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

本章節將描述將多維度二元特徵作為機器學習模型的輸入時，所面臨到的窘境與難題，並提及本研究的目的與架構。

## 1.1 研究背景

在機器學習的過程之中，處理輸入資料即時，時常會遭遇到字串型別特徵，像是對於受測者血型、居住城市的描述皆以字串形式呈現。為了將類別特徵輸入模型之中，則必須透過各種編碼方式（Encoding method）來對無法作為模型輸入的類別變數（Categorical variable）做可變編碼（Variable encoding）；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）來為各個種類產生新的虛擬變數（Dummy variable）：

表1 獨熱編碼後產生的虛擬變數。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | City | | | | |
|  | New York | London | Tokyo | Taipei | Shanghai |
| Sample 1 | 0 | 1 | 0 | 0 | 0 |
| Sample 2 | 1 | 0 | 0 | 0 | 0 |

這些虛擬變數之間將存在強烈的互斥關係，因而有高度的關聯性，並且稀釋了原先單一特徵的資訊。而類別特徵中的種類越多，也將產生越多的虛擬變數，使得模型訓練時也將消耗更多的記憶體與運算時間，導致模型難以收斂和訓練。但是在製造業當中，讀熱編碼過、同時虛擬欄位眾多的二元特徵資料卻相當常見，常用於描述製品於製造過程中通過的機台或是工序、以及品質管制當中表示合格與否的檢測項目；該如何前處理這些眾多且相互關聯的二元特徵也成了一大難題。



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式，對眾多相互關聯的二元特徵進行群組、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。主要目標可分為以下四點。

1. 縮減資料特徵個數：

維度的增加對於機器學習模型的成果有著深遠的負面影響，過多的維度將使得模型難以收斂、延長訓練資源與計算時間。透過對二元特徵群組後進行編碼，將能大幅度的縮減特徵總數；在縮減資料大小以利儲存的同時，也減低模型在資料讀取的時間。

1. 壓縮特徵資訊：

經由讀熱編碼過後，總體資訊不改變，但特徵總數的上升；意味著眾多二元特徵瓜分了原先單一特徵所包含的資訊。透過群組後編碼二元特徵，回復特徵平均的資訊含量與重要性。

1. 產生相互獨立的特徵：

多數的二元特徵與彼此有著強烈的互斥關係，而特徵與特徵之間的不

獨立導致的共線性問題，也代表著特徵能用以預測彼此。透過群組相關性、相似程度過高的特徵並編碼，來解決特徵與特徵嚴重共線性的問題。

1. 維持編碼過後資料的分類成果：

只包含二元特徵的資料，即只存在大量0、1的資料對於機器模型的訓練與預測也造成一大麻煩。當模型再進行優化與求解時大多仰賴梯度計算，而眾多的0將使模型無法計算梯度，導致模型的訓練緩慢、劣化預測成效等負面影響。

## 1.3 研究架構

本研究程序如下圖所示，主要目標為透過對於二元特徵的資料進行群組後的排序、與編碼，來壓縮資料並維持一定程度的分類成果。在緒論中簡介研究目標與架構；由第二章文獻回顧闡明變數編碼的用意與目的、多維度二元特徵資料在以機器學習模型分類時所遭遇的難題。第三章研究方法中，描述本研究如何對於多維度二元特徵資料群組特徵、排序組間特徵、以及編碼各個特徵組。第四章案例分析將透過產生的測試資料結合本研究所提出之方法，與原始資料一同給予機器學習模型做分類成果的比較。第五章結論建議部分則對於根據案例實施的成果做出總結，並歸納出未來本研究的後續發展方向。

# 文獻探討

本章節探討研究欲解決之問題，與研究之相關文獻

## 2.1 變數編碼

在甫收集完資料，進入機器模型建模之前，必須先經過的資料的前處理(Data Preprocessing)這一步驟。資料的清理與型態調整、資料探索與視覺化(EDA)或是特徵工程(Feature Engineering)都是本階段的目標；而變數編碼便旨在調整資料的型態，以轉換成適合模型的輸入，此過程也稱為(ETL, Extract-Transform-Load)。

