Introdução ao Machine Learning

TIpos de dados, Contexto das aplicações, Aplicações em casos de Regressão e Classificação

João Pedro Andrade Gomes da Silva
IEEE Computational Intelligence Society Student Chapter
Universidade de Brasília

May 2, 2025

Sumário

Tipos de dados

- Quando enfrentamos um problema de Machine Learning, a primeira coisa que devemos nos atentar são aos tipos de dados que estamos tratando.
- Existem, em linhas gerais, 2 tipos de dados: Dados Numéricos e Dados Categóricos

Exemplos

A seguir, iremos citar alguns exemplos de cada tipo de dado

- Dados Númericos: Preço de um imóvel, Salário médio de uma população, Peso, Altura...
- Dados Categóricos: Estado de vida (morto/vivo), Tipo de contrato, Departamento de Trabalho, Gênero

Dados Númericos

Os Dados Númericos se dividem em dois tipos distintos

- Dados Discretos: São dados numéricos, representados por números inteiros não negativos. Sua principal característica é a finitude. (Ex: Número de bolas de futebol utilizadas ao longo de uma partida)
- Dados Contínuos: São dados numéricos que podem assumir qualquer valor dentro de um intervalo, e podem ser divididos em partes, infinitamente. (Ex: Velocidade, Pressão, Distância)

Dados Categóricos

Por sua vez, os Dados Categóricos possuem 3 sub-categorias:

- Dados Ordinais: Varíaveis categorizáveis cuja a ordem é de suma relevância (Ex: Nível Educacional)
- Dados Binários: Varíaveis categorizáveis que tendem a seguir a lógica booleana, ou seja, assumem apenas um valor (Ex: Falso/Verdadeiro, Estado de vida)
- Dados Nominais: Váriaveis Categorizáveis que não obedecem a determinada ordenação (Ex: Gênero, Tipo Sanguíneo)

Lidando com dados faltantes

Grande parte das vezes em que estamos trabalhando com Bases de Dados complexas, nem sempre as receberemos de maneira completa. Sendo necessária a realização de uma geração sintética de dados Para isso, é necessário entendermos os diferentes tipos de "dados perdidos".

- Missing Completly at Random (MCAR) - Todas as variáveis tem a mesma probabilidade de serem "perdidas"
- Missing at Random (MAR) - Varíaveis com probabilidades distintas de serem perdidas
- Missing not at Random (MNAR) - São aqueles cuja ausência está diretamente relacionada ao próprio valor ausente. Ou seja, a falta do dado ocorre porque há algo inerente à variável que influencia sua não resposta.

Técnicas para verificar dados faltantes

Função	Descrição
`.isnull()`	Retorna um DataFrame pandas, onde cada valor é um booleano: `True` se o valor estiver ausente, `False` caso contrário.
`.notnull()`	Similar à função anterior, mas retorna `False` se o valor for `NaN` ou `None`, e `True` caso contrário.
`.info()`	Gera três colunas principais, incluindo "Non-Null Count", que exibe a quantidade de valores não ausentes para cada coluna.
`.isna()`	Semelhante a `.isnull()`, mas retorna `True` apenas quando o valor ausente é do tipo `NaN`.

Table: Principais funções para identificar valores ausentes no pandas.

Técnicas para substituir dados faltantes

Existem várias estratégias de imputação, e elas não devem ser utilizadas de forma indiscriminada. A adoção da abordagem correta pode evitar a introdução de vieses nos dados e a tomada de decisões equivocadas.
A tabela a seguir ilustra qual método de imputação utilizar com base no tipo de dado ausente. A lista de métodos não é exaustiva, mas estes são os mais comumente usados.

Tabela - Técnicas para substituir dados faltantes

Tipo de Dado Ausente	Método de Imputação
Missing Completly at Random (MCAR)	Média, Mediana, Moda ou qualquer outro método de imputação.
Missing at Random (MAR)	Imputação Múltipla, Imputação por Regressão.
Missing not at Random (MNAR)	Substituição por Padrão, Estimativa de Máxima Verossimilhança.

