Resumo – Aprendizado Supervisionado

Machine Learning: permite que sistemas aprendam e melhorem com a experiência, sem serem explicitamente programados.

Dados: Informações coletadas que servem como base para o aprendizado. Podem ser de todo tipo: tabelas, imagens, sons, vídeos, etc.

Algoritmos: Conjuntos de instruções que processam os dados para identificar padrões. São os algoritmos que produzem os modelos de machine learning.

Em outras palavras, o modelo de machine learning é um produto da execução do lgoritmo sobre uma base de dados.

Métricas de Avaliação: Critérios utilizados para medir a performance e precisão dos modelos. São importantes para validar a qualidade do modelo, como sua performance e seu valor de negócio.

O que é Aprendizado Supervisionado?

Modelo treinado para **determinar saídas com base nos dados de entrada.**

Objetivo: aprender função que mapeia entrada para saída, permitindo previsões precisas.

Utiliza rótulos (dados anotados) durante o treinamento.

Foco na capacidade de generalização para novos dados.

Por que o Aprendizado é Chamado de "Supervisionado"?

Presença de Dados Rotulados: O modelo é treinado com conjuntos de dados que incluem entradas e suas respectivas saídas corretas, permitindo que aprenda a mapear entradas para saídas desejadas.

Processo de Correção: Durante o treinamento, o modelo faz previsões e as compara com as saídas reais, ajustando-se com base nos erros para melhorar sua precisão.

Analogia com Ensino Tradicional: Assim como um aluno aprende com a orientação de um professor, o modelo aprende com os dados rotulados que atuam como supervisores.

Objetivo de Generalização: Após o treinamento, o modelo deve ser capaz de aplicar o conhecimento adquirido para prever corretamente saídas de novos dados não vistos anteriormente.

O que é Aprendizado Supervisionado?

Objetivo: atribuir categoria/classe (classificação) ou valor (regressão) a um conjunto de dados com base em suas características.

Classificação: Prever categorias ou classes.

Regressão: Prever valores numéricos.

Features vs Target

Features

- Também chamadas de variáveis preditoras ou variáveis independentes.
- Dados de entrada do modelo.
- Categóricas e Numéricas.

Target:

- Também chamadas de variável de saída, variável alvo, variável resposta, e outros.
- Dados de saída do modelo.
- y: variável de saída alvo (target).
- y_hat: variável de saída predita (prediction)

PROCESSO

1. Preparação das Features

2. Treinamento do Modelo

3. Validação

4. Utilização (predição)

Regressão Linear

O modelo é a linha que melhor se 'ajusta' aos pontos.

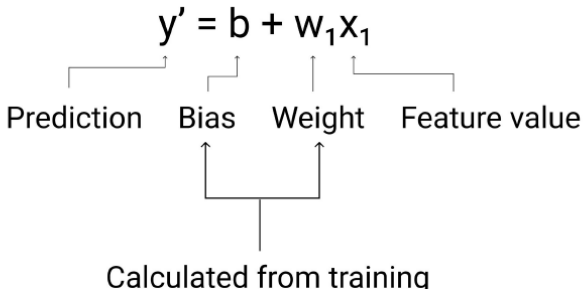
Equação da Regressão Linear

Definir a equação é encontrar os valores dos parâmetros m e b:

$y = m.x + b \rightarrow \quad y = (-3.6).x + 30$

Em Machine Learning, encontrar o valor dos parâmetros que melhor se ajusta aos dados, que melhor resolvem o problema, é chamado **treino**.

Treinar o Modelo = Encontrar o valor dos parâmetros



Treinar o modelo, é encontrar todos os valores de w (parâmetros), que encontre a curva que melhor se ajusta ao conjunto de dados.

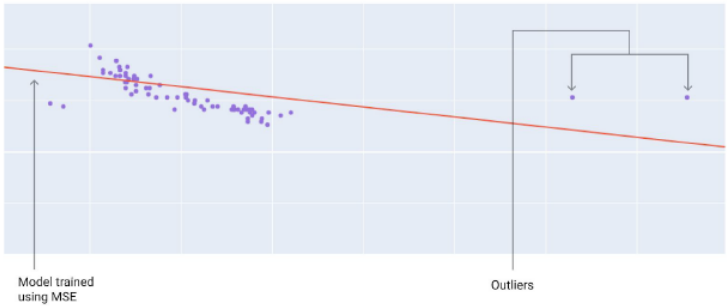
Função de custo

- Existem diferentes **loss function**, que são diferentes maneiras de medir o erro:

