HW2 MNIST

404410030 資工三 鄭光宇

環境設置：

使用 python 的 **sklearn** 套件，裡面有的 **SVM**, PCA, t-SNE 等工具完成分類問題，使用 matplotlib 來繪製圖表。

資料集：

使用這次作業指定的 **MNIST** 和最近流行的 **Fashion Mnist**。

MNIST：

28x28 大小的手寫數字，共有 10 種數字（0~9）。每筆資料先是他對應的 label（0~9），之後是 784 維的向量，以 ” index:value” 的方式表示。

Fashion Mnist：

來源：

github: https://github.com/zalandoresearch/fashion-mnist

鑑於 MNIST 手寫分類問題對於現代機器學習模型不夠難，所以有人發展出了這個 Fashion Mnist，如果是使用 SVM，最好的 benchmark 在測試集上，大約也只有 89%左右的 accuracy。

28x28 大小的灰階服飾圖片，共有 10 種服飾。每筆資料先是對應的 label（0~9），然後是 784 維的向量。

（如果是使用 python 可以用它 github 上提供的工具讀資料）

以下是Fashion Mnist的Label意義：

LABEL	DESCRIPTION
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

實驗結果：

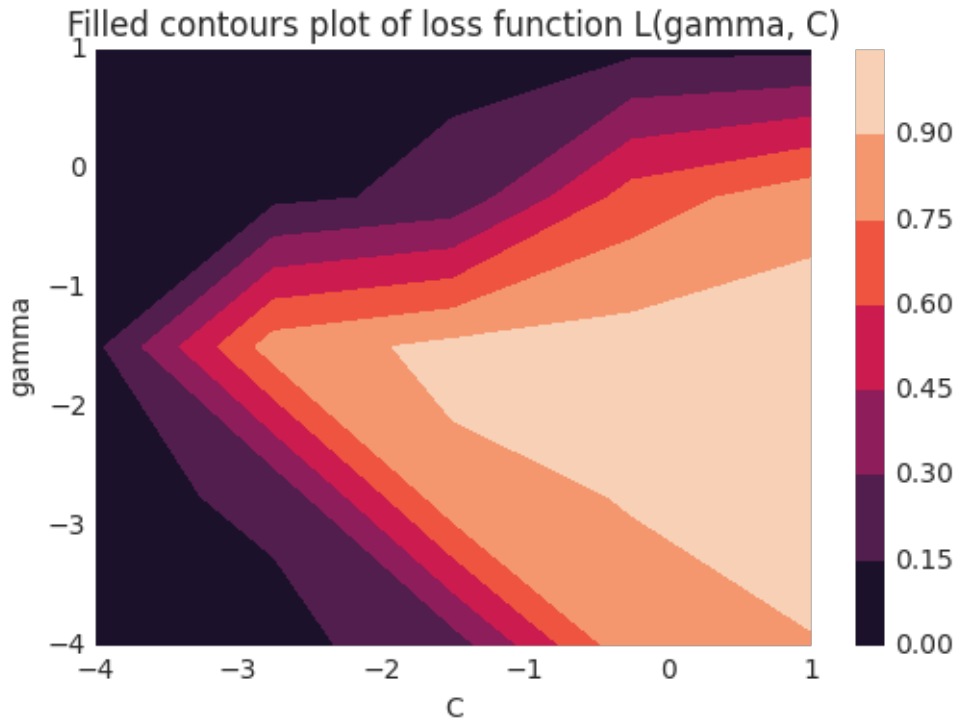
以下兩個實驗結果，SVM使用的kernel均為RBF (radial basis function)，因為不知道什麼原因，Polynomial kernel的SVM執行時間很長，所以沒有實驗使用Polynomial kernel的結果。

MNIST：

為了避免高維詛咒，把資料送進SVM前，我先將圖片數值scale到[0,1]的浮點數，之後用PCA將維度降至 20維，保留約 64% 的資訊量，經過一些測試，雖然只有 64% 資訊量，但SVM在此問題上仍然可以表現優秀。

為了找出較好的SVM超參數（C, gamma），以log尺度、5種C、5種gamma，共25種參數以3-fold Cross-Validation取平均accuracy做grid-search，看看是否能找到好的參數。

