ICP363
Revisão: Uma Introdução ao Aprendizado de Máquina

O que é Aprendizado de Máquina?

Definição:

O que é Aprendizado de Máquina?

Definição:

Uma subárea da Inteligência Artificial que desenvolve algoritmos capazes de "aprender" a partir de dados. O objetivo é que um sistema melhore sua performance através da experiência.

O que é Aprendizado de Máquina?

Definição:

Uma subárea da Inteligência Artificial que desenvolve algoritmos capazes de "aprender" a partir de dados. O objetivo é que um sistema melhore sua performance através da experiência.

Como isso funciona?

O que é Aprendizado de Máquina?

Definição:

Uma subárea da Inteligência Artificial que desenvolve algoritmos capazes de "aprender" a partir de dados. O objetivo é que um sistema melhore sua performance através da experiência.

Como isso funciona?

A ideia é que a máquina adquira conhecimento através de estudo, observação ou experiência. Isso nos permite resolver problemas onde é difícil antecipar todas as situações possíveis, como dirigir um carro ou reconhecer um objeto em uma foto.

O que é Aprendizado de Máquina?

Principais Subáreas:

- ***Inteligência Artificial (IA):*** O campo geral da construção de sistemas inteligentes.
- ***Aprendizado de Máquina (ML):*** A capacidade dos algoritmos de aprender com dados.
- ***Aprendizado Profundo (DL):*** Uma subárea do ML que usa modelos com múltiplas camadas de processamento para aprender representações de dados com vários níveis de abstração.

Tipos de Aprendizado

Aprendizado Supervisionado

Aprendizado não supervisionado

Aprendizado por reforço

Tipos de Aprendizado - Supervisionado

***Conceito*:** Treina algoritmos com dados rotulados, onde cada entrada tem uma saída correspondente correta. O algoritmo ajusta sua precisão através de uma função de perda até que o erro seja minimizado.

Tipos de Problemas:

***Regressão*:** Prever um valor contínuo. Exemplo: prever o preço de uma casa.

***Classificação*:** *Atribuir* dados a categorias específicas. Exemplo: classificar um e-mail como spam ou não spam.

Tipos de Aprendizado - Não supervisionado

Conceito: O algoritmo explora dados sem rótulos, buscando padrões e estruturas por conta própria.

Exemplo: *Agrupamento* (Clustering) para agrupar dados similares com base em suas características.

Tipos de Aprendizado - Por Reforço

Conceito:

Aprende através de recompensas e penalidades, permitindo que um agente interaja com um ambiente para maximizar uma recompensa total.

Perceptron - Supervisionado

Definição:

Um algoritmo de aprendizado de máquina supervisionado para classificação binária.

Componentes-Chave:

Entradas (x1,x2,...,xn): Valores de entrada para o modelo.
Pesos (w1,w2,...,wn): Coeficientes que representam a importância de cada entrada.
Viés (Bias, b): Um valor que ajusta o limiar de ativação do neurônio.
Soma Ponderada: A soma de cada entrada multiplicada pelo seu peso correspondente, mais o viés: $\sum_{i=1}^{n} w_i x_i + b$.
Função de Ativação: Uma função que decide a saída final (0 ou 1) com base na soma ponderada.

Perceptron (AND) - Supervisionado

Entrada x1	Entrada x2	Saída Esperada y
0	0	0
0	1	0
1	0	0
1	1	1

Inicialização

Pesos (weights): w1=0, w2=0

Viés (bias): b= -1.2

Taxa de Aprendizagem (learning rate): α=0.5

Perceptron (AND) - Supervisionado

1. Verificando os Três Primeiros Pontos:

Com os valores iniciais w1=0, w2=0 e b=−1.2, a soma ponderada para os três primeiros pontos será:

- Ponto (0, 0): $(0 \cdot 0)+(0 \cdot 0)+(-1.2)=-1.2$ -> Previsão **0** (correta).
- Ponto (0, 1): $(0 \cdot 0)+(0 \cdot 1)+(-1.2)=-1.2$ -> Previsão **0** (correta).
- Ponto (1, 0): $(0 \cdot 1)+(0 \cdot 0)+(-1.2)=-1.2$ -> Previsão **0** (correta).