Table: Métodos de imputação com base no tipo de dado ausente.

Definição e tipos de EDA

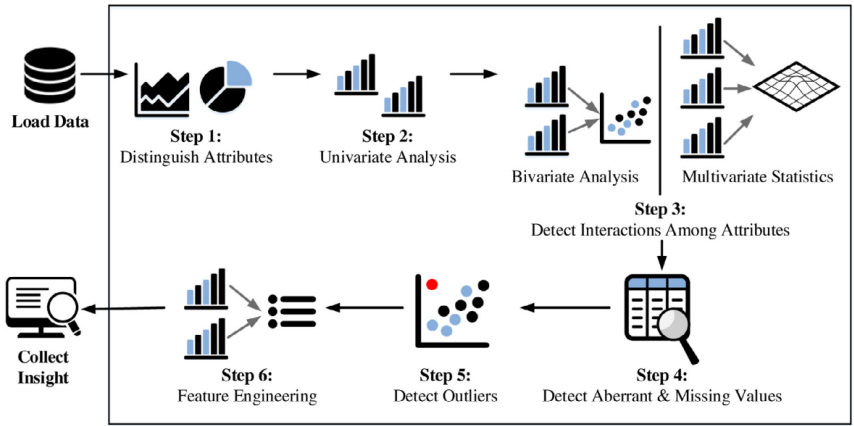

Figure: Fluxo lógico em análise exploratória de dados

Aprendizado Supervisionado

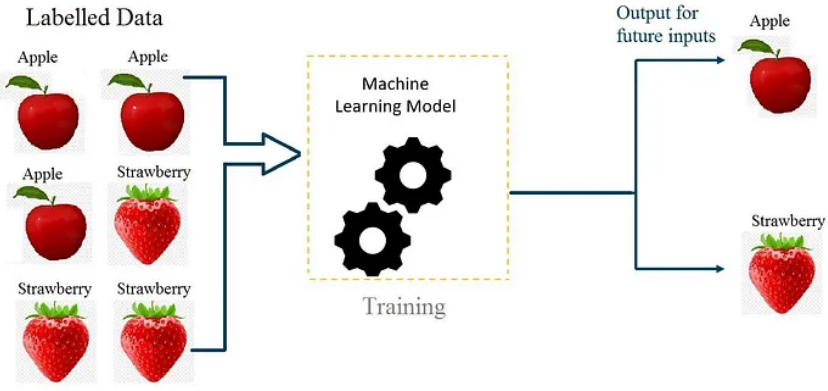

Figure: Aprendizado Supervisionado

Aprendizado Não Supervisionado

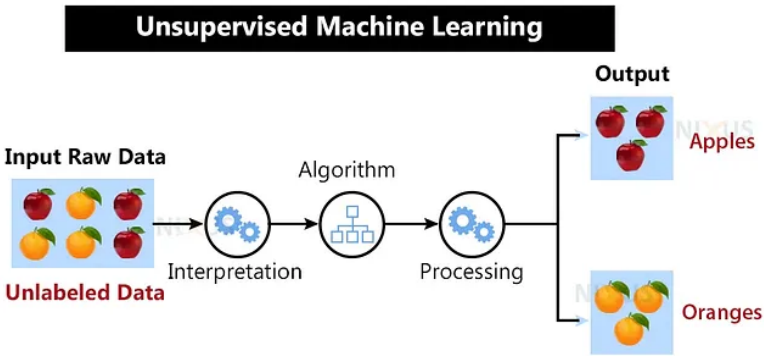

Figure: Aprendizado Não Supervisionado

Principais algoritmos

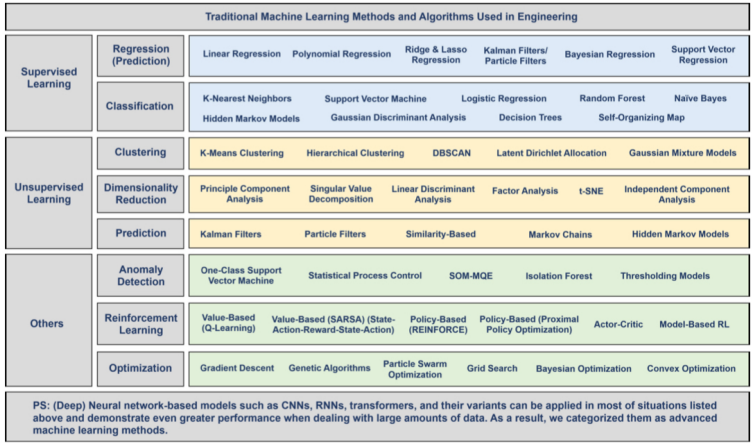

Figure: Estado da arte - Algoritmos e Métodos de ML usados na engenharia

Estado da arte - linhas de pesquisa

Research Direction	Feasibility	Impact	Leading Sector
Transfer learning & domain adaptation	High	Moderate	Academia + Industry
Similarity-based machine learning	High	Moderate	Academia + Industry
Synthetic data generation	Moderate	Moderate	Academia
Digital twin-based machine learning	Moderate	High	Academia + Industry
Stream-of-X	Moderate	High	Academia + Industry
Hybrid physics-based and data-driven models	Low	High	Academia
Multi-modal machine learning	Low	High	Academia
LLM and ILKM	Low	High	Academia + Industry
Foundation models for engineering AI	Very Low	Transformative	Academia + Government

Figure: Roteiro qualitativo para futuras direções de pesquisa em engenharia de IA

Definição

Para um problema de regressão, os dados de treinamento $\mathcal{D}_n$ estão na forma de um conjunto de n pares:

$$\mathcal{D}_n = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\},$$

onde $x^{(i)}$ representa uma entrada, geralmente um vetor d-dimensional de valores reais e/ou discretos, e $y^{(i)}$ é a saída a ser prevista, neste caso, um número real. Os valores de y são às vezes chamados de valores-alvo.
O objetivo em um problema de regressão é, dado um novo valor de entrada $x^{(n+1)}$, prever o valor de $y^{(n+1)}$. Problemas de regressão são um tipo de aprendizado supervisionado, pois a saída desejada $y^{(i)}$ é especificada para cada um dos exemplos de treinamento $x^{(i)}$.

Qual Método de Regressão usar?