在…中，探討了不同的編碼方式將對於ANN分類模型訓練完後的預測準確度有著影響。可見在選擇該特定的變數編碼前，必須先明白各個特徵本身的刻度關係，在選擇相對應的編碼方式，才能使機器學習模型能正確的識別特徵關係，並獲得更佳的分類成果。

|  |  |
| --- | --- |
| Encoding Technique | Accuracy (Percentage) |
| One Hot Coding | 90 |
| Ordinal Coding | 80 |
| Sum Coding | 95 |
| Helmert Coding | 89 |
| Polynomial Coding | 91 |
| Backward Difference Coding | 95 |
| Binary Coding | 90 |

而…，中則將變數的刻畫區分的相當詳盡，能依據不同經驗法則的辨別標準、與數值結構進行區分，分為名義尺度（Nominal scale）、順序尺度（Ordinal scale）、等距尺度（Interval scale）、以及比例尺度（Ratio scale）等四種不同的變數型態。在根據不同的變數型態選用適當的變數編碼方式，來對類別變數進行編碼。

|  |  |  |  |
| --- | --- | --- | --- |
| Scale | Basic empirical operations | Mathematical group structure | Permissible statistics(invariantive) |
| Nominal | Determination of equality | Permutation group | Number of cases  Mode  Contingency correlation |
| Ordinal | Determination of greater of less | Isotonic group | Median percentiles  Percentiles |
| Interval | Determination of equality of intervals of difference | General linear group | Mean  Standard deviation  Rank-order correlation  Product-moment correlation |
| Ratio | Determination of equality of ratios | Similarity group | Coefficient of Variation |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 類別 | 運算子 | 範例 |
| 名義尺度 | 非計量 | =、≠ |  |
| 順序尺度 | =、≠、>、< |  |
| 等距尺度 | 計量 | =、≠、>、<、+、− |  |
| 比例尺度 | =、≠、>、<、+、−、×、÷ |  |

在這四種變數型態中，順序尺度、等距尺度、以及比例尺度皆是以數值方式呈現，可直接交由機器學習模型作為輸入，唯獨名義尺度以字串、或是布林的型別出現，為此勢必需要進行變數編碼，以處理無法接受數學運算的變數型態，常見的手法有獨熱編碼（One-Hot encoding）、二進制編碼（Binary encoding）來轉換該類別特徵成許多的虛擬變數。

### 2.1.1獨熱編碼

讀熱（One-hot）為在數位電路與機器學習領域之中，描述一種位元組或是向量的表現形態。在同一樣本之中，互相關連的讀熱欄位群組中只允許存在一個1，其餘相關欄位必須為零；而在統計、經濟學中，這些相關的讀熱欄位則被稱呼為虛擬變數。

當面對的資料特徵並非數值、且種類之間沒有物理與特性上的順序時，便需以讀熱編碼進行變數的轉換，來避免模型誤解種類之間存在特定的關聯關係。讀熱編碼的想法為透過N個虛擬變數來描述原先的種類特徵。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 顏色 | 🡺 | 紅色 | 黃色 | 藍色 |
| 紅色 | 1 | 0 | 0 |
| 黃色 | 0 | 1 | 0 |
| 藍色 | 0 | 0 | 1 |
| 黃色 | 0 | 1 | 0 |

讀熱編碼在一定程度上協助了機器學習模型遭遇屬性變數的時的處理能力，雖然在同時間使得資料種維度上升。假如今天的種類特徵是在描述台灣的21個縣市，則讀熱編碼過後便會產生21個虛擬變數欄位；雖然提升了特徵總數，但總體的資訊卻沒有增加，意味著讀熱欄位的1將零散地被0所包圍，並散落在這些虛擬變數的之中，導致資料趨為稀疏。而稀疏的資料列間存在於完全的互斥關係，只要掌握樣本讀熱的特徵位置便能預測出其餘特徵數值，使得讀熱後的高維資料存在著嚴重的並行性與共線性等問題。

