資料前處理 -特徵縮放 與 獨熱編碼

Feature Scaling and One Hot Encoding

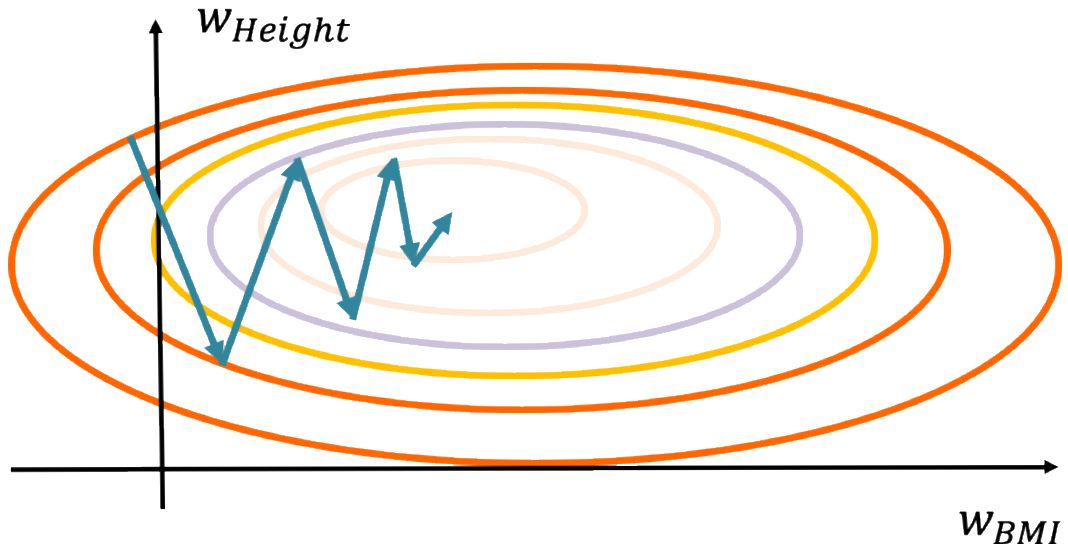
國立東華大學電機工程學系 楊哲旻

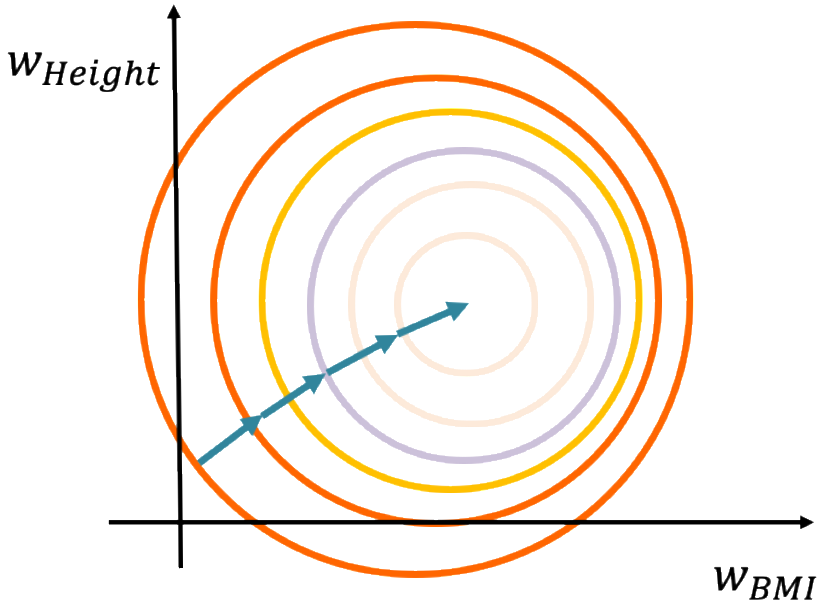
Outline

1. 特徵縮放
2. 重新縮放 與 平均值正規化
3. 獨熱編碼

資料前處理　01. 特徵縮放

若是**連續型**的資料使用**特徵縮放(Feature Scaling)**的方式，更能讓模型更加擬和數據。考量輸入值的範圍及單位通常較不一致，如身高(cm)與身體質量指數(kg/m^2)的範圍就有相當大的差別，此時會得到像左圖一樣相當狹長, 利用此種比例進行求解時，由於的更新不能太快，會使得更新過慢, 導致整體梯度下降法收斂過慢。





資料前處理　02. 重新縮放 與 平均值正規化

重新縮放(Rescaling)

將每一維特徵線性映射到目標範圍[a, b]，即將最小值映射為a，最大值映射為b，常用目標範圍為最小最大正規化(min-max normalization) [0, 1] 和 [-1, 1]

$$Range\ scaling\ :\quad x' = a + \frac{(x - min(x))(b - a)}{max(x) - min(x)}$$

平均值正規化(Mean normalization)

$$Mean\ normalization\ :\quad x' = \frac{x - average(x)}{max(x) - min(x)}$$

資料前處理　03. 獨熱編碼

離散型的資料則使用**獨熱編碼(One-Hot Encoder)**，它又稱為虛擬變數(Dummy variables)，常用於特徵與標籤中，讓每個特徵或標籤的類別彼此距離是相同的，二進制化的原因是不同值之間存在一些有正確的距離關係，但特徵數量會變多

編碼前的資料

ID	交通工具
0	大眾交通工具
1	走路
2	機車
3	走路
4	機車
…	…

排序型編碼

ID	交通工具
0	2
1	0
2	1
3	0
4	1
…	…

獨熱編碼

ID	走路	機車	大眾交通工具
0	0	0	1
1	1	0	0
2	0	1	0
3	1	0	0
4	0	1	0
…	…	…	…