結果如下：



（註記：兩個座標都是 log scale）

找到最好的參數在

$$C=10^{1.0},\ gamma=\ 10^{-1.5}$$

在這樣的參數下，training set 上的 10-fold Cross-Validation 可以達到 98±1% 左右，其中±1%的意思是，這十組測試 accuracy 的兩倍標準差為 ±1%。

訓練集上測試得 10-fold Cross-Validation 如下：

10-fold cross-validation
0.983516483516
0.980176578377
0.978003666056
0.981666666667
0.9755
0.978996499417
0.976829471579
0.978996499417
0.976821744205
0.984823215477
Accuracy: 0.98 (+/- 0.01)

之後，在測試集上面驗證效果。

測試集上的 Confusion Matrix 如下：

	0	1	2	3	4	5	6	7	8	9
0	973	0	1	1	0	2	1	1	1	0
1	0	1132	1	1	0	0	0	0	1	0
2	3	0	1012	4	1	1	0	7	4	0
3	0	1	2	994	0	4	0	4	4	1
4	0	0	1	0	965	0	4	0	1	11
5	2	0	0	8	1	870	3	1	5	2
6	5	4	1	0	4	4	938	1	1	0
7	0	5	11	2	2	0	0	997	0	11
8	2	0	3	4	3	5	2	2	951	2
9	2	4	0	5	11	5	0	8	1	973

對於每一個類別，效能評估基準如下表：

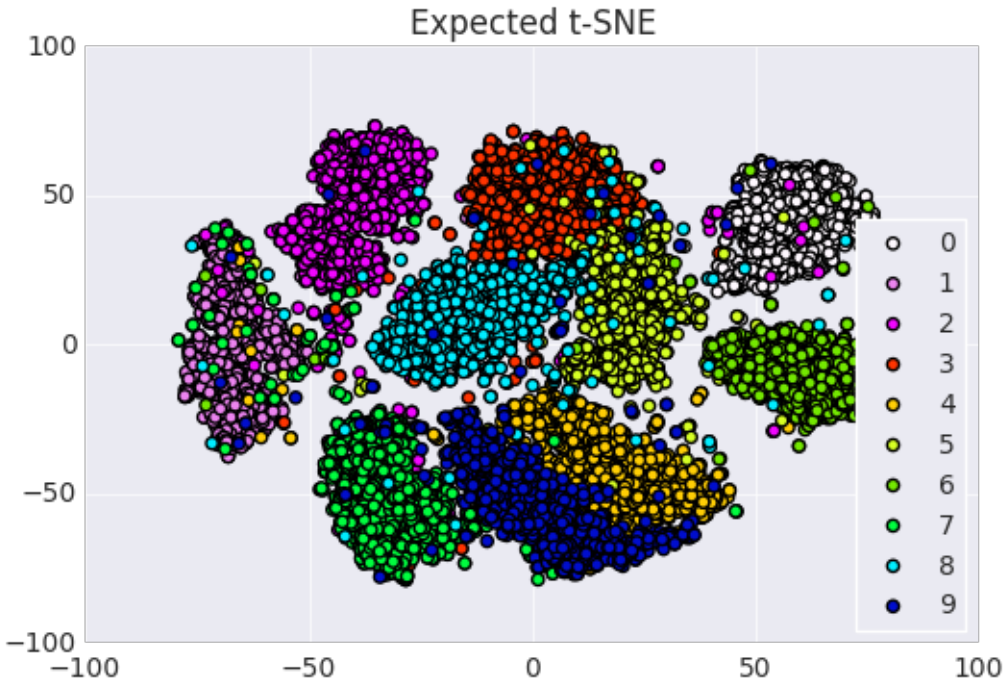
	precision	recall	F1-score	support
0	0.99	0.99	0.99	980
1	0.99	1.00	0.99	1135
2	0.98	0.98	0.98	1032
3	0.98	0.98	0.98	1010
4	0.98	0.98	0.98	982
5	0.98	0.98	0.98	892
6	0.99	0.98	0.98	958
7	0.98	0.97	0.97	1028
8	0.98	0.98	0.98	974
9	0.97	0.96	0.97	1009
avg/total	0.98	0.98	0.98	10000

