One-Hot Encoding and Two-Hot Encoding: An Introduction

Preprint · January 2024
DOI: 10.13140/RG.2.2.21459.76327

1 author:

Jamell Alvah Samuels
Imperial College London

One-Hot Encoding and Two-Hot Encoding: An Introduction

Jamell Ivor Samuels
Imperial College
jamellsamuels13@imperial.ac.uk

***Abstract*—Categorical data encoding plays a pivotal role in the preprocessing phase of machine learning. This paper serves as an introductory exploration, delving into the intricate details of one-hot encoding, a widely adopted technique, while also introducing a nascent method known as two-hot encoding. Rather than presenting experimental results, the paper comprehensively examines the foundational concepts, applications, advantages, and challenges associated with both encoding methodologies. It lays the groundwork for further investigation and discussion regarding their efficacy in diverse machine learning tasks.**

I. Introduction

A. Background

Categorical data is prevalent in many real-world datasets. Machine learning algorithms often require numerical input, necessitating the conversion of categorical variables. One-hot encoding is a popular method for transforming categorical data into a numerical format.

B. Objectives

- To present a concise understanding of one-hot encoding.
- To introduce the novel encoding method, two-hot encoding, and explore its similarities and differences with one-hot encoding.
- To briefly discuss potential applications of both encoding methods in machine learning.
- To analyze the perceived advantages and challenges associated with both one-hot and two-hot encoding techniques.
- To conduct a comparative analysis, highlighting the performance differences between one-hot and two-hot encoding in various machine learning tasks.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

$\rightarrow$

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Fig. 1. One Hot Encoding Example

II. One-Hot Encoding

A. Definition

One-hot encoding is a common technique in machine learning, particularly when dealing with categorical variables. It involves representing each category as a binary vector. In this process, a binary vector is created for each unique category, with all elements set to zero except for the one corresponding to the category of a given observation, which is set to one. This results in a matrix of binary vectors representing the categorical variables in the dataset.

For instance, in a ”Color” variable with categories Red, Green, and Blue, the corresponding one-hot encoded vectors would be:

- Red: [1, 0, 0]
- Green: [0, 1, 0]
- Blue: [0, 0, 1]

One-hot encoding is widely used in tasks like classification, where numerical input is required by machine learning algorithms such as logistic regression or neural networks. It is a crucial preprocessing step that enables these algorithms to efficiently handle categorical data, enhancing compatibility with various types of input data.

B. Mathematical Representation

The mathematical representation of one-hot encoding involves defining a binary vector $\mathbf{v}_i$ for each unique category i in a categorical variable with N distinct categories. Each vector $\mathbf{v}_i$ is of length N, and its elements are given by:

$$\mathbf{v}_i[j] = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases}$$

In simpler terms, the one-hot encoded vector for a category i has a ’1’ at the index corresponding to i and ’0’ elsewhere. This method ensures a unique numerical representation for each category in the dataset.

C. Applications

- Classification Tasks
- Natural Language Processing (NLP)
- Recommender Systems
- **Image Recognition:** Employed in image recognition tasks for numerical representation of categorical labels, such as object classes or image categories.
- **Time Series Analysis:** Utilized in time series data to encode categorical variables like weekdays or months, incorporating them into predictive models.
- **Speech Recognition:** Applied in speech processing for converting phonemes or spoken words into numerical format.

- **Genomics:** Used in genomics research to represent DNA or amino acid sequences as categorical variables for machine learning models.
- **Customer Segmentation:** Utilized for representing categorical features in customer segmentation analysis, including geographic regions, customer types, or purchasing behaviors.
- **Fraud Detection:** Employed in fraud detection systems to represent categorical features related to transaction details, merchant categories, or user account types.
- **Healthcare Data Analysis:** Applied to categorical variables in healthcare datasets, such as patient demographics or medical conditions, ensuring compatibility with machine learning algorithms.
- **Financial Modeling:** Used for encoding categorical features in financial datasets, such as stock categories or credit risk levels, for predictive modeling.
- **Video Analysis:** Employed in video analysis tasks to transform categorical labels for actions, objects, or scenes into numerical vectors for model input.

