Đồ án vấn đáp môn Học máy

Phân lớp ảnh chữ số viết tay bằng SVM

1) Thông tin chung:

1.1 Thành viên:

MSSV	Họ và tên	Email
1712003	Lê Nguyễn Đức Anh	1712003@student.hcmus.edu.vn
1712096	Đặng Hồng Minh	1712096@student.hcmus.edu.vn

1.2 Kế hoạch:

- 19/8/2020 – 20/8/2020: Tìm hiểu về thuật toán SVM. xem lý thuyết.
- 20/8/2020 – 21/8/2020: Tìm hiểu thư viện scikit-learn, cách sử dụng các hàm có sẵn.
- 22/8/2020 – 26/8/2020: Tiến hành train và test cho các trường hợp tập dữ liệu.
- 25/8/2020 – 26/8/2020: Ghi lại kết quả, báo cáo.

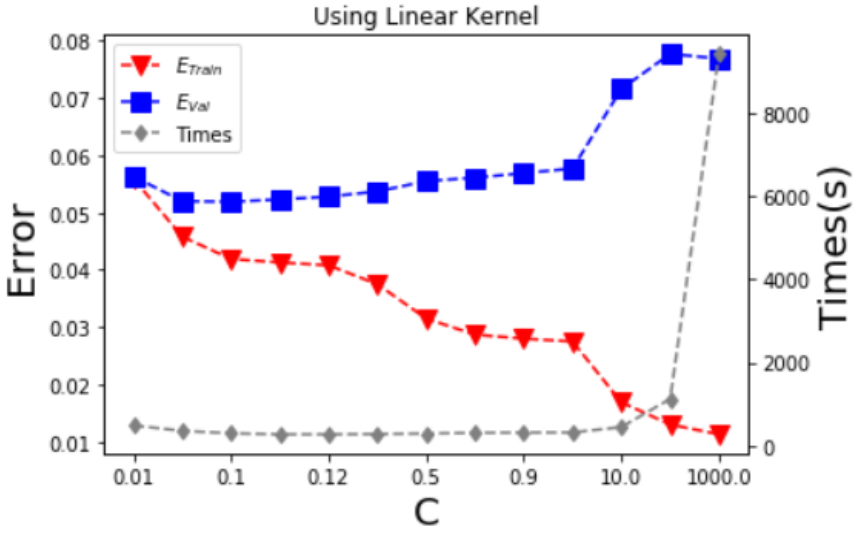
2) Huấn luyện SVM để phân lớp ảnh chữ số viết tay:

- **2.1. Dùng linear kernel:**

C	Etrain	Eval	Train Score	Validation Score	Times
0.01	0.0559	0.0563	0.94406	0.9437	8min8s
0.05	0.0458	0.052	0.95424	0.948	5min59s
0.1	0.0419	0.0519	0.95812	0.9481	4min59s
0.11	0.0413	0.0523	0.95866	0.9477	4min35s
0.12	0.0408	0.0528	0.95918	0.9472	4min35s
0.2	0.0375	0.0537	0.96248	0.9463	4min42s
0.5	0.0315	0.0555	0.96854	0.9445	4min54s
0.8	0.0287	0.0561	0.97132	0.9439	5min10s
0.9	0.028	0.0569	0.97204	0.9431	5min17s
1	0.0275	0.0577	0.97246	0.9423	5min19s
10	0.0169	0.0716	0.98308	0.9284	7min30s
100	0.013	0.0777	0.987	0.9223	18min46s
1000	0.0113	0.0769	0.98866	0.9231	2h37min6s

Bảng 1. Độ lỗi trên tập training, độ lỗi trên tập validation, thời gian huấn luyện

Biểu diễn đồ thị



Độ lỗi trên tập Train và Validation khi C thay đổi

Nhận xét:

Theo lý thuyết giá trị C ảnh hưởng đến quá trình học.

- C lớn sẽ dẫn đến trường hợp: bias thấp, variance cao → overfitting.

Bảng kết quả và biểu đồ cho thấy khi C tăng thì Etrain giảm (tốt cho tập huấn luyện), tuy nhiên Eval tăng → overfitting.

- C nhỏ sẽ dẫn đến trường hợp: bias cao, variance thấp → underfitting.

. Bảng kết quả cho thấy C quá nhỏ thì Eval có độ lỗi lớn → underfitting. Thậm chí là không tốt cho tập train (Etrain cũng có độ lỗi rất lớn).

