

✦ Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)



One-Hot Encoding สร้างตัวแปร Dummies สำหรับ Classification model



Sasiwut Chaiyadecha · [Follow](#)

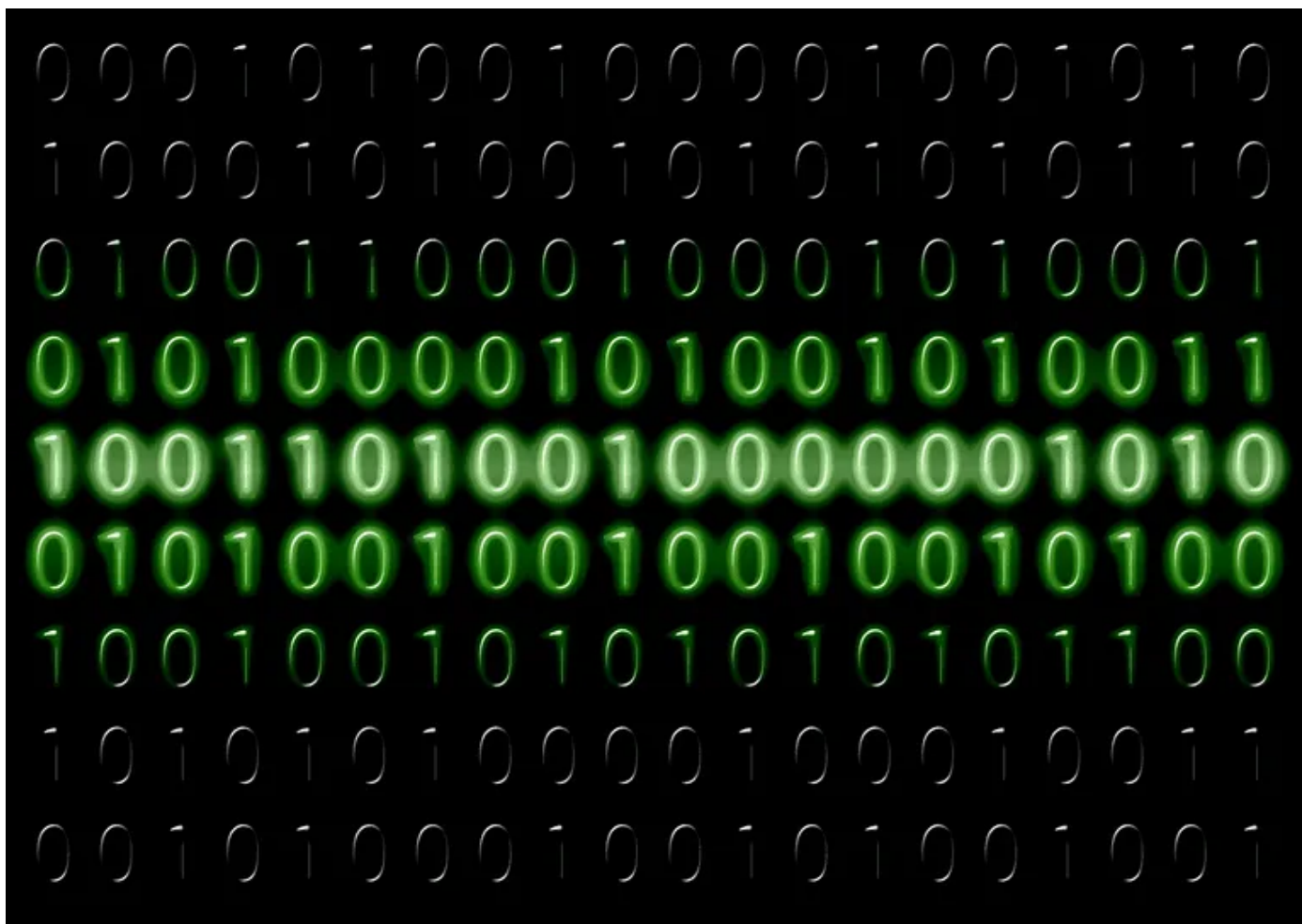
3 min read · Dec 23, 2020



10



One-Hot encoding หรือ One-Hot encoder แล้วแต่ว่าใครสะดวกเรียกในแบบไหน ซึ่งมันคือสิ่งเดียวกัน วันนี้มารู้จักกับการ Transform ข้อมูลในรูปแบบนี้ รวมไปถึงการใช้งานและเมื่อไหร่ที่ต้องใช้งาน One-Hot encoding



