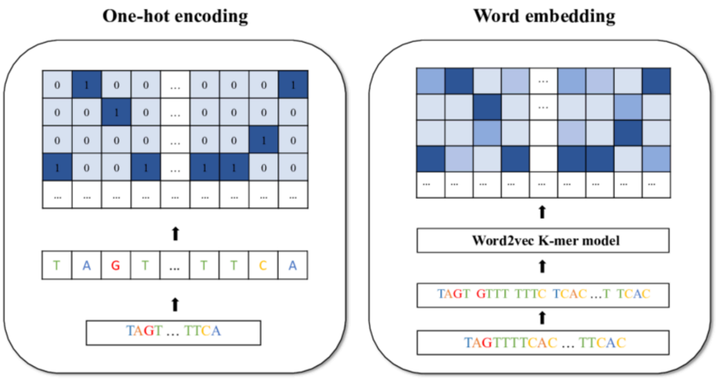



Difference between One-Hot Encoding vs Embedding

# What Is the Difference between One-Hot Encoding vs Embedding

**SURESH BEEKHANI**
Machine Learning Engineer | Data Scientist | AI Agents | GenAI

June 20, 2024

### Representing Categorical Data in Machine Learning Models

In machine learning, representing categorical data as numerical values is crucial. Two popular techniques for this are:

1. **One-hot encoding**
2. **Embedding**

## One-hot Encoding

One-hot encoding is a simple technique, suitable for small datasets with a limited number of categories.

## Embedding

On the other hand, Embedding is more suited for large datasets with high-cardinality categorical features. It is beneficial when capturing relationships between categories is important.

## Key Differences

## Dimensionality

- **One-hot encoding:** Increases the data dimensionality by creating a new binary column for each unique category. For N categories, it creates N new columns, leading to a high-dimensional, sparse representation. This can be inefficient and impractical for datasets with many categories.
- **Embedding:** Reduces the dimensionality by representing each category as a dense vector of lower dimensionality (e.g., 8, 16, 32 dimensions). This results in a compact representation that is more manageable and efficient.

### Relationship between Categories

- **One-hot encoding:** Treats each category as independent and orthogonal, meaning there is no inherent relationship between the categories. Each category is represented by a unique binary vector, where only one element is '1' and all others are '0'.

- **Embedding:** Captures semantic relationships and similarities between categories by placing similar categories closer together in the embedding space. This is achieved by learning dense vectors that represent the categories in a way that similar categories have similar vectors.

### Interpretability

- **One-hot encoding:** Vectors are easily interpretable since each column directly corresponds to a specific category. This makes it straightforward to understand and analyze the encoded data.
- **Embedding:** Vectors are dense and harder to interpret directly. The meaning of individual dimensions is not as clear as in one-hot encoding, as embeddings capture complex relationships learned from the data.

### Scalability

- **One-hot encoding:** Becomes inefficient and sparse when dealing with high-cardinality categorical features (many unique categories). This can lead to the curse of dimensionality, where the data space becomes so large that the learning algorithm struggles to generalize.
- **Embedding:** More scalable and efficient for high-cardinality features, as each category is represented in a fixed-size, lower-dimensional vector. This makes embeddings suitable for large datasets with many unique categories.

### Learning

- **One-hot encoding:** A simple deterministic process that does not require any learning. Each category is independently encoded without considering the data distribution or relationships between categories.
- **Embedding:** Learned from data, typically as part of the training process of a neural network. Embeddings are adjusted during training to capture the relationships between categories, making them data-driven and context-aware.

**SURESH BEEKHANI**

Machine Learning Engineer | Data Scientist | AI Agents | GenAI

+ Follow

## More articles for you


**Comparing Performances of Encoding Schemes o...**

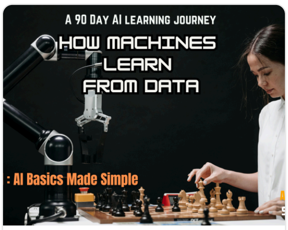
Muhammad Imran Khan

1


**Encoding in Machine Learning: Types, Usage, and Criteria fo...**

Daily Data Pill

11 · 2 comments




**Day 13 : How Machines Learn from Data – An Overview**

George Bonela

5


**Challenges and steps involved in solving Machine Learning -...**

Vinod Kumar GR

1