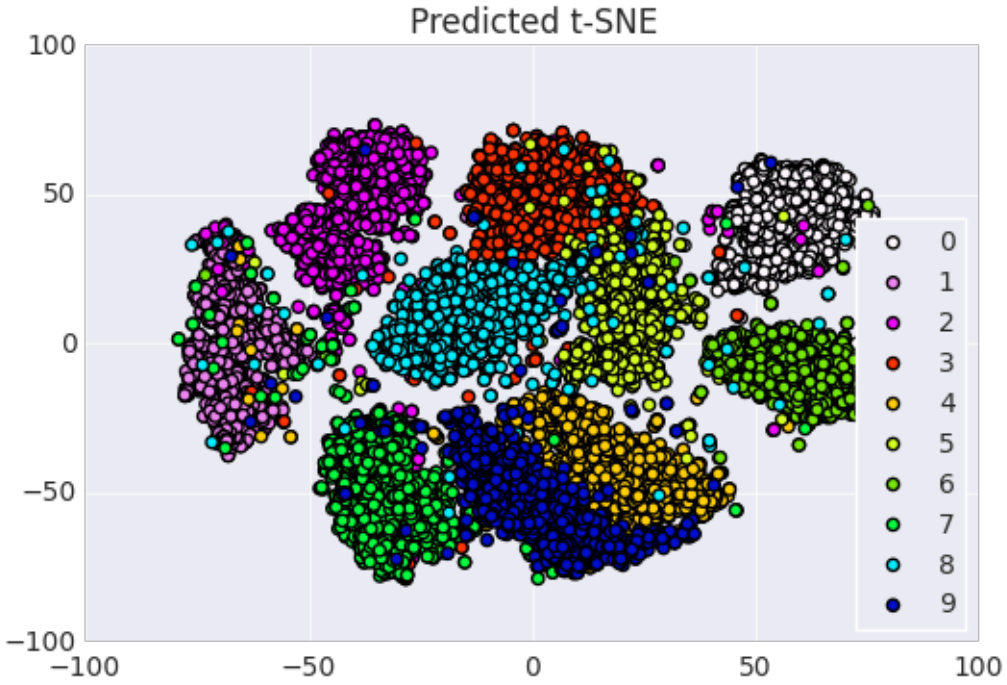
⇨ 總結測試測試集的 Accuracy：0.9805

可以看出效果還不錯。

資料可視化：

使用最近流行的 t-SNE 將資料降維、投影到 2D 平面上，使資料可視化。測試集上可視化結果如下：





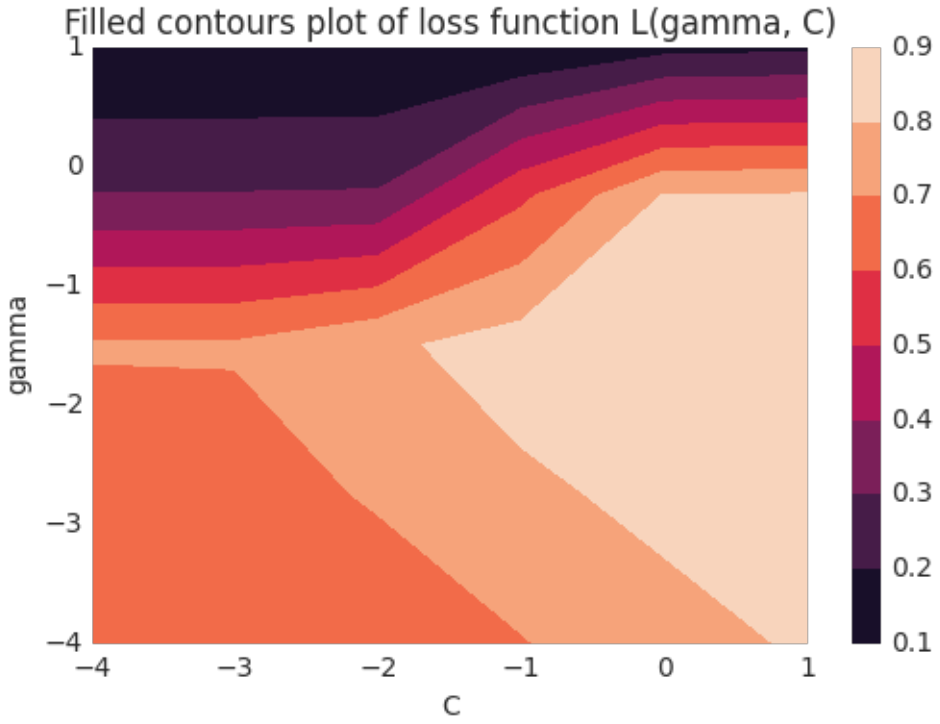
第二張圖片是 SVM 預測的結果，可以看出預測結果與 Ground Truth 很接近，SVM 很好地切開了各個類別。

