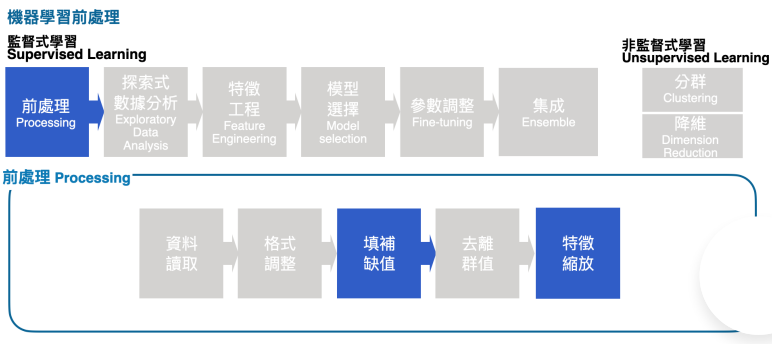
D11：常用的數值取代：中位數與分位數連續數值標準化

簡報閱讀 範例與作業 問題討論

數值填補與連續數值標準化

知識地圖



- 如何處理例外值
- 如何進行數據標準化

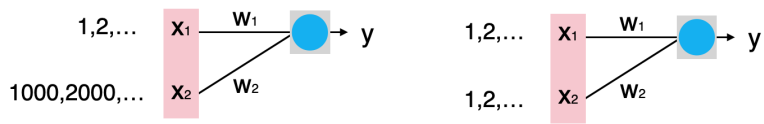
常用以填補的統計值

常用以填補的統計值	方法
中位數 (median)	np.median(value_array)
分位數 (quantiles)	np.quantile(value_arrar, q = …)
眾數 (mode)	scipy.stats.mode(value_array): 較慢的方法 dictionary method: 較快的方法
平均數 (mean)	np.mean(value_array)

連續型數值標準化

為何要標準化

改變一單位的 x2 對 y 的影響完全不同



是否一定要做標準化 (有沒有做有差嗎)

看使用的模型而定

- Regression model：有差
- Tree-based model：沒有太大關係

Requires little data preparation. Other techniques often require data normalization. Since trees can handle qualitative predictors, there is no need to

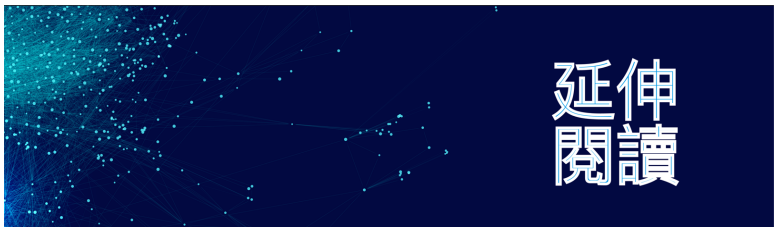
常用的標準化方法	公式
Z 轉換	$\frac{(x-mean(x))}{std(x)}$
空間壓縮	Y= 0 ~ 1, $\frac{x-min(x)}{max(x)-min(x)}$
	Y= -1 ~ 1, $(\frac{x-min(x)}{max(x)-min(x)}-0.5)*2$
	Y= 0 ~ 1, (針對特別影像), $\frac{x}{255}$

特殊狀況

有時候我們不會使用 min/max 方法進行標準化，而會採用 Qlow/Qhigh normalization

(如將空間壓縮第一例中的 min 改為 q1, max 改為 q99)

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

Is it a good practice to always scale/normalize data for machine learning?

> Is it a good practice to
>
> My understanding is that when some features have different
>
> stats.stackexchange.com

閱讀重點：

1. 某些演算法 (如 SVM, DL) 等，對權重敏感或對損失函數平滑程度有幫助者
2. 特徵間的量級差異甚大

- Bad

1. 有些指標，如相關不適合在有標準化的空間進行
2. 量的單位在某些特徵上是有意義的

解題時間

Sample Code &作業
開始解題

下一步：閱讀範例與完成作業