Regressões explicadas com Python

A proposta é elaborar um material que integre **comparativos de regressões** à uma série de scripts que inclua **erros comuns, causas e generalizações.**

# Introdução

A linguagem Python será utilizada para dar suporte aos módulos dessa série, que terão os seguintes níveis:

|  |  |  |
| --- | --- | --- |
| **Nível** | **Nome** | **Objetivo** |
| 0 | **Inocente** | Intuição visual e numérica sem código. |
| 1 | **Calculadora com Script** | Implementação manual dos cálculos (sem sklearn). |
| 2 | **Ferramental Pronto** | Aplicação com bibliotecas (sklearn, statsmodels). |
| 3 | **Análise Crítica** | Diagnóstico de erros, limites do modelo, testes estatísticos. |
| 4 | **Generalização** | Quando aplicar o modelo e como compará-lo com outros. |
| 5 | **Projeto Real** | Aplicação no mundo real com dataset aberto (ex.: Spotify). |

Cada tipo de regressão será um **capítulo autônomo**, sendo que no primeiro nível (‘Inocente’) buscará explicações simples utilizando gráficos e exemplos com pares ordenados. No próximo nível, serão implementados scripts com numpy e realizado simulações de erro quadrático médio (MSE). No terceiro nível, o objetivo do capítulo é a implementação de scripts com o uso de sklearn.linear\_model, PolynomialFeatures, LogisticRegression, etc. Além da geração de gráficos com matplotlib.

No quarto nível, elencará os erros comuns e diagnóstico deles. Por exemplo, quando possuí um MSE alto mesmo com R² bom resultando em overfitting causado pela seleção errada do grau do polinômio. Esse diagnóstico pode ser obtido por meio de gráfico dos resíduos e curva de aprendizado. No próximo nível, analisará as generalizações e como usá-las, seus limites e as comparações possíveis com outros tipos de regressões.

O último nível dedicara-se as aplicações reais para isso o kaggle será uma fonte valiosa. O pré-processamento e pipeline dessas regressões conterão explicações dos seus resultados para leigos e técnicos. Assim, esses níveis abarcarão as seguintes regressões:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Regressão** | **Forma Matemática** | **Uso Ideal** | **Relação com Transformações** | **Exemplo prático** |
| Linear Simples |  | Tendência linear clara | Nenhuma | Altura vs. peso |
| Linear Múltipla |  | Vários fatores | Requer normalização | Preço de aluguel |
| Polinomial |  | Curvas suaves, não lineares | Criação de features | Idade vs. renda |
| Logarítmica |  | Crescimento rápido, saturação | Transformação do x | Popularidade no tempo |
| Exponencial |  | Crescimento exponencial | Transformação de y | População, juros |
| Logística |  | Classificação binária | Escala de probabilidades | Doença (sim/não) |

Esse material foi organizado no github da seguinte forma:

* /scripts/ → Códigos organizados por tipo e nível.
* /datasets/ → Pequenos exemplos + datasets reais.
* /docs/ → Notas explicativas, fórmulas, PDFs.
* /visuais/ → Gráficos, esquemas, gabaritos de interpretação.
* /.venv → Ambiente virtual.
* /README.md → Estrutura da série com links internos.

Esse material foi elaborado como uma metodologia progressiva, isto é, cada nível constrói sobre o anterior. Buscando aplicar os conhecimentos à datasets reais desde o início para que isso consiga estimular os diagnósticos e análises transversais. Assim, incentiva-se a investigação das falhas e suas causas.

Nessa mesma direção elaborou-se uma seção ‘Erros iguais, causas diferentes’, que terá o seguinte conteúdo:

|  |  |  |  |
| --- | --- | --- | --- |
| **Erro** | **Causa no Inocente** | **Causa no Data Scientist** | **Solução Didática** |
| Previsões ruins | Dados mal anotados | Overfitting | Mostrar com gráfico de curva de validação |
| Gráfico sem sentido | Dados mal plotados | Escala errada nos eixos | Mostrar importância da normalização |
| Modelo não converge | Fórmula errada | Dados colineares | Diagnóstico com matriz de correlação |

Ao final da leitura desse material espera-se ter contribuído com os leitores na construção de habilidades que lhes permitam construírem os seus próprios comparativos de regressões e os erros comuns nas aplicações dessas generalizações.