Fashion Mnist：

為了避免高維詛咒，把資料送進 SVM 前，我先將圖片數值 scale 到[0,1]的浮點數，之後用 PCA 將維度降至 26 維，保留約 81% 的資訊量。

以 log 尺度、6 種 C、5 種 gamma，共 30 種參數以 3-fold Cross-Validation 取平均 accuracy 做 grid-search。

結果：



最好的參數：

$$C=10^{1.0},\ gamma=10^{-1.5}$$

與在 MNIST 資料集上相同。

10-fold Cross-Validation：

10-fold Cross-Validation
0.8925
0.892166666667
0.884166666667
0.891833333333
0.893166666667
0.889166666667
0.893166666667
0.887166666667
0.890833333333
0.882333333333
Accuracy: 0.89 (+/- 0.01)

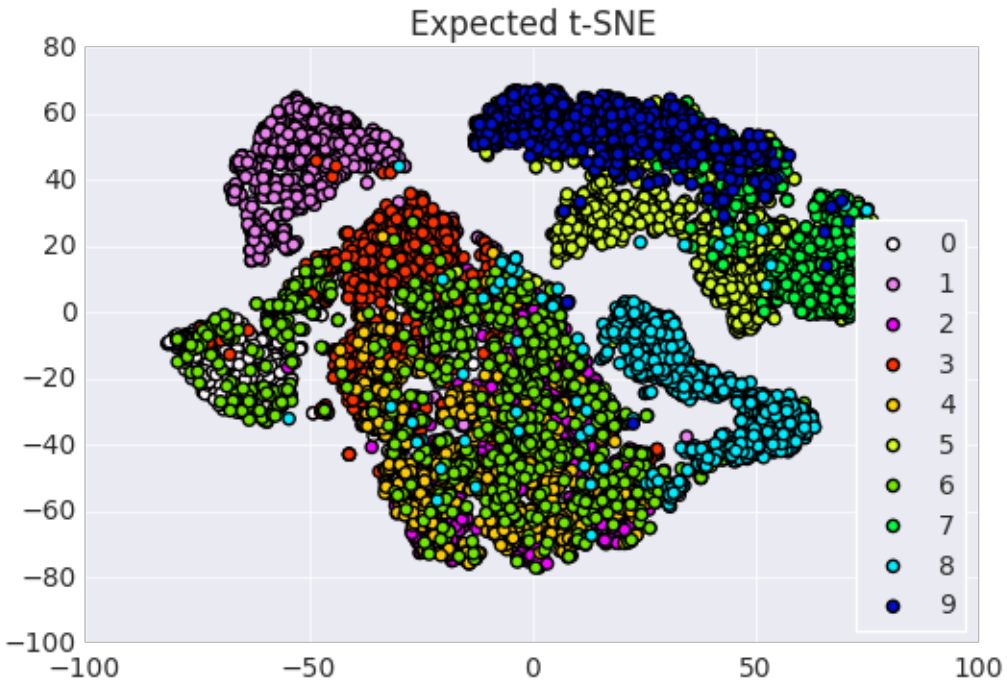
測試集上的 Confusion Matrix：

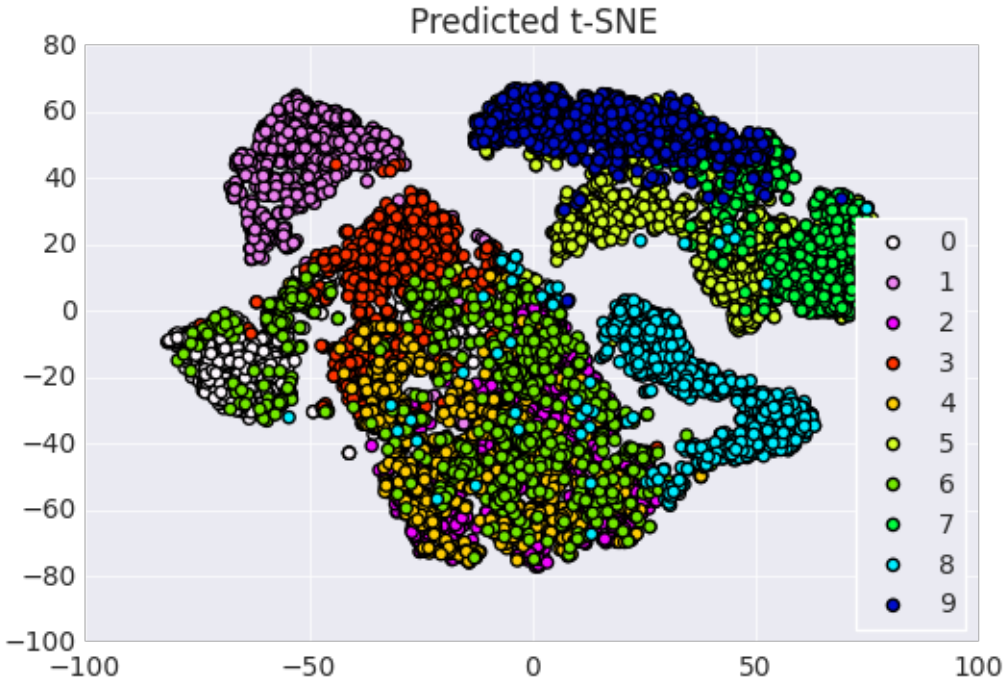
	0	1	2	3	4	5	6	7	8	9
0	852	1	15	23	4	1	96	0	8	0
1	6	970	1	16	4	0	3	0	0	0
2	16	0	805	11	88	0	78	0	2	0
3	24	9	13	892	36	0	23	0	3	0
4	0	0	89	25	825	0	59	0	2	0
5	0	0	0	1	0	939	0	43	1	16
6	132	0	89	28	75	0	662	0	14	0
7	0	0	0	0	0	21	0	949	0	30
