知識挖掘與資料工程導論期末報告

系級:資訊系 104 級
組別:15 組
組員:林蔚廷 F74001234、陳昱成 F74002256

1. 問題描述

本題目提供的 training data (train.csv)為 42000 張 28 pixel*28 pixel 灰階圖,圖片內容為手寫 1~9 的數字，每張圖有 784 筆灰階資料(0~255)，label 欄位代表每張圖上面的數字，pixel0~pixel783 欄位代表每個 pixel 的灰階。

如下圖

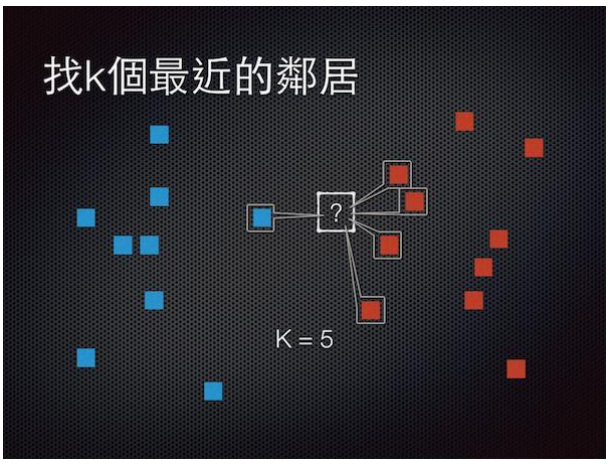
	A	B	C	D	E	F	G	H	I	J	K	L
1	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	pixel9	pixel10
2	1	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0
5	4	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	7	0	0	0	0	0	0	0	0	0	0	0
9	3	0	0	0	0	0	0	0	0	0	0	0
10	5	0	0	0	0	0	0	0	0	0	0	0
11	3	0	0	0	0	0	0	0	0	0	0	0
12	8	0	0	0	0	0	0	0	0	0	0	0
13	9	0	0	0	0	0	0	0	0	0	0	0
14	1	0	0	0	0	0	0	0	0	0	0	0
15	3	0	0	0	0	0	0	0	0	0	0	0

題目要求我們利用他們提供的 training data 去分辨 test.csv 的 28000 張圖上個別的數字，如下圖。

	A	B	C	D	E	F	G	H	I	J	K	L
1	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	pixel9	pixel10	pixel11
2	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0

2. 預測方法

(1) K Nearest Neighbor



我們將每組 test.csv 的資料對 train.csv 做距離的計算，然後取前 K 筆最小的，再看哪個 label 最多，就決定其數字。

(2) Jaccard Distance

在 KNN 裡面需要計算距離，我們計算距離的方式是用 Jaccard Distance，但由於 Jaccard Distance 只適用在二元的 data，因此我們先對 train.csv 和 test.csv 作事前資料處裡，由於他們 pixel 的數據都是 0~255，於是我們把大於 100 的數據都設成 1，小於則為 0，然後就可以做 Jaccard Distance 的計算了。

3. 成果展示

(1) K = 1 ， Score = **0.96829**

(2) K = 3 ， Score = **0.96857**

(3) K = 10， Score = **0.96557**