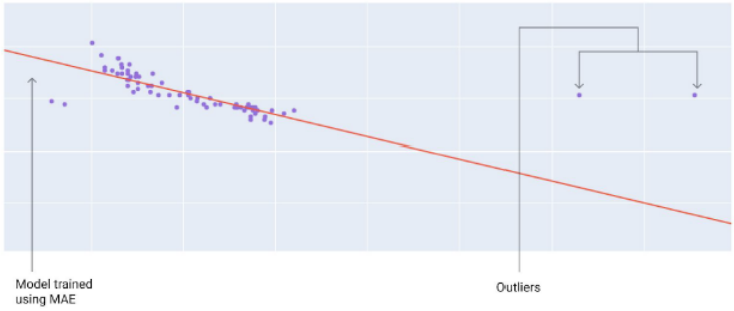
Tipo de Loss	Propósito principal	Equação
Perda L_1	Tornar o modelo **mais robusto a outliers:** cada erro contribui linearmente, então valores extremos não distorcem tanto o ajuste. Boa quando você quer **minimizar o desvio absoluto** total.	$\sum \lvert \text{valor real} - \text{valor previsto} \rvert$
Erro absoluto médio (MAE)	Fornecer uma **métrica fácil de interpretar,** na mesma unidade da variável-alvo, e ainda manter robustez razoável a outliers. Útil para avaliar desempenho quando você quer saber o erro médio típico.	$\frac{1}{N}\sum \lvert \text{valor real} - \text{valor previsto} \rvert$
Perda L_2	Priorizar a **redução de grandes erros**: como o erro é elevado ao quadrado, desvios maiores recebem penalidade bem maior. **É mais sensivel a outliers.**	$\sum (\text{valor real} - \text{valor previsto})^2$
Erro quadrático médio (MSE)	Servir como **função-objetivo padrão** em regressão e métrica para comparação entre modelos. Mantém as propriedades suaves da L_2, mas normaliza pela quantidade de exemplos, facilitando a leitura do valor e a convergência em otimização por gradiente.	$\frac{1}{N}\sum (\text{valor real} - \text{valor previsto})^2$

Função de custo

- Um modelo treinado com MSE aproxima o modelo dos outliers.



- Um modelo treinado com MAE fica mais distante dos outliers.



Função de Custo vs Métrica de Avaliação

Função de custo = função de perda = função objetivo = loss function

Função de custo ≠ Métrica de avaliação

- Função de perda (loss)
  - Usada durante o treinamento para minimizar o erro
  - Foco em eficiência computacional
  - Exemplo: MSE (Mean Squared Error)
- Métrica de desempenho (metric)
  - Usada para avaliar a qualidade do modelo nos dados de teste
  - Foco em interpretação prática
  - Exemplos:
    - R^2: proporção da variância explicada
    - MAE: erro médio absoluto em unidades reais
    - RMSE: raiz do erro quadrático médio

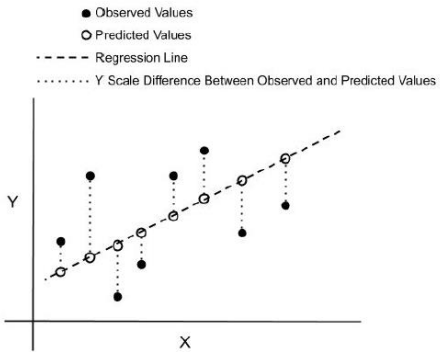
Principais Métricas para Regressão

Erro Quadrático Médio (MSE - Mean Squared Error)

- Calcula a média dos quadrados das diferenças entre as previsões e os valores reais.
- Utilidade: Útil para penalizar erros maiores, destacando discrepâncias entre previsões e valores reais.
- Faixa: De 0 até o infinito (quanto maior, pior).

