國立臺灣大學工學院工業工程學研究所

碩士論文

Institute of Industrial Engineering

College of Engineering

National Taiwan University

Master Thesis

針對二元變數提升預測表現的種類變數編碼方式

Encoding Methods for Binary to Categorical Variable for Prediction Performance Enhancement

楊雲皓

Yun-Hao Yang

指導教授：藍俊宏 博士

Advisor: Jakey Blue, Ph.D.

中華民國111年06月

June, 2022

# 摘要

人工智慧、機器學習與深度學習在近年來被廣泛的應用於各個領域當中；不論是影像辨識、自然語言處理或是機台數值預測。遍及了製造業、金融業、市場銷售、與影像醫學辨識的各領域實作者，皆前仆後繼地設法將機器學習應用於遭遇的問題上，來提升日常工作的效率與準確度。然而，機器學習的模型表現並不端看模型建置的技巧與超參數的調教與設置，資料的前處理與編碼方式對於模型表現也有著極為深遠的影響。例如，在處理含有字串的類別特徵時，我們便需使用獨熱編碼 (One-hot encoding)來將類別變數中的字串特徵轉換成二元變數，以利於作為輸入資料供模型讀取。但若是類別變數的種類繁多，進行獨熱編碼後將產生眾多的二元變數特徵，如此將稀釋該特徵的資訊、並面臨維數災難 (Curse of dimensionality)這一窘境。此外編碼出的二元特徵也不全然與分類、回歸有關聯關係；更甚者，二元變數特徵本身在一定程度上便違反了機器學習演算法的假設。

本研究提出了監督式與非監督式的編碼方式協助二元資料轉換為整數型別的類別變數，以求解決以上提及的問題。透過二元特徵之間的關聯關係、PCA權重等方式群組特徵，再根據各組間特徵自身的屬性進行排序後編碼成整數的類別變數。力求在縮減維度、提升處理速度的同時，維持模型的整確性與變數的可解釋性。

**關鍵字**：類別變數、獨熱編碼、監督式／非監督式學習、二元變數編碼

# Abstract

AI techniques have recently been widely applied to the tasks of image recognition and natural language processing. Practitioners from fields such as manufacturing, finance, marketing, and radiology are eager to implement AI methods to enhance daily efficiency and effectiveness. However, AI method performance depends on not only the modeling skills and hyperparameters tuning but also the data preprocessing and encoding. While handling categorical variables, one-hot encoding is commonly used to convert strings into binary features, which can then serve as the input for model training/testing. If the number of categorical levels is large, it consequently creates a large number of features, and the curse of dimensionality would be an essential concern. Furthermore, the one-hot encoding features are created based on the levels of categorical variables and do not guarantee to be related to the classification/regression tasks. Not to mention that the binary feature values often violate the assumptions in machine learning algorithms.

In this research, we develop unsupervised and supervised encoding methods to tackle the aforementioned issues. In unsupervised encoding, we compare the feature properties, such as the column sparsity, PCA-weight, and feature importance, for consolidating related features into a semi-continuous one via binary encoding. In supervised encoding, an optimization scheme is proposed to incorporate the performance improvement of the classifier/regressor and the consolidating orders of the binary features. It is expected to reduce the number of binary features significantly as well as to enhance the classification/regression accuracy through inputting the consolidated features.

**Keywords**—categorical variable, one-hot encoding, supervised/unsupervised encoding, binary encoding

# 目錄

(只要樣式定義完善，目錄最後可以自動產生)

# 圖目錄

(只要樣式定義完善，目錄最後可以自動產生)

# 表目錄

(只要樣式定義完善，表目錄最後可以自動產生)

# 緒論

本章節將描述將二元變數特徵作為機器學習模型的輸入時，所面臨到的窘境。

## 1.1 研究背景

在機器學習的過程之中，難免會遭遇到字串型別的類別特徵。為了將類別特徵輸入模型之中，必須透過各式編碼方式 (Encoding methods)轉換類別變數；例如，當面臨以字串描述體積的類別特徵：{「大」、「中」、「小」}時，我們可以依照相對體積的順序關係，將其編碼為{「大：3」、「中：2」、「小：1」}如此便能作為模型的輸入，此為序號編碼（Ordinal encoding）。然而，當今天面臨的是描述城市種類的類別變數：{「紐約」、「倫敦」、「東京」}時，由於城市之間並不存在著明顯的順序關係，若是編碼成{「紐約：1」、「倫敦：2」、「東京：3」}會使得訓練模型誤解城市之間的關聯性。為此可以考慮使用頻率編碼 (Frequency encoding)或是目標編碼 (Target encoding)，透過變數出現的頻率或是百分比取代變數本身，在種類較少、比例平衡的資料集當中可採用以上兩種編碼方式，編碼成：{「紐約：0.3」、「倫敦：0.2」、「東京：0.5」}。倘若面對的特徵種類繁多，或含有比例相近，則會使這兩種編碼優勢消失：{「紐約：0.2」、「倫敦：0.2」、「東京：0.2」、「台北：0.3」、「上海：0.1」}從而將不同種類的變數混淆。

當面臨類別變數不存在順序關係、種類多且出現比例相近時，大多數的情形下只能透過讀熱編碼 (One-Hot encoding)來為每個種類產生新的虛擬變數(Dummy variable)：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| City | | | | |
| New York | London | Tokyo | Taipei | Shanghai |
| 0 | 1 | 0 | 0 | 0 |

這些虛擬變數間將存在互斥關係，因而有高度的關聯性，並且稀釋了原先單一特徵的資訊。而特徵中的種類越多，也將產生更多的虛擬變數，在模型訓練時也將消耗更多的記憶體與運算時間。但在製造業當中，讀熱編碼過、同時虛擬欄位眾多的資料卻相當常見，常見於製品於製造過程中通過的機台或是工序，該如何前處理這些相互關聯的二元特徵也成了一大問題。



圖1 Lakers Logo.

內文

表1 表格範例。

|  |  |  |
| --- | --- | --- |
| 這是表格 | 這是表格 | 這是表格 |
| 這是表格 | 這是表格 | 這是表格 |
| 這是表格 | 這是表格 | 這是表格 |
| 這是表格 | 這是表格 | 這是表格 |

內文

### 1.1.1 (使用標題3樣式)

內文

## 1.2 研究動機與目的

本研究試圖透過監督以及非監督式的方法，對眾多相互關聯的二元變數進行群組、排序與編碼後，以求大幅度縮減資料維度、壓縮資料資訊、縮減模型讀取時間、並在一定程度上維持或提升機器學習模型對於資料的分類結果。

* 對於原先包含眾多的二元虛擬變數資料進行維度縮減：
* 與原先二元虛擬變數資料相比，維持或提升新編碼過後資料的分類成果：

## 1.3 研究架構

# 文獻探討

本章節

# (方法創見)

內文

|  |  |
| --- | --- |
|  | (11) |

(方程式最後再把外框改為透明即可)

# 案例研討

內文

# 結論與建議

內文

# 參考文獻列表

APA格式，依照第一作者姓的字母排序，第一作者同姓則比第二作者，如果兩論文作者全部相同，則以年代較舊者先列。各式文獻範本如下：

<https://apastyle.apa.org/style-grammar-guidelines/references/examples>

<https://subjectguides.library.american.edu/c.php?g=675993&p=4847486>

# 附錄 A (如果有)