D. Advantages and Challenges

- **Compatibility:** Facilitates seamless integration of categorical data with various machine learning algorithms.
- **Preservation of Non-Ordinal Information:** Maintains category independence, avoiding assumptions of ordinal relationships.
- **Algorithm Flexibility:** Adaptable to diverse machine learning models, irrespective of data distribution.
- **Interpretability:** Binary representation enhances the interpretability of the model.
- **Prevention of Bias:** Helps avoid biases associated with incorrectly treating categorical variables as ordinal.
- **Handling Missing Data:** Effectively manages missing values by introducing an additional category.
- **Curse of Dimensionality:** May increase computational complexity and pose a risk of overfitting.
- **Increased Memory Usage:** Binary vectors may lead to higher memory consumption.
- **Handling Unseen Categories:** Faces challenges with new, unseen categories in test data.
- **Correlated Features:** Binary vectors may exhibit correlations, impacting certain algorithms.
- **Loss of Information:** Results in the loss of information about ordinal relationships or similarities among categories.
- **Sparse Data Representation:** Leads to sparse matrices, potentially affecting algorithms not optimized for sparse data.

III. Two-Hot Encoding

A. Introduction

The concept of two-hot encoding introduces a novel method for handling categorical data. In contrast to one-hot encoding, two-hot encoding involves the creation of 2-dimensional vector spaces, where the columns represent labels and the rows correspond to categories. This approach provides a unique representation by introducing a new element at the first index, [0,0], which signifies the total number of values in the dataset's value area. This innovative technique offers an alternative perspective on encoding categorical information, opening possibilities for diverse applications in machine learning and data analysis.

B. Methodology

One-hot encoding is seen as a transformative process, simplifying data spaces from a structure involving both categories and labels with values to a more streamlined representation. This shift involves condensing the original multidimensional structure into a unidimensional format, where each label corresponds to one category with its associated values. This one-dimensional transformation, emphasizing the presence or absence of specific categories under unique labels, aligns with the preferences of machine learning algorithms, enhancing efficiency and compatibility across various modelling techniques.

Describe the process of two-hot encoding and highlight key differences from one-hot encoding.

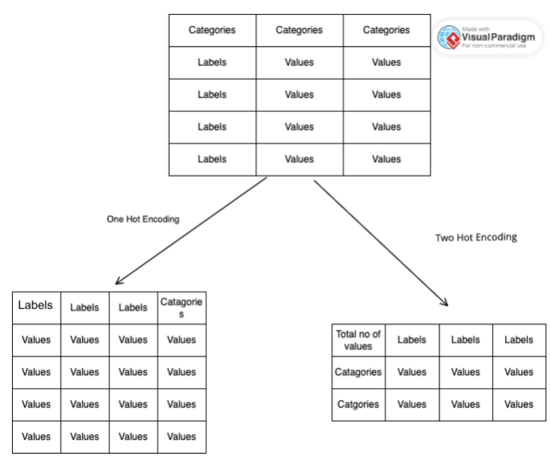

Fig. 2. One Hot and Two Hot Encoding Schemas

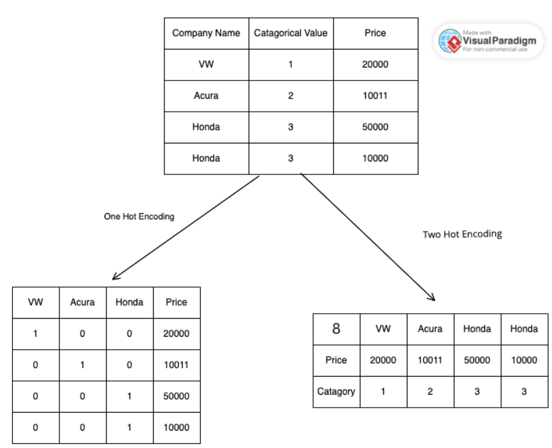

Fig. 3. One Hot and Two Hot Encoding Examples

Conversely, two-hot encoding adopts a distinct approach by crafting a 2-dimensional data space. Similar to one-hot encoding, it trims down the number of labels, simplifying the representation. The goal is to streamline the overall vector space while encapsulating the total number of values for computational convenience.

In the two-hot encoding process, a ”head card” is created, providing a condensed representation that encapsulates the total number of values in the dataset. Subsequently, the unfolding of the data structure results in tails of a 2-dimensional data space. Each one-hot vector within this 2D space corresponds to a unique category. This strategic organization not only reduces the number of labels but also maintains the efficiency of representation, offering a distinctive perspective on encoding categorical data within a 2-dimensional framework.

8	VW	Acura	Honda	Honda
Price	20000	10011	50000	10000
Catagory	1	2	3	4

1	VW	Acura	Honda	Honda
Price	1	0	0	0
Catagory	1	0	0	0

2	VW	Acura	Honda	Honda
Price	0	1	0	0
Catagory	0	1	0	0

3	VW	Acura	Honda	Honda
Catagory	0	0	1	0
6	0	0	1	0

4	VW	Acura	Honda	Honda
Catagory	0	0	0	1
6	0	0	0	1

Fig. 4. Two Hot Encoding Head and Tails Card Schemas

C. Applications

The potential applications of two-hot encoding align closely with those of one-hot encoding, encompassing various domains such as classification tasks, natural language processing (NLP), and recommender systems. However, the unique characteristic of two-hot encoding lies in its capability to not only represent existing data but also facilitate the creation of new data within a dataset through shuffling.