Perceptron (AND) - Supervisionado

Todos os três primeiros pontos são classificados corretamente. **Não há ajustes**.

2. A Única Atualização Necessária:

- **Ponto (1, 1), Saída Esperada = 1.**
- **Soma Ponderada:** $(0 \cdot 1)+(0 \cdot 1)+(-1.2)=-1.2$.
- **Previsão:** $-1.2 \leq 0$, então a previsão é **0**.
- **Erro:** A previsão está incorreta (0=1). Vamos ajustar os parâmetros com $\alpha=0.5$.

3. O Ajuste Perfeito:

- $w1 = w1 + \alpha(y-y')x1 = 0+0.5(1-0) \cdot 1 = 0.5$
- $w2 = w2 + \alpha(y-y')x2 = 0 + 0.5(1-0) \cdot 1 = 0.5$
- $b = b + \alpha(y-y') = -1.2 + 0.5(1-0) = -0.7$

Ao final da primeira época, os novos parâmetros do modelo são: **w1=0.5, w2=0.5, b=−0.7**.

Regressão Linear - Supervisionado

Definição:

Modelo estatístico que examina a relação linear entre duas ou mais variáveis.

Regressão Linear Simples:

$$Y \approx \beta_1 X + \beta_0$$

Encontrar a reta que melhor se ajusta a um conjunto de dados de entrada (X) e saída (Y)

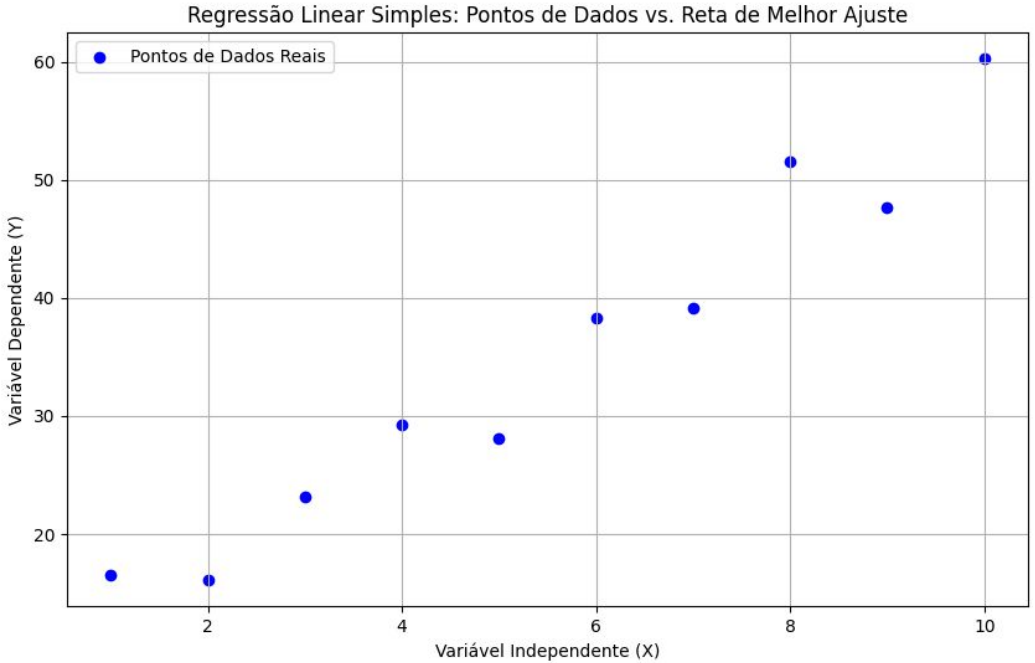
- β_0 **- (Intercepto):** O valor de Y quando X é zero.
- β_1 **- (Slope):** O quanto Y muda para cada unidade que X muda