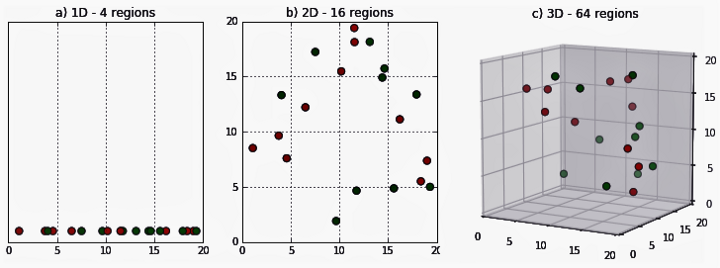
### 2.1.2二進制編碼

相較於獨熱編碼，二進制編碼以更少的虛擬變數描述了相同數量的特徵種類，缺點則在於面對無序特徵時，二進制編碼後的虛擬變數無法有效的解釋變數所包含的意義。

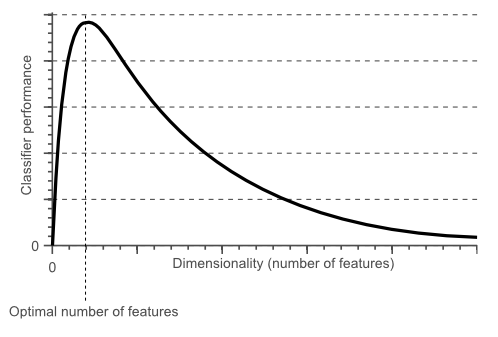
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 居住城市 | 順序編碼 |  | 二進制編碼 | | |  | 獨熱編碼 | | | | |
| 台北 | 0 |  | 0 | 0 | 0 |  | 0 | 0 | 0 | 0 | 0 |
| 桃園 | 1 |  | 0 | 0 | 1 |  | 0 | 0 | 0 | 0 | 1 |
| 新竹 | 2 |  | 0 | 1 | 0 |  | 0 | 0 | 0 | 1 | 0 |
| 台中 | 3 |  | 0 | 1 | 1 |  | 0 | 0 | 1 | 0 | 0 |
| 台南 | 4 |  | 1 | 0 | 0 |  | 0 | 1 | 0 | 0 | 0 |
| 高雄 | 5 |  | 1 | 0 | 1 |  | 1 | 0 | 0 | 0 | 0 |

## 2.2 維度災難

維度災難（Curse of dimensionality）是描述在樣本總數不改變時，當特徵（即樣本的空間維度）增加時，將面臨到的難題；包括資料分布範圍增大，而導致樣本之間距離增大、數據變為稀疏；因而造成高維空間中的樣本數量不足，低維度的空間特性無法推廣至高維空間，使歐式距離的計算與對資料的常態假設失去效用。



而過多的維度將使得數機器學習模型擬合訓練資料中的噪音誤差、模型訓練使用、調整的參數增加，無法對測試資料做出適當的泛化而導致過擬合的情形；難以收斂與有效的訓練模型之外，同時也將導致訓練時間的增加，分類成效也將隨著維度上升而下降。