การทำงานของ Machine Learning หรืออาจคุ้นเคยกว่าในคำว่า Model ล้วนแล้ว แต่เป็นการทำงานบน “ตัวเลข” คงมีบางคนเกิดคำถามขึ้นมาว่า แล้วข้อมูลที่ใส่เข้าไปใน Neural network มันมีทั้งรูปภาพ วิดีโอ เสียง หรือแม้แต่ตัวหนังสือข้อความ... จริงอยู่ที่ลักษณะทางกายภาพของข้อมูลเหล่านั้นไม่ใช่ตัวเลข แต่เมื่อส่งผ่านข้อมูลเข้าไปในโมเดล ข้อมูลต้องถูกเปลี่ยนตัวเลข ไม่ว่าวิธีใดก็วิธีหนึ่ง ซึ่งเรียกว่าการทำงาน Encoding ซึ่ง One-Hot encoding เป็นการ Encoding ในอีกลักษณะหนึ่ง

What is One-Hot encoding?

One-Hot encoding คือการทำข้อมูลที่ถูกเก็บในลักษณะ Categorical ทั้งในลักษณะที่มีลำดับ (Ordinal number) และไม่มีลำดับ (Nominal number) เปลี่ยนให้อยู่ในรูปแบบของ Binary values ที่มีค่า 0 หรือ 1 เท่านั้น

	location	menu	price
0	[1, 1, 1, 0]	[1, 1, 0]	[1, 0]
1	[0, 1, 0, 0]	[0, 0, 1]	[1, 0]
2	[1, 1, 0, 1]	[1, 1, 1]	[0, 1]

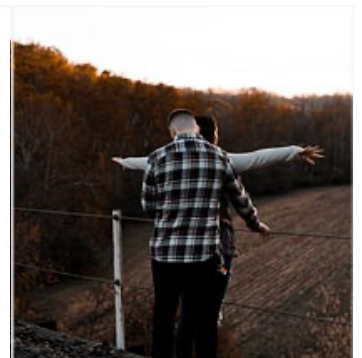
ตัวอย่างของการทำ one hot encoding

การจัดการกับ Categorical variable สามารถทำได้หลายวิธี แต่วันนี้ขอพูดถึงแค่วิธี One-Hot encoding เพียงอย่างเดียว จริง ๆ แล้วก่อนหน้านี้ เคยเขียนเกี่ยวกับการแปลงข้อมูลอีกประเภทหนึ่ง ในตอนที่ทำ Credit scoring model โดยในครั้งนั้นได้มีการแปลงข้อมูลเป็น WOE ก่อนทำ Logistic regression model สามารถย้อนกลับไปอ่านได้จาก Link ข้างล่าง

Titanic dataset #2: WOE และ IV ในการหาความสำคัญของตัวแปร

หาความสำคัญของตัวแปรแต่ละตัวด้วย WOE และ IV

medium.com



ซึ่งไม่ว่าจะเป็นการจัดการข้อมูลด้วยวิธีไหน แต่ใจความสำคัญของเทคนิคนี้คือ การเปลี่ยนข้อมูลให้ดูในรูปตัวเลขสำหรับงาน Machine Learning

Types of One-Hot encoding

One-Hot encoding แบ่งออกได้เป็น 2 ประเภทใหญ่ ๆ คือ

1. One-Hot encoding ซึ่งเป็นประเภทที่ได้อธิบายไปแล้วทั้งหมด
2. Dummy Variable Encoding ซึ่งมีลักษณะเหมือนกับ One-Hot encoding ทั้งหมด แต่แตกต่างกันที่จำนวน Column น้อยกว่า 1 Column

