Projeto Final

Visualização de Dados

FGV EMAp

Joel Perca
Mariana Fernandes Rocha
Paula Eduarda de Lima

Rio de Janeiro, 2025

Contents

1 Introdução

Este projeto tem como objetivo principal ampliar a experiência prática no desenvolvimento, implementação e avaliação de métodos de visualização, abordando um problema de visualização concreto e propondo uma solução inovadora e criativa. O desafio reside em transformar modelos abstratos e conjuntos de dados multifacetados em insights compreensíveis e interativos. Para este propósito, utilizamos o conjunto de dados do Censo Habitacional da Califórnia de 1990. O dataset contém informações sobre casas em distritos específicos da Califórnia, incluindo estatísticas resumidas baseadas nos dados do censo de 1990, com variáveis como longitude, latitude, idade média das casas, número total de cômodos, população, renda média, valor médio das casas e proximidade com o oceano. Nossa visualização busca desmistificar conceitos fundamentais do aprendizado de máquina, com foco específico na explicação de árvores de decisão binária.

2 Dados

O conjunto de dados original contém informações sobre casas em distritos específicos da Califórnia, com estatísticas resumidas baseadas nos dados do censo de 1990. Para o propósito específico de nossa visualização, que foca na classificação de casas entre Sacramento e São Francisco, realizamos uma seleção e transformação das variáveis. O objetivo era concentrar-nos nas características intrínsecas da casa e na sua localização para a tarefa de classificação binária. Consequentemente, foram deletadas variáveis que não contribuíam diretamente para a descrição das características da propriedade ou que eram redundantes para a classificação geográfica após a introdução da variável city.

As variáveis utilizadas em nossa visualização são:

- **total_rooms**: Representa o número total de cômodos dentro de um quarteirão. Esta variável contribui para entender o tamanho ou a capacidade de uma propriedade.
- **total_bedrooms**: Indica o número total de quartos dentro de um quarteirão. Complementa total_rooms ao fornecer uma medida mais específica sobre a composição dos imóveis.
- **households**: Corresponde ao número total de famílias que residem em unidades habitacionais dentro de um quarteirão. Oferece contexto sobre a densidade populacional e o uso das propriedades.
- **median_house_value**: É o valor médio das casas dentro de um quarteirão, medido em dólares americanos. Embora não seja o alvo da nossa classificação, esta variável é um indicador socioeconômico relevante que pode estar correlacionado com a localização.

- **city**: Esta é uma variável-alvo criada para o nosso problema de classificação. Ela indica se uma casa está localizada em Sacramento ou São Francisco.

Essa seleção permitiu que a visualização se concentrasse nas relações entre as características das casas e a sua classificação geográfica.

3 Metodologia e construção da visualização

A metodologia central adotada para a construção da nossa visualização é a abordagem "ScrollStory". Este formato permite que a visualização gráfica permaneça fixa no centro da tela, enquanto o usuário rola através do conteúdo textual explicativo.

A estrutura técnica da nossa aplicação foi desenvolvida utilizando SvelteKit, um framework web moderno que facilita a criação de interfaces de usuário reativas e eficientes. Cada ScrollStory organiza seu layout em duas colunas: uma para o texto dos passos (por exemplo, Step1.svelte e Step2.svelte) e outra para o componente da visualização (Viz1.svelte e Viz2.svelte), fixado verticalmente com position: sticky.

A construção da visualização focou em ilustrar os fundamentos de como uma árvore de decisão classifica os dados, empregando as variáveis selecionadas do conjunto de dados do Censo Habitacional da Califórnia de 1990: total_rooms, total_bedrooms, households, media_house_value e a variável alvo city (Sacramento ou São Francisco).

Em resumo, a construção da visualização utilizou uma abordagem de storytelling interativo baseada em scroll e ferramentas de desenvolvimento web modernas para traduzir os conceitos abstratos do treinamento de árvores de decisão em uma experiência de aprendizado clara e visualmente intuitiva.

4 Conclusão

A visualização interativa desenvolvida serve como uma ferramenta didática inovadora para desmistificar o algoritmo de Árvores de Decisão, focando na tarefa de classificação binária de casas no Censo Habitacional da Califórnia de 1990. Esta solução aborda um problema de visualização, transformando conceitos complexos em uma experiência de aprendizado acessível e envolvente.

Os conceitos de aprendizado de máquina visualizados incluem:

- Classificação: O problema central da visualização é categorizar (classificar) casas como pertencentes a Sacramento ou São Francisco, uma tarefa de classificação binária.
- Características (Features): As variáveis do conjunto de dados (total_rooms, total_bedrooms, households, median_house_value) são apresentadas como "dimensões", "features" ou "preditores", que o modelo usa para tomar decisões.

- Divisões e Limites: A visualização demonstra como uma árvore de decisão identifica "limites" nos dados, que são os "pontos de divisão".
- Ramificações: As decisões da árvore são representadas como "instruções se-então" (if-then statements) ou "ramificações", que dividem os dados em dois "ramos" com base em um valor.
- Pureza e Ganho de Informação: A visualização tem como objetivo mostrar como as decisões são tomadas para tornar os ramos resultantes o mais "homogêneos" ou "puros" possível.
- Overfitting: A visualização, assim como as referências, busca contextualizar o problema do overfitting, onde o modelo aprende detalhes irrelevantes dos dados de treinamento, resultando em desempenho menos ideal em dados não vistos.

Em suma, a visualização capacita o usuário a compreender de forma intuitiva como uma árvore de decisão classifica dados, visualizando o fluxo dos dados através da estrutura da árvore à medida que as previsões são feitas.

Trabalhos Futuros

Para aprimorar ainda mais esta ferramenta educacional e abordar as limitações das árvores de decisão simples, várias direções para trabalhos futuros podem ser exploradas:

- Controle de poda interativo para evitar overfitting, o usuário escolhe os parâmetros de profundidade e pureza para realizar a poda.
- Extensão para Random Forests.
- Adaptação para regressão.
- Entrada de dados customizados pelo usuário, a possibilidade do usuário poder enviar seu dataset e o interface ser capaz de constriur a árvore de classificação e realizar a poda.

5 Referências

References

[1] R2D3. *A Visual Introduction to Machine Learning.* Disponível em: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/.

[2] MLU. *Decision Trees – Explained Visually.*
Disponível em: https://mlu-explain.github.io/decision-tree/.

[3] MLU. *California Housing Prices*.

Disponível em: https://www.kaggle.com/datasets/camnugent/california-housing-prices/data. Acesso em: jun. 2025.