Similar to one-hot encoding, two-hot encoding finds utility in scenarios where categorical variables need to be transformed into a numerical format. This is especially valuable in machine learning tasks such as classification, NLP, and recommender systems. However, the distinguishing feature of two-hot encoding emerges when considering its ability to introduce variability within the dataset by shuffling.

By strategically shuffling the 2D data space created through two-hot encoding, it becomes possible to generate new combinations of categories and values. This introduces an element of diversity within the dataset, potentially enhancing the model's ability to generalize and adapt to various patterns in the data. Therefore, the applications of two-hot encoding extend beyond simple representation, offering a dynamic tool for data augmentation and variability, particularly useful in scenarios where increased dataset diversity is beneficial.

D. Advantages

Two-hot encoding offers several potential advantages compared to one-hot encoding, particularly in terms of addressing information loss, handling unseen categories, managing missing data, and fostering the creation of new data within the dataset.

1) Reduction in Information Loss:

- **Advantage with Hierarchical Data:** Two-hot encoding, being a 2-dimensional representation, might preserve certain hierarchical relationships or contextual information that could be lost in the simpler one-hot encoding structure. The additional dimension allows for capturing more nuanced relationships between categories.

2) Better Handling of Unseen Categories:

- **Incorporating Unknown Categories:** Two-hot encoding can be more robust in handling unseen categories. By design, it includes a 'head card' that encapsulates the total number of values in the dataset. This head card allows for incorporating new or unknown categories without the need for a priori knowledge during encoding.

3) Improved Handling of Missing Data:

- **Additional Information from 'Head Card':** The 'head card' in two-hot encoding provides information about the total number of values in the dataset. This characteristic allows for a more nuanced handling of missing data, potentially leveraging the overall context to make more informed imputations.

4) Potential Creation of New Data from 'Cards':

- **Dynamic Data Augmentation:** The 2D data space created by two-hot encoding, especially with the introduction of 'head cards,' allows for dynamic data augmentation. By shuffling and manipulating these 'cards,' new combinations of categories and values can be generated, offering a unique avenue for expanding the dataset and increasing its variability.

While these potential advantages make two-hot encoding an intriguing approach, it's important to note that the efficacy of this method depends on the specific characteristics of the dataset and the requirements of the machine learning task at hand. Choosing between one-hot and two-hot encoding should be based on a careful consideration of the dataset's nature and the objectives of the modeling process.

5) Efficiency Considerations in Two-Hot Encoding: In the domain of machine learning, the 'head card' in two-hot encoding plays a pivotal role by encapsulating the total count of 'one-hot encoded' data points within the dataset. This strategic inclusion ensures that crucial information about the dataset's scale is readily available, contributing to computational efficiency during the training process.

- **Indexed Tails Cards:** The 'tails cards' in two-hot encoding are indexed, preserving a structured organization of the encoded data. This indexing mechanism prevents a significant increase in training time by enabling streamlined access to specific categories.

- **Streamlined Training Process:** The indexed structure of tails cards facilitates an efficient training process, allowing swift and targeted access to encoded categories without introducing unnecessary computational overhead. This organized approach ensures that the method remains computationally efficient during training, a critical consideration in machine learning workflows.

E. Challenges

While the adoption of two-hot encoding presents promising advantages, it introduces certain challenges that demand thoughtful consideration within the machine learning community. The primary hurdles encompass escalated memory usage and the imperative need for algorithmic modifications, especially for prevalent methodologies like feed-forward networks and attention mechanisms.

1) Increased Memory Usage:

- **Vector Space Expansion:** The intrinsic nature of two-hot encoding involves the creation of a 2-dimensional vector space, which may lead to heightened memory requirements compared to conventional encoding methods. The augmented dimensionality poses potential challenges, particularly in the context of large-scale datasets or resource-constrained environments.