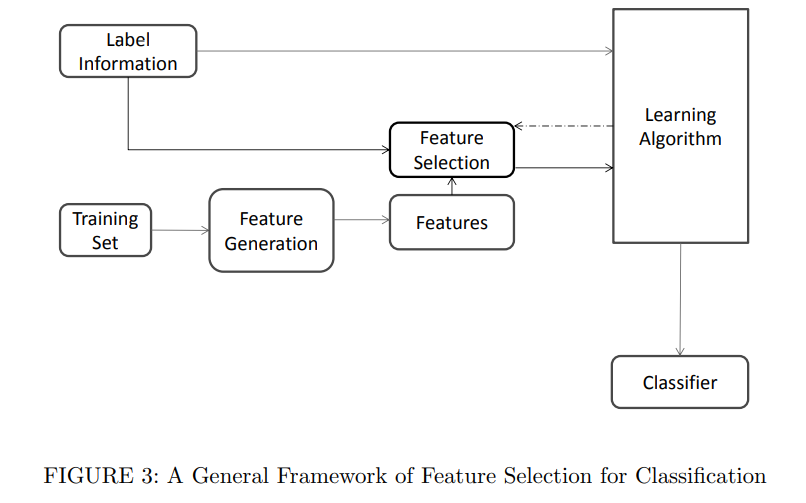
然而，隨著近年來資料分析技術與接器學習的盛行，高維度資料集也變得相當的廣泛、常見。而該如何處理高維度資料也成了及富有研究價值的主題。目前通常透過降低維度（Dimension reduction）做為高維資料的前處理方式，而降維方法可分為以下三類：

1. 正則化回歸：Lasso、Ridge
2. 線性降維：主成分分析（PCA）、線性判別分析（LDA）、MDS
3. 非線性降維：局部線性嵌入（LLE）、T-SNE

其目的皆在以降低變數、維度的個數來描述原先的高維度資料，同時保留原先資料樣本之間的特性與關聯性，便於資料視覺化與模型訓練。

## 2.3 特徵選取

特徵選取（Feature selection）旨在假設數據之中含有許多冗餘或無關的特徵，並透過從資料集中移除部分不具備夠多信息的特徵，從原有資料集的特徵之中挑選出最具代表性、富有資訊的重要特徵子集合。若特徵選取得當，包含了極具鑑別能力的最優特徵子集，便能達到簡化機器學習模型的訓練時長、避免過度擬合、提升模型準確度，以及便於理解特徵於模型輸出之間的關聯關係等目的。



依據不同特徵選取的方式，可以區分為三類：

1. 過濾法（Filter）：透過指標評比每一特徵，並依據閾值或者預選定特徵個數案大小排名選取；這些指標可能為皮爾森相關係數、解釋變異等。
2. 包裝法（Wrapper）：用模型測試、評比若干不同的特徵子集，並依據分數排除或選取特徵。
3. 嵌入法（Embedded）：與過濾法相似；即在模型與演算法訓練的同時計算各特徵權重與指標分數，依此進行選取。

### 2.3.1 主成分分析（PCA）

### 2.3.2 Correlation

## .4 評估指標

# 針對多維度二元特徵資料的變數編碼

經由第二章文獻探討得以發現，以多維度二元特徵資料作為機器學習模型的輸入時所遭遇到的難題，

## 3.1依據特徵工程群組二元特徵

群組對於具備有相同物理意義、具有相關性、或是重要度相近的二元特徵。若

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

(方程式最後再把外框改為透明即可)

## 3.2 排序組間特徵

群組完相關特徵後，將依據特徵本身屬性，再對各組內特徵做排序，來調動編碼過後的數值。這些屬性包括但不限於特徵總值，gini impurity，或以隨機指派的方式作為排序依據，而後對比不同的排序方式對於分類結果的影響。

除了以特徵屬性作排列依據之外，也可以將此問題描述為一最佳化問題，嘗試以不同優化手法進行求解，例如基因演算、捷思法等等。甚至是先預訓練模型，再依據模型給出的特徵重要度進行排列。

## 3.3 對各特徵組進行二進碼十進數編碼

針對各組排序過後的二元特徵組，進行二進碼十進數編碼，產生新的整數型別的類別變數，新資料的特徵個數將等於原先的二元特徵群組數，然而，新編碼過後的資料因為原先數字1的分布稀疏，也將導致新資料全距過大、且分布稀疏；例如：若組內二元特徵的個數為十個，則此群組編碼出的整數類別變數全距範圍將達到0-1024。為此，可透過排名編碼為了改善編碼後資料之間的稀疏程度，同時避免儲存過大正整數，導致整數溢位等問題。

# 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)