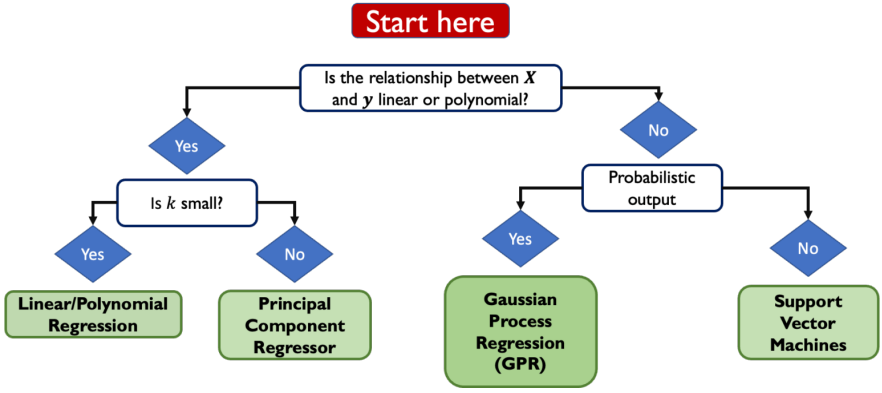

Figure: Métodos de Regressão

Regressão Linear

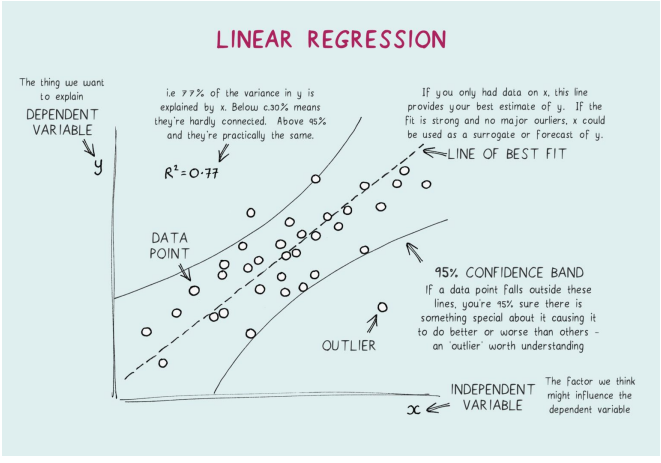

Figure: Overview - Regressão Linear

Métricas de Avaliação em Regressão Linear - Coeficiente de Determinação

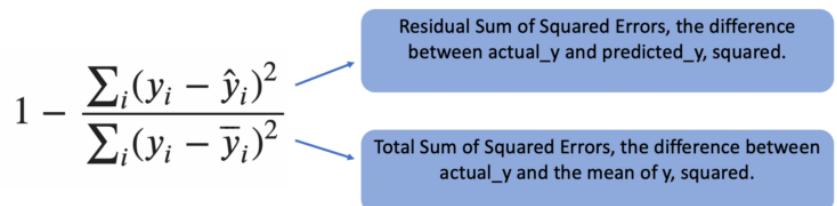

Figure: Fórmula - Coeficiente de Determinação

Métricas de Avaliação em Regressão Linear - Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Figure: Mean Squared Error

Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\Sigma_{i=1}^{N}\left|y_i - \hat{y}_i\right|$$

Figure: Mean Absolute Error

Questionamentos...

- Qual das 3 métricas é menos sensível a Outliers? E qual é a mais sensível?
- Qual a principal diferença entre as 3 métricas? (No sentido de "o que cada uma diz a respeito da performance do seu modelo)
- Cite contextos em que você acredita que uma das métricas diz mais que as outras quanto a analíse da performance

Overview - Modelos de Regressão

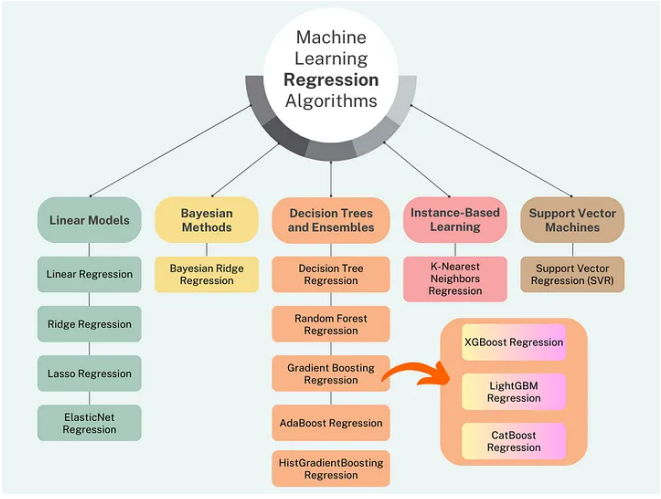

Figure: Diferentes tipos de modelos de Regressão

O que é um algoritmo de classificação?

Classificação é um método de aprendizado de máquina supervisionado em que o modelo tenta prever o rótulo correto de um determinado dado de entrada. Na classificação, o modelo é totalmente treinado usando os dados de treino e, em seguida, é avaliado com os dados de teste antes de ser utilizado para realizar previsões em novos dados ainda não vistos.