ขอยกตัวอย่างการทำ One-Hot encoding เพื่อความเข้าใจที่มากขึ้น สมมติว่ามีนักเรียนอยู่ทั้งหมด 4 คนคือ A B C และ D นักเรียนแต่ละคนสามารถรับลูกอมได้คนละ 1 เม็ด โดยที่มีลูกอมอยู่ทั้งหมด 3 สีคือ แดง เขียว และน้ำเงิน

Name	Candy
A	R
B	G
C	B
D	R

raw data ที่เก็บเป็น categorical

Name	R	G	B
A	1	0	0
B	0	1	0
C	0	0	1
D	1	0	0

data หลังจากเปลี่ยนเป็น one-hot encoding

ข้อดีของการทำ One-Hot encoding แน่นอนว่าการแปลงข้อมูลให้เครื่องคอมพิวเตอร์สามารถเรียนรู้และเข้าใจ Logic ที่เป็นตัวเลขได้ ซึ่งรวมไปถึงการคำนวณทางคณิตศาสตร์ต่าง ๆ ส่วนข้อเสียคือการใช้ทรัพยากรเครื่องที่เพิ่มขึ้นตามจำนวนข้อมูล เพราะการทำ One-Hot encoding ข้อมูลถูกแปลงให้อยู่ในรูป Sparse Matrix หรือ Matrix ที่ประกอบไปด้วยเลข 0 เป็นจำนวนมาก ดังนั้นหากข้อมูล Categorical มีหลายประเภท ย่อมทำให้ต้องการ Memory ในการทำงานเยอะขึ้น

pandas.get_dummies()

ใน Python มี Library เพื่ออำนวยความสะดวกในการทำ One-Hot encoding อยู่หลายตัว โดยตัวแรกที่ขอแนะนำคือ `pandas.get_dummies()` อย่างที่คุ้นเคยกันดีอยู่

แล้วว่า Pandas เน้นการทำงานบน DataFrame เป็นหลัก ดังนั้นการทำงานของคำสั่งนี้คือ การแปลง Categorical variable ให้อยู่ในรูปของตาราง 0, 1 ซึ่งสามารถนำไปใช้งานต่อในเชิงการคำนวณ หรือการพัฒนาโมเดลได้

Code ด้านบนเป็นตัวอย่างการแปลง One-Hot encoding ด้วย Pandas



Write



```
1 import pandas as pd
2
3 #One-Hot encoding
4 data = {'Students': ['A', 'B', 'C', 'D'], 'Candy': ['R', 'G', 'B', 'R']}
5 df = pd.DataFrame(data)
6 dummies = pd.get_dummies(df['Candy'])
7 df_onehot = pd.concat([df, dummies], axis = 1)
8 df_onehot
```

onehot1.py hosted with ♥ by GitHub

[view raw](#)

ซึ่งจาก Code ด้านบน ให้ผลลัพธ์เท่ากับตัวอย่างที่วาดรูปเอาไว้ โดยนักเรียนแต่ละคนมีลูกอมคนละสีที่แตกต่างกัน ยังมีวิธีการทำ One-Hot อีกวิธีที่เรียกว่า Dummy Variable Encoding โดยที่จำนวน Column ลดลงไป 1 Column

```
1 #Dummy Variable Encoding
2 dummies = pd.get_dummies(df['Candy'], drop_first = True)
3 df_var = pd.concat([df, dummies], axis = 1)
4 df_var
```

onehot2.py hosted with ♥ by GitHub

[view raw](#)

Code ด้านบนมีการใส่ Option เพิ่มเติมคือ `drop_first = True` ดังนั้นผลลัพธ์ที่ได้คือ Column One-Hot ที่มีขนาดลดลงไป 1 Column กล่าวคือจะมี Column (0,0) เกิดขึ้นมาเนื่องจากจำนวนสีของลูกอมมีทั้งหมด 3 สี ดังนั้นถ้าไม่ใช่ทั้งสีเขียวหรือสีแดง สีที่เป็นไปได้สีเดียวคือสีน้ำเงิน สามารถดูได้จากตัวอย่างด้านล่าง

Q Search this file...

