人工智能之自動化光學檢測實務

Yi-Yung Chen

1

適用影片單元01

資料預處理(Data Preprocessing)

- 下載數據集
- 導入資料庫
- 導入數據集
- 缺失數據
- 分類數據
- 特徵縮放
- 將數據分成訓練集與測試集
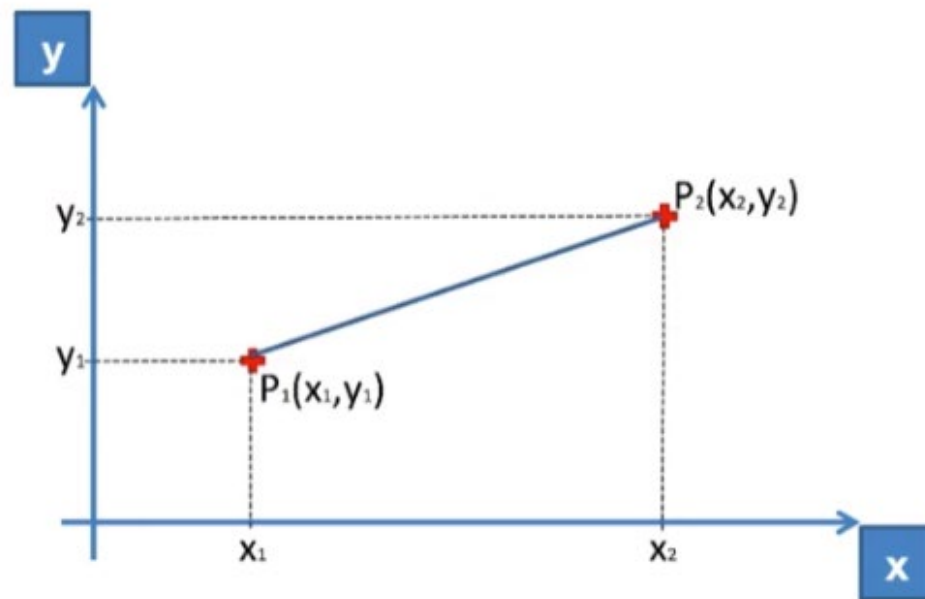
數據資料

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

數據資料_自變量、應變量

自變量			因變量
Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

歐式距離 (Euclidean Distance)

Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

數據資料_缺失數據

缺失數據

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

數據資料_分類數據

分類數據

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

分類數據

- 文字轉換為數字
- 避免數字對原分類產生大小差異
  - 虛擬編碼(Dummy Encoding)

例如：
衣服尺寸 vs. 庫存
衣服尺寸 vs. 成本

Country
France
Spain
Germany
Spain
Germany
France
Spain
France
Germany
France

France	Spain	Germany
1	0	0
0	1	0
0	0	1
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
0	0	1
1	0	0

數據資料_特徵縮放

特徵縮放

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

特徵縮放

Standardisation	Normalisation
$x_{stand} = \dfrac{x - mean(x)}{StandardDeviation(x)}$	$x_{norm} = \dfrac{x - min(x)}{max(x) - min(x)}$
標準化：平均=0, 標準差=1	正規化：0~1

數據資料_訓練集、測試集

訓練/測試

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

感謝聆聽