Day 25 特徵工程

類別型特徵 - 均值編碼

出題教練 陳明佑

本日知識點目標

- 知道當類別特徵與目標明顯相關時，該用什麼編碼方式
- 知道均值編碼可能有什麼問題
- 知道應該使用何種方式修正均值編碼的問題

知識地圖 特徵工程 類別型特徵 - 均值編碼

特徵工程

監督式學習
Supervised Learning

前處理 Processing → 探索式數據分析 Exploratory Data Analysis → 特徵工程 Feature Engineering → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

非監督式學習
Unsupervised Learning

分群 Clustering

降維 Dimension Reduction

特徵工程 Feature Engineering

概論

數值型特徵

類別型特徵

時間型特徵

填補缺值 → 去離群值 → 去偏態 → 特徵縮放

類別型特徵處理

時間型特徵處理

特徵組合

特徵篩選

特徵評估

均值編碼 (1 / 2)

額外線索：如果類別特徵看起來與目標值有顯著相關，應該如何編碼?

行政區	房產價位
大安區	4500萬
南港區	1500萬
大安區	3500萬
大安區	2500萬
南港區	1800萬
文山區	2000萬

?

均值編碼 (2 / 2)

均值編碼 (Mean Encoding) ：使用目標值的平均值，取代原本的類別型特徵

*在部分模型中，使用均值編碼作為類別型特徵預設編碼方式

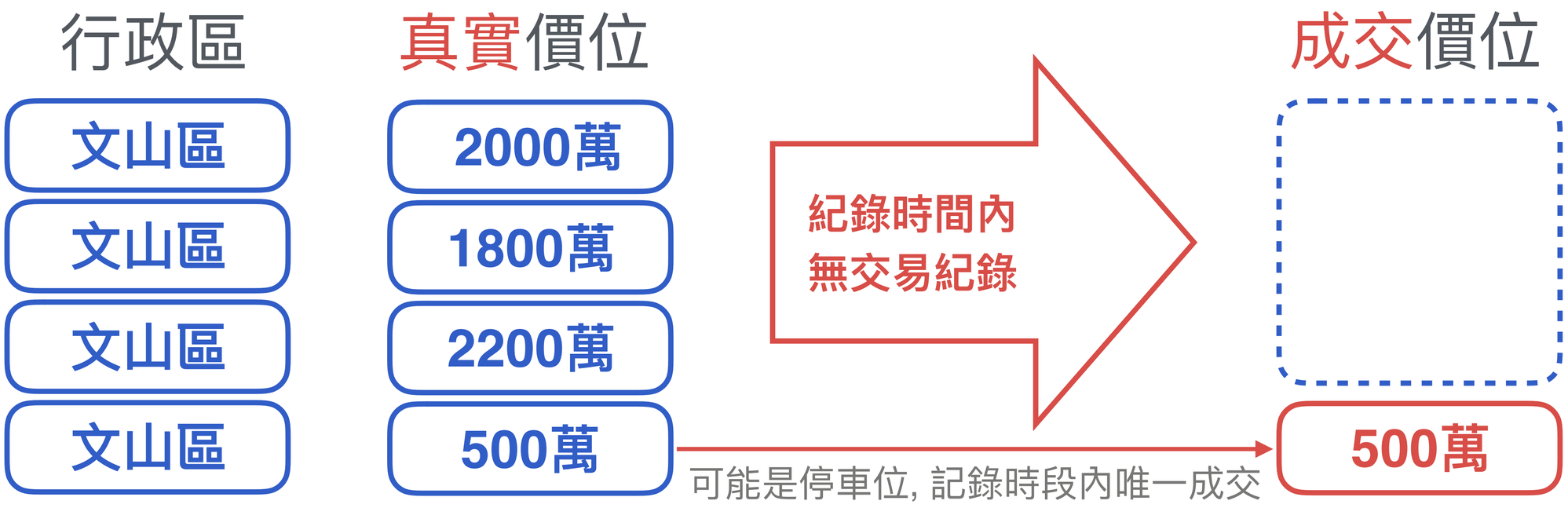
行政區	房產價位
大安區	4500萬
南港區	1500萬
大安區	3500萬
大安區	2500萬
南港區	1800萬
文山區	2000萬

行政區－均值編碼

3500萬	$= \frac{4500+3500+2500}{3}$ 萬
1650萬	$= \frac{1800+1500}{2}$ 萬
3333.3萬	
3333.3萬	
1650萬	
2000萬	= 2000 萬

平滑化 (Smoothing) (1 / 2)

如果交易樣本非常少, 且剛好抽到極端值, 平均結果可能會有誤差很大



想想看：這個問題如何解決?

平滑化 (Smoothing) (2 / 2)

因此, 均值編碼還需要考慮紀錄筆數, 當作可靠度的參考

行政區	價位均值	記錄筆數	可靠度
文山區	500萬	X1	低
大安區	3000萬	X100	高
南港區	1700萬	X10	中

- 當平均值的可靠度低時, 我們會傾向相信全部的總平均
- 當平均值的可靠度高時, 我們會傾向相信類別的平均
- 依照紀錄筆數, 在這兩者間取折衷

平滑化公式與小提醒

均值編碼平滑化

$$\text{新類別均值} = \frac{\text{原類別平均} * \text{類別樣本數} + \text{全部的總平均} * \text{調整因子}}{\text{類別樣本數} + \text{調整因子}}$$

*調整因子用來調整平滑化的程度，依總樣本數調整

小提醒：均值編碼容易 overfitting

雖然均值編碼符合直覺, 並且也是強大的編碼方式
但實際上使用時很容易 overfitting (即使使用了平滑化)
所以需確認是否適合再使用 (用 cross validation 確認使用前後分數)

重要知識點複習

CUPOY

- 當類別特徵與目標明顯相關時，該考慮採用**均值編碼**
- 知道均值編碼最大的問題，在於**相當容易 Overfitting**
- **平滑化**的方式能修正均值編碼容易 Overfitting 的問題，但效果有限，因此仍須**經過檢驗**後再決定是否該使用均值編碼

延伸閱讀

除了每日知識點的基礎之外，推薦的延伸閱讀能補足學員們對該知識點的了解程度，建議您解完每日題目後，若有多餘時間，可再補充延伸閱讀文章內容。

推薦延伸閱讀

平均數編碼 ：針對高基數定性特徵(類別特徵)的數據處理/ 特徵工程

知乎 網頁連結

- 就實務上而言，均值編碼的意義在於當一個特徵有明顯意義，但是類別數量特別多(這裡說的「高基數」)時可能有用，但最麻煩的點在於極度容易OverFitting， 所以需要不同的平滑化方式。
- 在課程內使用平均因子的方法只是其一，這邊的內容也介紹了另一種較複雜的平滑化方式，提供同學參考。

解題時間

It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題