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values



Principais Métricas para Regressão

Raiz do Erro Quadrático Médio (RMSE - Root Mean Squared Error)

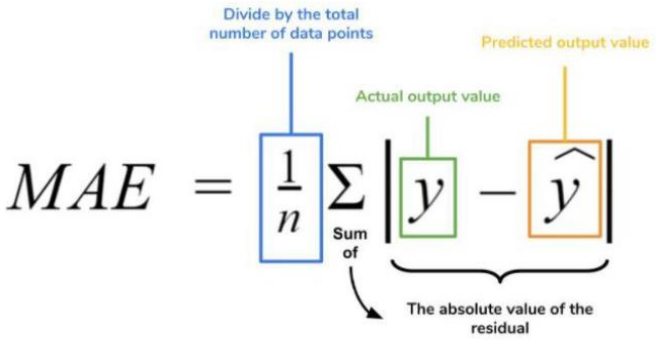
- É a raiz quadrada do MSE e fornece uma medida do erro **em unidades originais**.
- Utilidade: Oferece uma métrica de erro mais facilmente interpretável em comparação com o MSE.
- Faixa: De 0 até o infinito (quanto maior, pior).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Principais Métricas para Regressão

Erro Médio Absoluto (MAE - Mean Absolute Error)

- Calcula a média das diferenças absolutas entre as previsões e os valores reais.
- Utilidade: Menos sensível a valores discrepantes do que o MSE, sendo útil quando se deseja evitar que valores extremos distorçam a métrica.
- Faixa: De 0 até o infinito (quanto maior, pior).

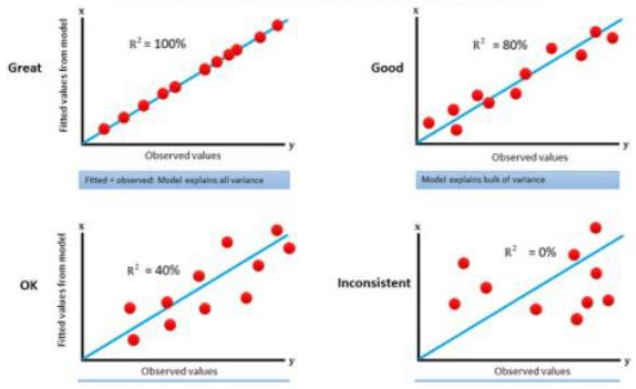


MAE vs RMSE

- **MAE (Erro Médio Absoluto):**
  - Use o MAE quando você deseja ter uma ideia clara da magnitude média dos erros.
  - O MAE é menos sensível a valores discrepantes, pois considera apenas as diferenças absolutas.
  - É uma boa escolha quando você quer uma métrica simples e fácil de interpretar, especialmente se houver outliers nos dados.
- **RMSE (Raiz do Erro Quadrático Médio):**
  - Prefira o RMSE quando erros grandes devem ser penalizados mais fortemente.
  - O RMSE amplifica o efeito de grandes erros devido ao processo de elevar ao quadrado as diferenças.
  - É útil em situações em que você quer dar mais peso a erros maiores, como quando a precisão é crítica e erros significativos são particularmente indesejáveis.

Coeficiente de Determinação (R^2 - R-squared)

- Avalia a proporção da variabilidade nos dados que é explicada pelo modelo.
- Utilidade: Fornece uma medida da qualidade global do modelo, quanto mais próximo de 1, melhor o ajuste.
- Faixa: De 0 a 1 (quanto maior, melhor).



Erro Percentual Absoluto Médio (MAPE - Mean Absolute Percentage Error)

- Calcula a média das porcentagens das diferenças absolutas entre as previsões e os valores reais.
- Utilidade: Útil quando a precisão relativa é mais importante do que a precisão absoluta.
- **Permite comparar séries com valores absolutos diferentes.**
- Faixa: De 0% até o infinito (quanto maior, pior).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i}$$