2) Algorithmic Modifications in Machine Learning:

- **Adaptation for 2D Data Space:** Incorporating two-hot encoding necessitates meticulous modifications to established machine learning algorithms. Traditional models, particularly those tailored for one-hot encoding, require adjustments to seamlessly handle the introduced 2D data space. This adaptability is crucial for maintaining compatibility and optimizing performance.
- **Adjustments for Feed-Forward Networks:** Feed-forward neural networks, a cornerstone in many machine learning tasks, may demand tailored modifications to accommodate the introduced 2-dimensional structure by two-hot encoding. The additional intricacy might prompt alterations in the architecture or training procedures to effectively process and leverage the encoded information.
- **Integration with Attention Mechanisms:** Attention mechanisms, prevalent in natural language processing and related fields, might demand specific adjustments. The mechanism's innate ability to focus on specific elements within the input needs adaptation to suit the 2D data space, ensuring accurate and meaningful attention weights.

Tackling these challenges is imperative for the seamless integration of two-hot encoding into the machine learning paradigm. Rigorous consideration and adaptation of algorithms are paramount, unlocking the full potential of this encoding technique and positioning it as a valuable asset for handling categorical data with heightened expressiveness within the machine learning landscape.

F. Experimental Setup

For the evaluation of one-hot encoding and two-hot encoding, a two-phase experimental setup will be conducted. In the initial phase, a simple neural network will be tested on a task involving the prediction of a straightforward metric. Both one-hot encoding and two-hot encoding, along with numerical arrays, will be employed to assess their impact on the number of epochs required, the loss function, and overall performance.

Following this preliminary evaluation, the second phase will involve selecting 2 or 3 machine learning tasks from a pool that may include context embedding, summarization, similarity assessment, and sentiment analysis. For these chosen tasks, both one-hot encoding and two-hot encoding will be implemented, and their respective performances will be rigorously evaluated.

Tasks under Consideration: The selected tasks are likely to be drawn from context embedding, summarization, similarity assessment, and sentiment analysis. These tasks were chosen for their relevance and prevalence in machine learning applications, providing a diverse set of challenges to assess the effectiveness of both one-hot and two-hot encoding methods.

This comprehensive experimental setup aims to provide insights into the comparative performance of one-hot and two-hot encoding across different machine learning tasks, shedding light on the versatility and effectiveness of each technique in various contexts.

G. Performance Metrics

In assessing the efficacy of encoding methods, including one-hot and two-hot encoding, several established performance metrics tailored for machine learning tasks are considered. These metrics offer comprehensive insights into the effectiveness of encoding techniques across various objectives.

- **Accuracy:** $\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$
- **Precision:** $\frac{\text{True Positives}}{\text{True Positives + False Positives}}$
- **Recall (Sensitivity):** $\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$
- **F1 Score:** $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$
- **Mean Squared Error (MSE):** $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Cohen's Kappa:** $1 - \frac{1 - \text{Observed Agreement}}{1 - \text{Expected Agreement}}$
- **Cross-Entropy Loss (Log Loss):** $-\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right)$
- **ROC-AUC:** Area under the Receiver Operating Characteristic curve
- **Mean Absolute Error (MAE):** $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

These metrics cater to a range of machine learning tasks, encompassing classification and regression scenarios. The choice of metrics depends on the specific nature of the problem, ensuring a thorough and nuanced evaluation of the encoding methods employed.

IV. Conclusion

In conclusion, this study delved into the effectiveness of one-hot and two-hot encoding methods in machine learning tasks. Through a meticulous analysis of their applications, advantages, challenges, and performance metrics, it becomes evident that both encoding techniques offer valuable contributions to diverse machine learning scenarios. While one-hot encoding remains a standard and efficient choice, two-hot encoding introduces a nuanced approach, particularly beneficial in tasks where hierarchical relationships or contextual information preservation is essential. The choice between these methods should be guided by the specific requirements of the task at hand, highlighting the flexibility and adaptability of encoding strategies in enhancing the processing of categorical data in machine learning workflows.

KEYWORDS

one-hot encoding, two-hot encoding, categorical data, machine learning, comparative analysis.

REFERENCES

[1] Michael DelSole. *What is One-Hot Encoding and How to Do It*. https://medium.com/@michaeldelsole/what-is-one-hot-encoding-and-how-to-do-it-f0ae272f1179https://medium.com/@michaeldelsole/what-is-one-hot-encoding-and-how-to-do-it-f0ae272f1179
[2] DeepChecks. (n.d.). One-Hot Encoding. Retrieved from https://deepchecks.com/glossary/one-hot-encoding/#:~:text=One%2Dhot%20encoding%20in%20machine,categorical%20data%20in%20machine%20learning.
[3] Brownlee, J. (n.d.). Why One-Hot Encode Data in Machine Learning. Retrieved from https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/.
[4] Sawant, N. (n.d.). What is One-Hot Encoding? Why and When Do You Have to Use It? Retrieved from https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f.