1	Students	Candy	G	R
2	A	R	0	1
3	B	G	1	0
4	C	B	0	0
5	D	R	0	1

onehot1.tsv hosted with ♥ by GitHub

[view raw](#)

sklearn.preprocessing.OneHotEncoder()

Library ต่อมาที่สามารถใช้ในการทำ One-Hot encoding ในภาษา Python ได้เช่นกันคือ Scikit-learn ซึ่งเป็น Library ที่มีชุดคำสั่งมากมายเกี่ยวกับงาน Machine learning โดยคำสั่งที่ใช้จัดการเกี่ยวกับ Categorical variable อยู่ภายใต้ `.preprocessing` และให้ทำการเลือก `.OneHotEncoder()`

```

1 import numpy as np
2 from sklearn.preprocessing import OneHotEncoder
3
4 #One-Hot encoding
5 encoding = OneHotEncoder()
6 encoding.fit(np.array(df['Candy']).reshape(-1, 1))
7 dummies = encoding.transform(np.array(df['Candy']).reshape(-1, 1)).toarray()
8 df_onehot2 = pd.DataFrame(np.column_stack([df, dummies]), columns = [i for i in df
9 df_onehot2

```

onehot3.py hosted with ♥ by GitHub

[view raw](#)

การทำ One-Hot encoding ของ Scikit-learn มีข้อแตกต่างจาก Pandas อยู่บ้าง เพราะ Library ไม่ได้ Return ค่าออกมาเป็น DataFrame แต่จะเก็บไว้ในรูปแบบของ Encoder แทน ดังนั้นถ้าต้องการให้ผลลัพธ์จากการทำ One-Hot encoding ออกมาอยู่ในรูปของ DataFrame เพื่อใช้ทำงานต่อ ๆ ไป อาจต้องมีขั้นตอนการดึงค่าออกมาเพิ่ม

การใช้งาน `OneHotEncoder()` จาก Scikit-learn ถ้ามีข้อมูลมีเพียง 1 มิติ (*Feature*

ที่ต้องการเปลี่ยนเป็น One-Hot มี Column เดียว) ต้องมีการ `.reshape()` ให้อยู่ในรูป 1 มิติ ก่อนเริ่มทำงาน โดยให้ค่าเป็น `(-1, 1)`

เนื่องจากต้องการเก็บค่า Encoder เอาไว้เพื่อใช้สำหรับข้อมูลในครั้งต่อไป ดังนั้นจำเป็นต้องมีการ `.fit()` ข้อมูลก่อนที่จะ `.transform()` ถ้าใช้คำสั่ง `.fit_transform()` ไปทีเดียวเลย จะไม่สามารถเก็บ Encoder ระหว่างทางได้ ผลลัพธ์ที่ได้ออกมาอยู่ในรูปของ Numpy array ก็สามารถเปลี่ยนเป็น DataFrame เพื่อใช้ในการต่อไปได้เช่นกัน

ใน `OneHotEncoder()` ก็สามารถทำ Dummy Variable Encoding ได้เช่นกัน โดยใส่ Option `drop = 'first'` ค่า Return ที่ออกมาจะได้เหมือนกับ `drop_first = True` จาก Pandas

ข้อดีของการใช้ `.OneHotEncoder()` จาก Scikit-learn คือสามารถเก็บลักษณะการ Encode เพื่อใช้กับชุดข้อมูลอื่นได้ โดยที่ไม่จำเป็นต้องสร้าง DataFrame อย่างที่ Pandas ทำ

จากตัวอย่างใหม่ โดยให้มีนักเรียนใหม่เพิ่มขึ้นมาอีก 4 คน และให้มีลูกอมคนละสีเหมือนเดิม แต่ต่างจากข้อมูลก่อนหน้านี้คือลูกอมมีเพียง 2 สีคือ สีแดงและสีเขียวเท่านั้น ถ้าต้องใช้ Encoder จากที่เคยใช้กับข้อมูลก่อนหน้านี้กับข้อมูลชุดใหม่สามารถใช้ `encoding` ตัวเดิม และใช้ `.transform()` ข้อมูลใหม่ได้เลย ผลลัพธ์จากการ Encoding ที่ออกมา จะถูกอ้างอิงจากข้อมูลชุดเดิม

```
array([[0., 0., 1.],  
       [0., 0., 1.],  
       [0., 1., 0.],  
       [0., 1., 0.]])
```