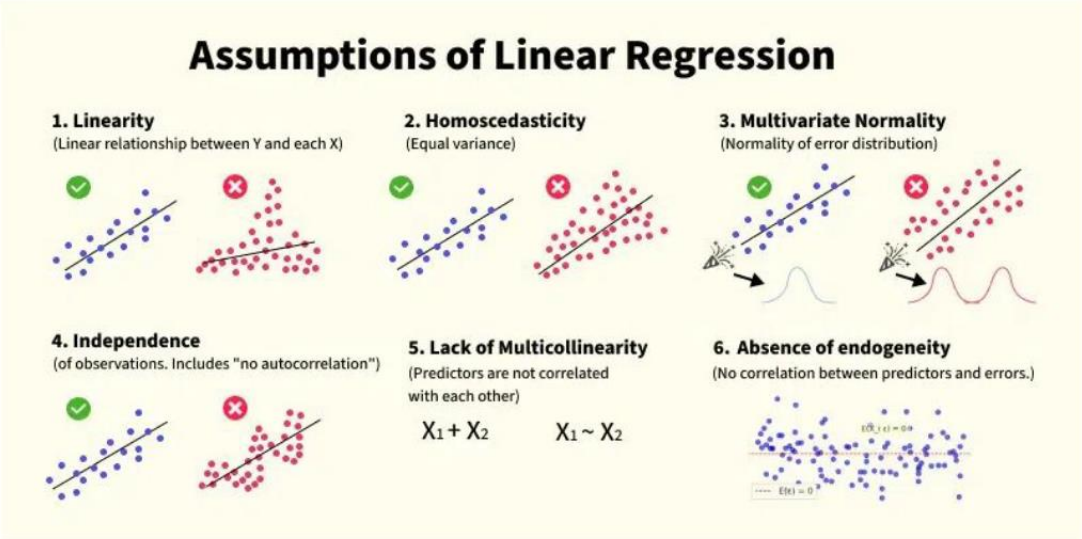
A_i is the actual value

F_i is the forecast value

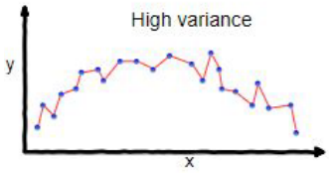
n is total number of observations

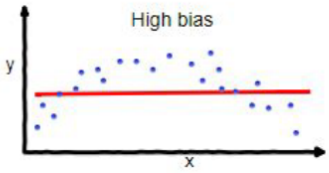
Suposições da Regressão Linear

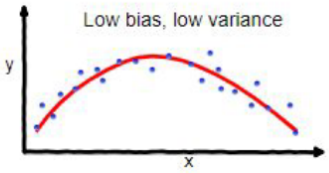
Em Resumo



Underfitting, Overfitting, Just right







overfitting

Também chamado de "high variance," ocorre quando um modelo é excessivamente complexo e se ajusta demais aos dados de treinamento, incluindo o ruído.

Um modelo overfit se adapta perfeitamente aos dados de treinamento, mas geralmente tem um desempenho ruim em dados de teste.

Pode ser mitigado com técnicas como regularização e aumento de dados.

underfitting

Também conhecido como "high bias," ocorre quando um modelo é muito simples para capturar os padrões nos dados.

Um modelo underfit não se ajusta bem aos dados de treinamento e tende a ter baixo desempenho em dados de teste.

Pode ser causado pela escolha de um modelo muito simples ou falta de dados de treinamento.

Good balance

Ocorre quando o modelo captura com precisão as relações nos dados, sem under ou overfitting.

Resulta em um desempenho equilibrado e confiável nos dados de treinamento e teste.

Encontrar um bom ajuste geralmente envolve ajustar a complexidade do modelo e usar técnicas como validação cruzada e regularização para equilibrar bias e variância.

Underfitting na Regressão Linear

- Acontece quando o modelo é **simples demais** para capturar o padrão dos dados.
- Na regressão linear, isso pode ocorrer quando:
  - Você usa **poucas variáveis** (ex: usa apenas x1, mas a variável alvo depende de x1, x2 e x3).
  - A relação real entre as variáveis é **não linear**, e um modelo linear não consegue representar a curva.
- **Sintomas**:
  - **Erro alto no treino** e **erro alto na validação**.
  - O modelo tem desempenho ruim mesmo nos dados em que foi treinado.
- **Solução:**
  - Usar mais variáveis que podem ajudar a entender melhor a variável de saída.

Overfitting na Regressão Linear

- Acontece quando o modelo é **complexo demais** e começa a ajustar o **ruído** dos dados em vez do padrão real.
- Na regressão linear, isso pode ocorrer quando:
    - Você inclui **muitas variáveis**, inclusive irrelevantes.
    - Você **não usa regularização** (como Ridge ou Lasso) em dados com muitas variáveis.
- **Sintomas**:
    - **Erro baixo no treino**, mas **erro alto na validação**.
    - O modelo vai muito bem nos dados de treino, mas generaliza mal para novos dados.
- **Solução:**
    - Usar métodos de regularização L1 (Lasso) ou L2 (Ridge).

Importar pacotes

Importar dados

Identificar variável y

Exploração de dados (distribuição, analise descritiva, outlier, correlação)

Preparar dados para modelagem

Treino e teste (80/20)

Regressão linear simples

Análise das métricas

Visualização gráfica

Validações

Regressão múltipla