Regressão Linear-Encontrar a melhor reta

Calcular a Soma Residual dos Quadrados (RSS) -> soma dos quadrados dos resíduos

$$\sum(y\_real - y\_previsto)^2$$

X	1	2	3	4	5	6	7	8	9	10
Y	20.4	19.7	27.7	28.7	32.8	43.6	40.2	52.6	55.4	61.1



$Y = aX+c$, $a = 5$, $c = 10$

$Y = 5X + 10 + ruído(normal) \rightarrow N(0,3)$

X	1	2	3	4	5	6	7	8	9	10
Y	20.4	19.7	27.7	28.7	32.8	43.6	40.2	52.6	55.4	61.1

$Y' = b^1X + b^0 + err$

$y\_i = b^1x\_i + b^0 + err\_i$, $i = \{1,...,10\}$

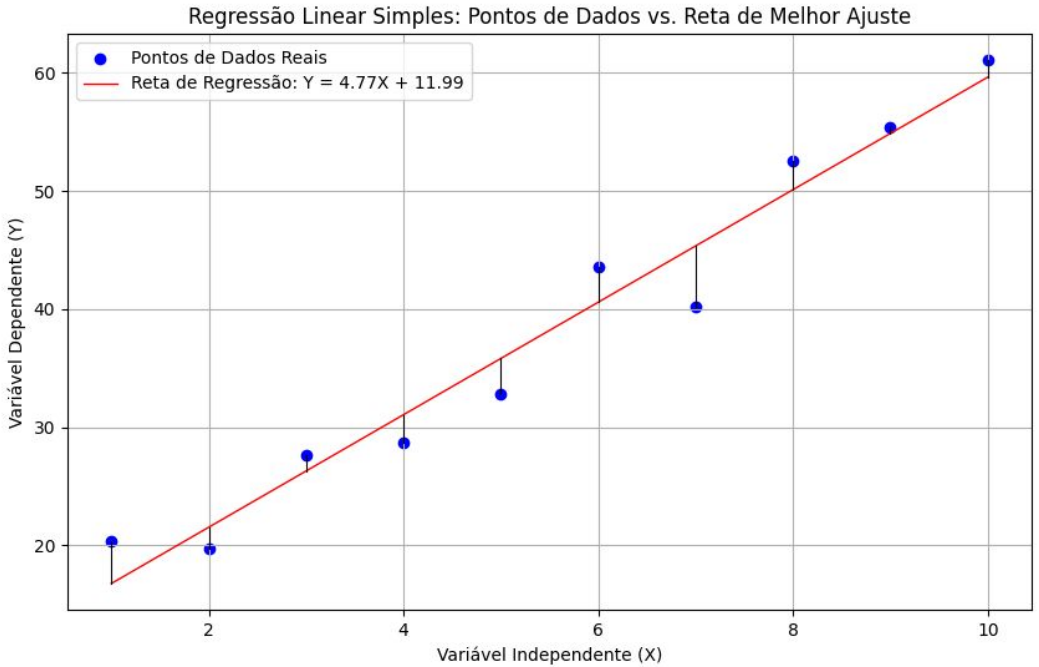
$err\_i = (y\_ireal - y\_iprev)$

$Erro = sum(err\_i)^2$ [RSS - Soma dos resíduos quadrados]