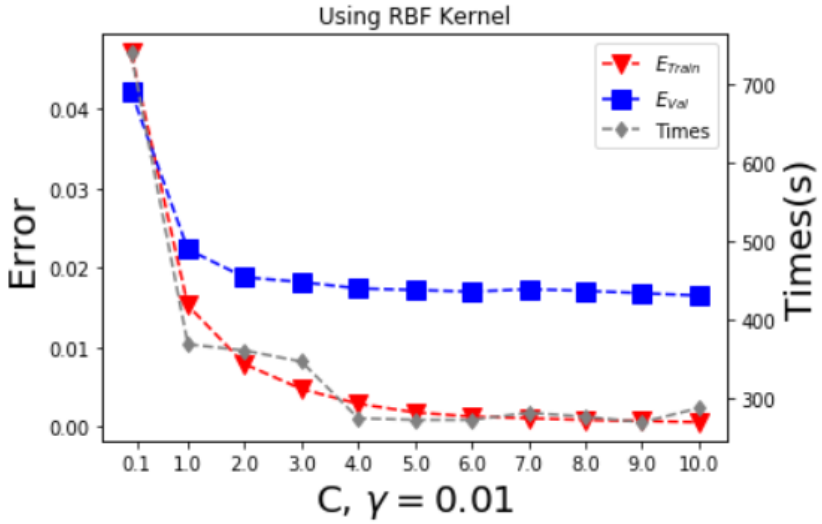
.⇒ Vậy dựa vào bảng kết quả ta chọn hàm dự đoán cuối cùng với Eval nhỏ nhất là C = 0.1

- **2.1. Dùng Gaussian/RBF kernel**

C	γ	Etrain	Eval	Train Score	Validation Score	Times
0.1	0.001	0.09824	0.0861	0.90176	0.9139	26min5s
	0.01	0.04702	0.0422	0.95298	0.9578	12min19s
	0.1	0.28952	0.3125	0.71048	0.6875	1h10min35s
1	0.001	6.422e-02	0.0589	0.93578	0.9411	10min5s
	0.01	1.526e-02	0.0223	0.98474	0.9777	6min9s
	0.1	4.000e-05	0.448	0.99996	0.9552	2h4min42s
2	0.01	0.00788	0.0188	0.99212	0.9812	6min1s
3	0.01	0.00476	0.0182	0.99524	0.9818	5min47s
4	0.01	0.0029	0.0174	0.9971	0.9826	4min35s
5	0.01	0.00182	0.0172	0.99818	0.9828	4min33s
6	0.01	0.00128	0.017	0.99872	0.983	4min33s
7	0.01	0.00108	0.0173	0.99892	0.9827	4min42s
8	0.01	0.00086	0.0171	0.99914	0.9829	4min37s
9	0.01	0.00074	0.0168	0.99926	0.9832	4min30s
10	0.001	0.03794	0.0408	0.96206	0.9592	5min29s
	0.01	0.00058	0.0165	0.99942	0.9835	4min48s
	0.1	0	0.0434	1	0.9566	1h54min7s

Bảng 2. Độ lỗi trên tập training, độ lỗi trên tập validation, thời gian huấn luyện

Biểu diễn đồ thị



Độ lỗi trên tập Train và Validation khi C thay đổi, $\gamma = 0.01$

Nhận xét:

- C, γ lớn sẽ dẫn đến trường hợp: bias thấp, variance cao → overfitting.

  Bảng kết quả và biểu đồ cho thấy khi C và γ càng lớn thì Etrain nhỏ và Eval lớn → overfitting. Cho kết quả rất tốt cho tập huấn luyện nhưng không phản ánh được tính tổng quát (không mô tả được cho các dữ liệu chưa gặp).

- C, γ nhỏ sẽ dẫn đến trường hợp: bias cao, variance thấp → underfitting.

  Bảng kết quả và biểu đồ cho thấy khi C và γ càng nhỏ thì Etrain lớn và Eval lớn → underfitting, không mô tả được dữ liệu.

.⇒ Vậy dựa vào bảng kết quả ta chọn hàm dự đoán cuối cùng với Eval nhỏ nhất là C = 10, $\gamma = 0.01$

3) Đánh giá SVM

Với Linear Kernel, C = 0.01, có kết quả:

Training score	0.95812
Training error	0.04188
Testing score	0.9463
Testing error	0.0537
Times	4min 26s

Với RBF Kernel, C = 10, $\gamma = 0.01$, có kết quả:

Training score	0.99942
Training error	0.00058
Testing score	0.982
Testing error	0.018
Times	4min 44s