

D12：數值型特徵-補缺失值與標準化



簡報閱讀



範例與作業



問題討論

數值型特徵- 補缺失值與標準化 >

知識地圖 >

本日知識點目標 >

填補缺值 >

複習：填補缺值的方式 >

為何要標準化 >

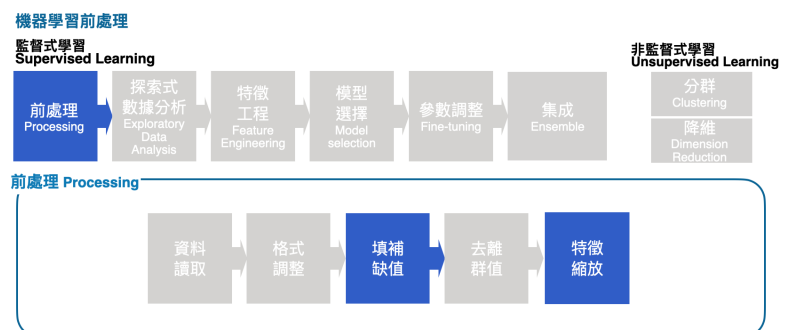
複習：標準化 / 最小最大化 >

標準化 / 最小最大化適用場合 >

數值型特徵- 補缺失值與標準化



知識地圖



本日知識點目標

本日知識點目標

- 資料當中，缺失值應該怎麼補？
- 補缺失值時該注意什麼？
- 將資料標準化的意義在哪裡？
- 什麼時候該用標準化？什麼時候又該用最大最小化呢？

填補缺失值

看答案前先想一想：下列幾種缺失值怎麼補最好？
(問號表示缺失值)

停車位	房間數	屋齡	行政區
True	1	11	?
?	?	20	南港區
?	2	32	大安區
True	1	?	南港區
True	?	25	?
?	1	?	文山區

停車位	房間數	屋齡	行政區
True	1	11	南港區(眾數)
False	0	20	南港區
False	2	32	大安區
True	1	屋齡總平均	南港區
True	0	25	南港區(眾數)
False	1	屋齡總平均	文山區

沒有False但應該有，推測False表示成為空白

沒有0但應該有，推測0表示成為空白

屋齡不可能為空值，推測應該是資料遺失或者漏填，故可取總平均或中位數

行政區不可能為空值，推測應該是漏填，故可取行政區眾數或另創一值

最重要的是欄位的領域知識與欄位中的非缺數

複習：填補缺值的方式

填補統計值

1. 填補平均值(Mean)：數值型欄位，偏態不明顯
2. 填補中位數(Median)：數值型欄位，偏態很明顯
3. 填補眾數(Mode)：類別型欄位

填補指定值 - 需對欄位領域知識已有了解

1. 補 0：空缺原本就有 0 的含意，如前頁的房間數
2. 補不可能出現的數值：類別型欄位，但不適合用眾數時

填補預測值 - 速度較慢但精確，從其他資料欄位學得填補知識

1. 若填補範圍廣，且是重要特徵欄位時可用本方式
2. 本方式須提防 **overfitting**：可能退化成為其他特徵的組合

為何要標準化

想一想：競賽中的給分，如果發生下列情形... 要如何修正呢？

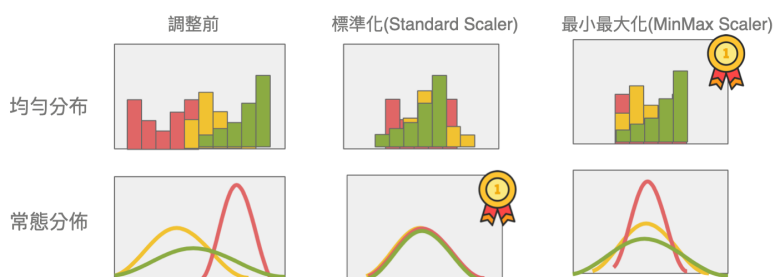
10	7	7	8.5
10	5	8	7.6
0	8	9	8.3
0	6	9	8.1
10	7	8	8.6

評審1	原始值	修正後	評審4	原始值	修正後
MIN	0	0		8.1	0.5
MAX	10	1		8.5	0.9
	10	1	MIN	7.6	0
	0	0		8.3	0.7
	0	0		8.1	0.5
	10	1	MAX	8.6	1

標準化：以合理的方式，平衡特徵間的影響力
此處範例為「最大最小化」，一般常用的方法還有「標準化」

複習：標準化 / 最小最大化

- 標準化 (Standard Scaler)：假定數值為常態分佈，適合本方式平衡特徵
- 最小最大化 (MinMax Scaler)：假定數值為均勻分佈，適合本方式平衡特徵



標準化 / 最小最大化適用場合

- **樹狀模型或非樹狀模型(參考今日練習題)**

非樹狀模型：如線性迴歸, 羅吉斯迴歸, 類神經...