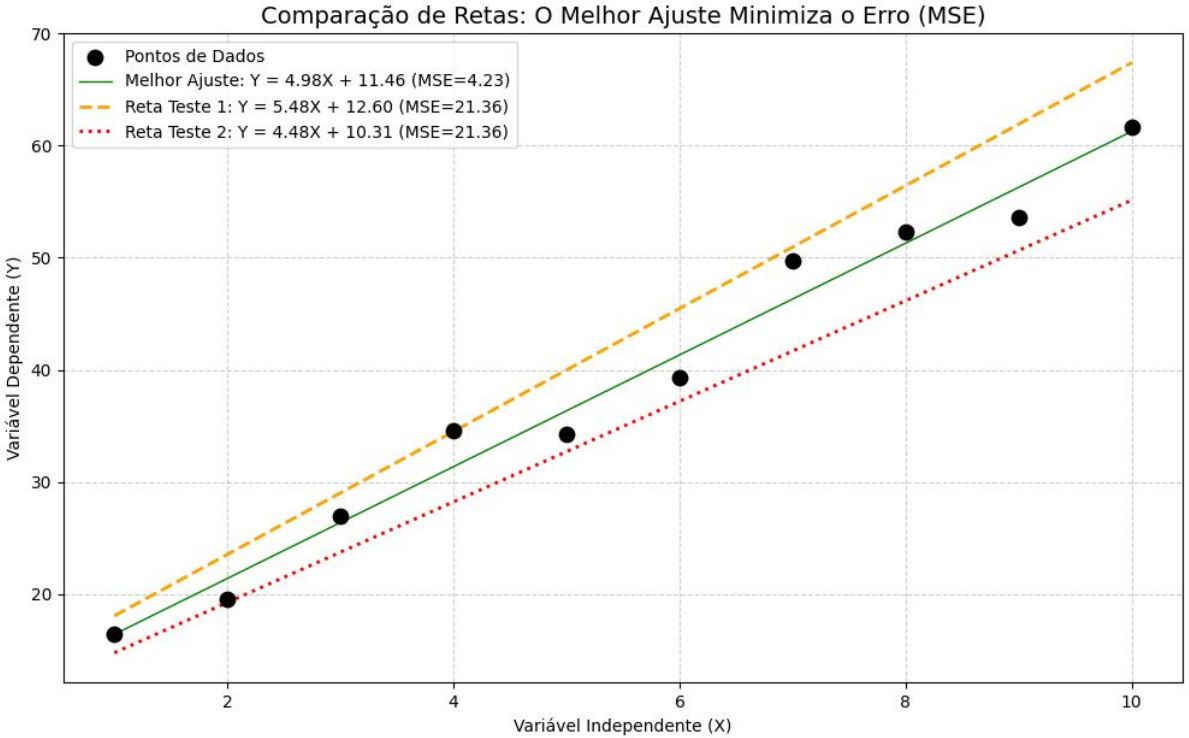
$$\hat{\beta}_1 = \frac{n * \sum_{i=1}^{n} y_i * x_i - \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i}{n * \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i * \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i * x_i}{n * \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

12



Comparação de Retas



Regressão Linear com Múltiplas Variáveis

Conceito:

Quando temos mais de uma variável de entrada. A equação é generalizada para a forma matricial $Y \approx X\hat{\beta}$

Matriz de Entradas (X):

Inclui uma coluna de 1s para o intercepto, permitindo que o modelo aprenda o termo independente. Sem essa coluna, a reta seria forçada a passar pela origem, o que limita o modelo.

- **Dataset**: $\{(x_1^1, x_1^2, \cdots, x_1^k, y_1), (x_2^1, x_2^2, \cdots, x_2^k, y_2), \cdots, (x_n^1, x_n^2, \cdots, x_n^k, y_n)\}$
- **Atributos de entrada**:
  - X^1 com valores: $x_1^1, x_2^1, ..., x_n^1$
  - X^2 com valores: $x_1^2, x_2^2, ..., x_n^2$
  - $\cdots$
  - X^k com valores: $x_1^k, x_2^k, ..., x_n^k$
- **Atributos de saída**: Y com valores: $y_1, y_2, ..., y_n$.
- $$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2^1 & x_2^2 & \cdots & x_2^k \\ \cdots & & & & \\ 1 & x_n^1 & x_n^2 & \cdots & x_n^k \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_k \end{bmatrix}$$
- $Y \approx X\hat{\beta}$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \approx X\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 + x_1^1\hat{\beta}_1 + x_1^2\hat{\beta}_2 + \cdots + +x_1^k\hat{\beta}_k \\ \hat{\beta}_0 + x_2^1\hat{\beta}_1 + x_2^2\hat{\beta}_2 + \cdots + +x_2^k\hat{\beta}_k \\ \cdots \\ \hat{\beta}_0 + x_n^1\hat{\beta}_1 + x_n^2\hat{\beta}_2 + \cdots + +x_n^k\hat{\beta}_k \end{bmatrix}$$

Regressão Linear com Múltiplas Variáveis

$Y \approx X\hat{\beta}$

$X^T Y \approx (X^T X)\hat{\beta}$, onde X^T é a matriz transposta de X

$(X^T X)^{-1} X^T Y \approx (X^T X)^{-1}(X^T X)\hat{\beta}$, onde $(X^T X)^{-1}$ é a inversa da matriz $X^T X$

$(X^T X)^{-1} X^T Y \approx I\hat{\beta}$, onde I é a matriz identidade

$\hat{\beta} \approx (X^T X)^{-1} X^T Y$

Regressão Linear com Múltiplas Variáveis

Exemplo Prático (Publicidade): Usando o dataset Advertising.csv.

