Explorando Técnicas de Encoding: uma competência essencial para Cientistas de Dados

No mundo da ciência de dados, lidar com variáveis categóricas é uma tarefa comum e crucial. As variáveis categóricas, ao contrário das variáveis numéricas, não podem ser diretamente utilizadas em modelos de aprendizado de máquina e exigem um processo de codificação (encoding) para serem transformadas em um formato numérico que os algoritmos possam interpretar.

Existem várias técnicas de encoding que um cientista de dados deve conhecer para selecionar a mais adequada dependendo do contexto do seu problema. Vamos explorar sete técnicas essenciais de encoding de variáveis categóricas:

Técnica	Descrição	Melhor Aplicação	Evitar Aplicar	Exemplo
One-Hot Encoding	Cada categoria é representada por um vetor binário de 0s e 1s. Cada categoria obtém sua própria característica binária. (Codificação One-Hot)	Quando há poucas categorias em uma variável.	Em variáveis com muitas categorias, pois pode causar alta dimensionalidade.	`Color = [Red, Green, Blue]` `-> Red: [1, 0, 0], Green: [0, 1, 0], Blue: [0, 0, 1]`
Dummy Encoding	Similar ao One-Hot Encoding, mas elimina uma categoria para evitar multicolinearidade. (Codificação Dummy)	Similar ao One-Hot, mas com a vantagem de reduzir a dimensionalidade ligeiramente.	Pode ainda causar problemas de dimensionalidade se houver muitas categorias.	`Color = [Red, Green, Blue]` `-> Red: [1, 0], Green: [0, 1], Blue: [0, 0]`
Effect Encoding	Similar ao Dummy Encoding, mas altera a linha com todos zeros para -1. (Codificação por Efeito)	Quando se deseja capturar a presença e a ausência de categorias.	Em dados onde a presença de uma categoria é extremamente rara.	`Color = [Red, Green, Blue]` `-> Red: [1, 0], Green: [0, 1], Blue: [-1, -1]`
Label Encoding	Cada categoria recebe um rótulo numérico único. (Codificação por Rótulo)	Variáveis ordinais onde há uma ordem inerente entre as categorias.	Em variáveis nominais, pois pode introduzir uma ordenação inexistente.	`Color = [Red, Green, Blue]` `-> Red: 1, Green: 2, Blue: 3`
Ordinal Encoding	Similar ao Label Encoding, mas com ênfase na ordem dos valores. (Codificação Ordinal)	Em variáveis onde há uma relação de ordem entre as categorias.	Em variáveis categóricas não ordinais.	`Size = [Small, Medium, Large]` `-> Small: 1, Medium: 2, Large: 3`
Count Encoding	Baseia-se na frequência de cada categoria. (Codificação por Contagem)	Quando se deseja considerar a frequência das categorias.	Quando a distribuição das categorias não é significativa para o problema.	`City = [New York, Los Angeles, Chicago]` `-> New York: 30, Los Angeles: 20, Chicago: 15`
Binary Encoding	Combinação de One-Hot e Ordinal, converte valores ordinais em binário e depois em características binárias. (Codificação Binária)	Em variáveis com alta cardinalidade, reduzindo a dimensionalidade.	Pode ser complexo de interpretar.	`Color = [Red, Green, Blue]` `-> Red: [0, 1], Green: [1, 0], Blue: [1, 1]`

Esperamos que você tenha achado esta comparação útil. Seja você um iniciante no mundo da Ciência de Dados ou um profissional experiente, entender as nuances entre as técnicas de encoding pode abrir novas possibilidades em suas análises. Se você tiver alguma dúvida ou sugestão, não hesite em entrar em contato. Queremos garantir que nossos conteúdos sejam sempre relevantes e úteis para você.

Fique atento à nossa próxima edição, onde continuaremos a explorar ferramentas, técnicas e dicas valiosas para aprimorar suas habilidades em Ciência de Dados. Até lá, continue explorando e aprendendo!

Saudações,

Prof. Dr. Dilermando Piva Jr

Coordenador de Ciência de Dados para Negócios / Fatec Votorantim

E-mail: f301.cdn@fatec.sp.gov.br