O clássico caso do classificador de e-mails

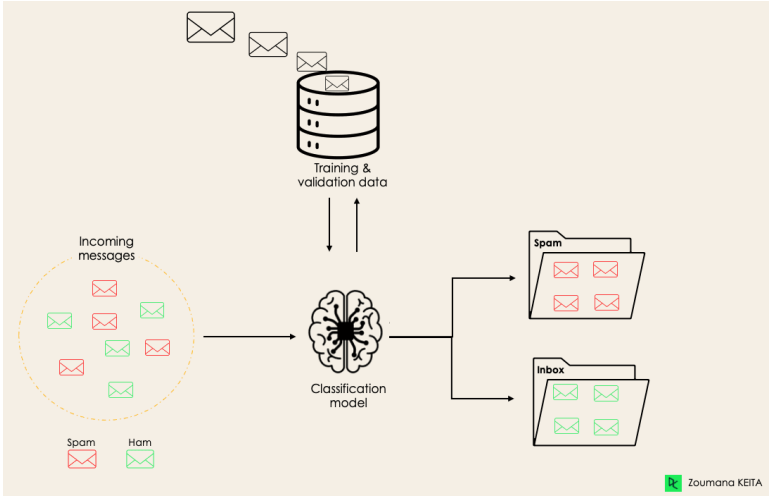

Figure: Classificador de e-mails

Lazy Learners Vs. Eager Learners

- Eager Learner: São algoritmos de aprendizado de máquina que primeiro constroem um modelo a partir do conjunto de dados de treinamento antes de fazer qualquer previsão em dados futuros. Eles gastam mais tempo durante o processo de treinamento devido à sua "ansiedade" em obter uma melhor generalização ao aprender os pesos, mas exigem menos tempo para fazer previsões.
- Lazy Learner: Não criam nenhum modelo imediatamente a partir dos dados de treinamento — e é daí que vem o aspecto "preguiçoso". Eles apenas memorizam os dados de treinamento, e, toda vez que há a necessidade de fazer uma previsão, procuram o vizinho mais próximo em todo o conjunto de treinamento, o que os torna muito lentos durante a fase de previsão.

Lazy Learners Vs. Eager Learners

Característica	Lazy Learners	Eager Learners
Abordagem de Treinamento	Memoriza todos os dados de treinamento durante o treinamento.	Cria uma representação generalizada durante o treinamento.
Processo de Predição	Procura por instâncias semelhantes durante a predição e aplica seus rótulos.	Aplica diretamente a representação aprendida para realizar a predição.
Adaptabilidade a Novos Dados	Adapta-se rapidamente a novos dados sem necessidade de reentreinamento.	Menos adaptável a novos dados; pode ser necessário reentreinamento.
Velocidade de Predição	Pode ser mais lenta, especialmente com grandes conjuntos de dados.	Predições mais rápidas devido ao modelo pré-treinado.
Predições Offline	Requer acesso aos dados de treinamento durante a predição.	Pode fazer predições offline ou sem os dados de treinamento.
Tratamento de Relações Complexas	Eficaz no tratamento de relações complexas e não lineares.	Mais adequado para padrões e relações bem definidos.
Representação do Modelo	Sem representação de modelo fixa; depende dos dados memorizados.	Requer uma representação de modelo fixa aprendida durante o treinamento.

Table: Comparação entre Lazy Learners e Eager Learners

Métricas de avaliação em algoritmos de classificação

- Acurácia: Útil em casos que sua varíavel alvo está balanceada; expressa, em linhas gerais, o quanto o seu modelo acertou em comparação com todas as previsões feitas.
- Precisão: De todos os dados classificados como positivos, quantos são realmente positivos.
- Recall: Qual a porcentagem de dados classificados como positivos comparado com a quantidade real de positivos que existem em nossa amostra.
- F1-score: essa métrica une precisão e recall afim de trazer um número único que determine a qualidade geral do nosso modelo.

Métricas de avaliação em algoritmos de classificação

Métrica	Descrição	Fórmula
Acurácia	Indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente.	$\frac{VP + VN}{VP + FV + FP + FN}$
Precisão	Dentre todas as classificações da classe positiva que o modelo fez, quantas estão corretas.	$\frac{VP}{VP + FP}$
Recall /Revocação/ Sensibilidade:	Dentre todas as classificações da classe positiva como valor esperado, quantas estão corretas.	$\frac{VP}{VP + FN}$
F1-Score	Média harmônica entre precisão e *Recall*.	$\frac{2 * Precisão * Recall}{Precisão + Recall}$

Figure: Métricas de avaliação em algoritmos de classificação