ในทางกลับกัน ถ้าใช้ `pandas.get_dummies()` ผลลัพธ์ที่ออกมาจะถูกคำนวณบนข้อมูลชุดใหม่เท่านั้น ข้อมูลชุดใหม่ไม่มีลูกอมที่เป็นสีน้ำเงิน ดังนั้น Column One-Hot จึง Return ค่าออกมาเพียง 2 Columns เนื่องจากการทำงานของทั้ง 2 คำสั่งมีรายละเอียดที่แตกต่างกัน การใช้งานควรคำนึงถึงผลลัพธ์สุดท้ายเสมอ

Conclusion

สำหรับตัวอย่างการทำ One-Hot encoding ด้วย Pandas และ Scikit-learn สามารถจบลงสั้น ๆ ได้เพียงเท่านั้น ส่วนการเก็บค่า Encoder สามารถใช้ Pickle ในการเก็บการเข้ารหัสไว้ได้ ซึ่งคงมีโอกาสนำมาในครั้งต่อ ๆ ไป

[One Hot Encoding](#)[One Hot Encoder](#)[Classification Models](#)[Encoding](#)[Machine Learning](#)



Written by Sasiwut Chaiyadecha

699 Followers

Manager - Financial Risk Model

Follow



More from Sasiwut Chaiyadecha

Sasiwut Chaiyadecha in Analytics Vidhya

How to install dlib library for Python in Windows 10

Installation dlib library on Windows 10 by using pip install NOT conda install

3 min read · May 14, 2020



423



17



Sasiwut Chaiyadecha

รู้จัก Principal Components Analysis (PCA) ในเชิง Practical

การทำ Principal Components Analysis ด้วย Sklearn

3 min read · Nov 18, 2020




9



2



 Sasiwut Chaiyadecha

ARIMA Model ตอนที่ 1: เข้าใจ ARIMA แบบ Practical

รู้จักกับ ARIMA Model ในเชิง Practical

2 min read · Apr 13, 2020

 31 

 Sasiwut Chaiyadecha

Credit scoring / Rating model evaluation metrics

AUC, GINI, KS, Odds analysis, IV Analysis, PSI and CSI


5 min read · Aug 3, 2022

 5 

See all from Sasiwut Chaiyadecha

Recommended from Medium

 Brandon Wohlwend


Decision Tree, Random Forest, and XGBoost: An Exploration int...

In the digital age, data has emerged as a critical currency, driving growth and...

13 min read · Jul 24

 37 

 ...

 Frauke Albrecht in Towards Data Science

Decision Trees for Classification —Complete Example

A detailed example how to construct a Decision Tree for classification

8 min read · Jan 1

 242  3

 ...

Lists

Predictive Modeling w/ Python

20 stories · 651 saves

Natural Language Processing

920 stories · 436 saves

Practical Guides to Machine Learning

10 stories · 732 saves

The New Chatbots: ChatGPT, Bard, and Beyond


12 stories · 226 saves

 Paresh Patil

The ultimate guide to Encoding Numerical Features in Machine...

Table of Contents:




5 min read · Aug 18

 3   Amy Pajak

t-SNE: t-distributed stochastic neighbor embedding

An overview of t-SNE as a dimensionality reduction technique


11 min read · Jun 28

 10   Huda Saleh

Categorical Variables Encoding

Categorical variables are a common type of data encountered in machine learning...



6 min read · Sep 9

 3   Ding Han

How to tackle dataset class imbalance for NLP

In the field of machine learning, class imbalance poses a significant challenge...

7 min read · Jul 1

See more recommendations

[Help](#) [Status](#) [About](#) [Careers](#) [Blog](#) [Privacy](#) [Terms](#) [Text to speech](#) [Teams](#)