Regressão Simples: Compare a correlação entre TV e sales com a de radio e sales.

sales = 0.048 * TV + 7.033

sales = 0.203 * radio + 9.312

Regressão Múltipla: Mostre como a relação de newspaper com sales muda quando TV e radio também estão no modelo, revelando que o gasto com jornais tem pouca influência quando os outros já são considerados.

sales = 0.046 * TV + 0.189 * radio - 0.001 * newspaper + 2.939.

Preparação de Dados e Seleção de Features

Por que preparar os dados?

Dados de fontes heterogêneas podem conter problemas como inconsistências e falta de valores. Dados de qualidade são cruciais para análises confiáveis

Tarefas de Pré-processamento:

- **Limpeza:** Lidar com dados ausentes ou outliers.
- **Integração:** Unir dados de diferentes fontes.
- **Redução:** Diminuir o tamanho do dataset sem perder qualidade, como a redução de dimensionalidade.
- **Transformação:** Alterar a representação dos dados, como a normalização.

Preparação de Dados e Seleção de Features

Seleção de Features: O processo de escolher as variáveis de entrada mais importantes para a previsão.

Vantagens: Reduz o overfitting, melhora o desempenho e diminui o tempo de treinamento.

Métricas de Correlação:

Pearson: Mede a correlação linear entre duas variáveis numéricas.

Spearman: Mede a correlação monotônica usando os postos dos valores, sendo menos sensível a outliers.

Matriz de Confusão e Métricas de Classificação

Matriz de Confusão: Uma tabela que resume o desempenho de um algoritmo de classificação, mostrando os acertos e erros do modelo.

Componentes da Matriz:

Verdadeiro Positivo (VP): O modelo previu corretamente a classe positiva.

Verdadeiro Negativo (VN): O modelo previu corretamente a classe negativa.

Falso Positivo (FP): O modelo previu positivo, mas estava incorreto (erro tipo I).

Falso Negativo (FN): O modelo previu negativo, mas estava incorreto (erro tipo II).

Matriz de Confusão e Métricas de Classificação

Acurácia: Mede a proporção de acertos totais.
(VP+VN)/(VP+VN+FP+FN)

Precisão: Foca nos acertos das previsões positivas.
VP/(VP+FP)

Revocação (Recall): Foca em quantos dos casos positivos reais o modelo capturou.
VP/(VP+FN)

Medida F1: A média harmônica entre Precisão e Revocação, útil para balancear ambos os erros.
F1 = 2 x (Precision x Recall) / (Precision + Recall)

Matriz de Confusão e Métricas de Classificação

a)Acurácia

b)Precisão

c) Revocação (Recall)

d) Medida F1

1. Base de dados de biometria digital para bancos
2. Base de dados que detecta câncer
3. Base de dados balanceada que classifica emails como spam e não spam
4. Base de dados para classificar avaliações de produtos como positivas ou negativas para uma empresa de e-commerce

Matriz de Confusão e Métricas de Classificação

a)Acurácia (3)

b)Precisão (1)

c) Revocação (Recall) (2)

d) Medida F1 (4)

1. Base de dados de biometria digital para bancos
2. Base de dados que detecta câncer
3. Base de dados balanceada que classifica emails como spam e não spam
4. Base de dados para classificar avaliações de produtos como positivas ou negativas para uma empresa de e-commerce

Matriz de Confusão e Métricas de Classificação

a)Acurácia (3) (Base de dados balanceada que classifica emails como spam e não spam.)

b)Precisão (1) (O custo de um falso positivo, permitir que a pessoa errada acesse a conta, é extremamente alto.)

c) Revocação (Recall) (2) (O custo de um falso negativo, não detectar o câncer quando ele existe, é fatal e o mais alto de todos.)

Matriz de Confusão e Métricas de Classificação

d) Medida F1 (4) (Base de dados para classificar avaliações de produtos como positivas ou negativas para uma empresa de e-commerce.)

Falso Positivo: Classificar uma avaliação negativa como "positiva". A empresa pode subestimar a insatisfação do cliente, o que é um problema grave.

Falso Negativo: Classificar uma avaliação positiva como "negativa". A empresa pode pensar que um produto é ruim, quando na verdade os clientes o amam, o que pode levar a decisões de negócios erradas.

Dúvidas?