等，標準化 / 最小最大化後對預測**會有影響**

樹狀模型：如決策樹, 隨機森林, 梯度提升樹...

等，標準化 / 最小最大化後對預測**不會有影響**

- **標準化 / 最小最大化 使用上的差異**

標準化：轉換不易受到極端值影響

最小最大化：轉換容易受到極端值影響

註：因此，去過離群值的特徵，比較適用最大最小化

重要知識點複習

- 補缺失值的方法因**特徵類型**與**缺的意義**不同，會有許多不同補法，需要因資料調整，無法一概而論
- 除了上面兩點，補缺失值還要注意盡量**不要破壞資料分布**
- 標準化的意義：**平衡**數值特徵間的影響力
- 因為**最大最小化**對**極端數值**較敏感，所以如果資料不會有極端值，或已經去極端值，就適合用最大最小化，否則請用標準化

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

掘金：Python數據分析基礎：數據缺失值處理



首頁

沸點

話題

小冊

活動

搜索掘金



Python數據科學

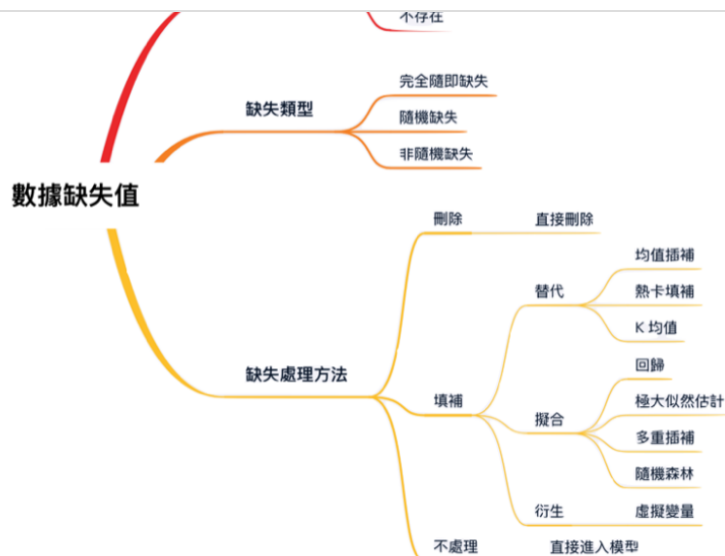
2018年07月28日 閱讀2227

關注

【Python數據分析基礎】：數據缺失值處理

作者：xiaoyu

這篇文章更詳細地介紹了各種缺失值的種類，以及處理的各種方式優缺點，如果要徹底搞懂缺失值的話，這是一份不錯的補充資料。但是我們還是要強調：補缺是因資料而異，所以熟悉方法與觀察資料本身都是同樣重要的，因此這在實務上會是一個辛苦環節。



數據標準化 / 歸一化normalization

數據標準化 normalization

連續型特徵還有一種處理方式是，先分桶/分箱（如等頻/等距的分）進行

blog.csdn.net

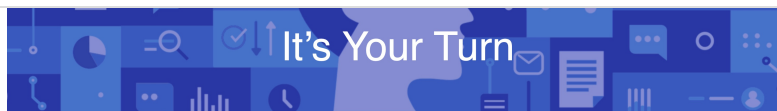
本文重點如下圖，介紹了標準化 / 最大最小化以外的一些特徵縮放方式，雖然文中也提到這兩種就是最常見的方式了，但是其他幾種方式也是不錯的參考，提供同學查閱。

常見的數據歸一化方法

最常用的是 min-max 標準化 和 z-score 標準化。

min-max 標準化(Min-max normalization)/0-1 標準化(0-1 normalization)/線性函數歸一化/離差標準化

解題時間



Sample Code & 作業
開始解題



[下一步：閱讀範例與完成作業](#)