8	5	0	3	3	3	3	2	4	976	1
9	0	0	0	0	0	13	1	39	0	947

測試集上的效能：

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.82	0.85	0.84	1000
1	0.99	0.97	0.98	1000
2	0.79	0.81	0.80	1000
3	0.89	0.89	0.89	1000
4	0.80	0.82	0.81	1000
5	0.96	0.94	0.95	1000
6	0.72	0.66	0.69	1000
7	0.92	0.95	0.93	1000
8	0.97	0.98	0.97	1000
9	0.95	0.95	0.95	1000
AVG/TOTAL	0.88	0.88	0.88	10000

使用 t-SNE 降維、可視化：





可以看出，Fashion Mnist 的資料降維後，看起來非常的複雜、難分，也許是因為資料真的比較複雜，SVM 在這個資料集上面的表現，不如 MNIST 上好，也許使用最近的 CNN 會得到比較好的結果？

結論：

這次使用完 SVM 後，了解到它是個優秀的算法，在 MNIST 上能夠輕鬆得到 97~98%的正確率，但也了解到它也有一些限制，例如在新的資料集上表現地較不理想。

參考資料：

我的 github:

https://github.com/peter0749/Multimedia-Content-Analysis

Fashion Mnist：

https://github.com/zalandoresearch/fashion-mnist

bayesian-optimization

https://github.com/thuijskens/bayesian-optimization