre a eficácia do SVM para resolver o problema proposto.

Critérios de Avaliação

- **Critérios de Avaliação:**
- Os grupos serão avaliados de acordo com os seguintes critérios:

1. Qualidade do Treinamento do Modelo (30%)

1. Aplicação correta do SVM no dataset.
2. Ajuste dos parâmetros e escolha adequada do kernel.

2. Análise das Métricas de Desempenho (20%)

1. Avaliação completa dos resultados com acurácia, precisão, recall, e F1-score.

3. Exploração de Hiperparâmetros (15%)

1. Teste e análise do impacto de diferentes valores de **kernel** e **C** no modelo.

Critérios avaliação

1. **Clareza e Organização da Apresentação (20%)**
    1. Explicação detalhada e clara dos métodos utilizados, resultados e conclusões.
2. **Discussão e Propostas de Melhoria (15%)**
    1. Reflexão crítica sobre os resultados e sugestões para otimizar o desempenho do modelo.

- **Entrega e Apresentação:**
- **Data de entrega do relatório e código**: 29/04
- **Data da apresentação**: 22/04
- Cada grupo terá 25 minutos para apresentar seus resultados, seguidos de 5 minutos de perguntas e respostas.

OPÇÕES DE DATASET

- Escolha um dos cinco datasets abaixo para o desenvolvimento do seminário:

1. CIFAR-10 Dataset

1. Descrição: Conjunto de imagens coloridas de 32x32 pixels, categorizadas em 10 classes, incluindo aviões, automóveis, pássaros, gatos, etc.
2. Tamanho: Grande (60.000 imagens de treino e teste).

2. MNIST - Digits Classification

1. Descrição: Conjunto de imagens de dígitos manuscritos (0-9) para reconhecimento de dígitos.
2. Tamanho: Médio (60.000 imagens de treino).

OPÇÕES DE DATASET

3. **Titanic: Machine Learning from Disaster**
    1. Descrição: Base de dados sobre passageiros do Titanic, com o objetivo de prever se um passageiro sobreviveu ou não, com base em características como idade, sexo e classe de viagem.
    2. Tamanho: Médio.
4. **Breast Cancer Wisconsin Dataset**
    1. Descrição: Conjunto de dados de imagens de tumores para classificação de câncer como benigno ou maligno.
    2. Tamanho: Médio.
5. **Fashion MNIST**
    1. Descrição: Conjunto de imagens de roupas classificadas em 10 categorias (sapatos, camisas, etc.).
    2. Tamanho: Grande (60.000 imagens de treino).