

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.

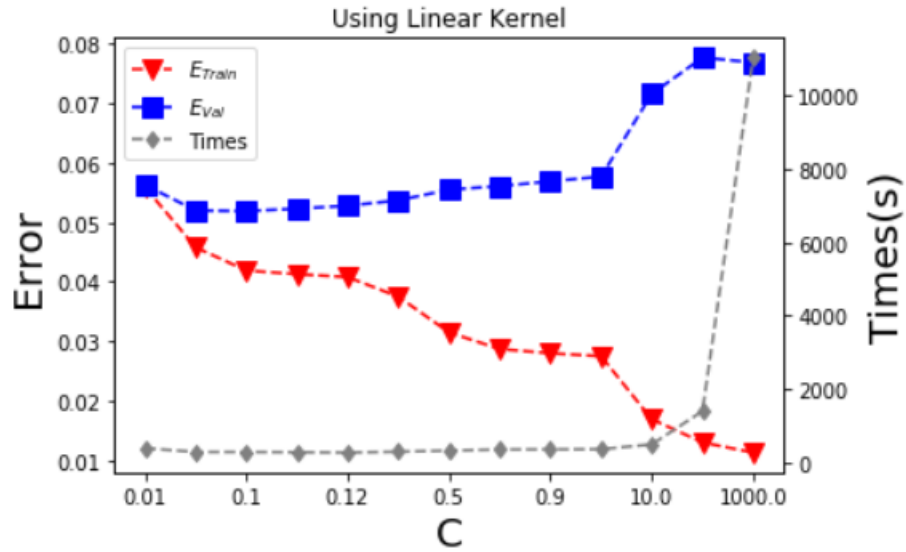
2) Huấn luyện SVM để phân lớp ảnh chữ số viết tay:

• 2.1. Dùng linear kernel:

C	Etrain	Eval	Train Score	Validation Score	Times
0.01	0.0559	0.0563	0.94406	0.9437	6min29s
0.05	0.0458	0.052	0.95424	0.948	4min46s
0.1	0.0419	0.0519	0.95812	0.9481	4min46s
0.11	0.0413	0.0523	0.95866	0.9477	4min39s
0.12	0.0408	0.0528	0.95918	0.9472	4min30s
0.2	0.0375	0.0537	0.96248	0.9463	4min53s
0.5	0.0315	0.0555	0.96854	0.9445	5min23s
0.8	0.0287	0.0561	0.97132	0.9439	5min54s
0.9	0.028	0.0569	0.97204	0.9431	5min56s
1	0.0275	0.0577	0.97246	0.9423	6min1s
10	0.0169	0.0716	0.98308	0.9284	8min12s
100	0.013	0.0777	0.987	0.9223	23min20s
1000	0.0113	0.0769	0.98866	0.9231	3h4min5s

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 \rightarrow overfitting.

Bảng kết quả và biểu đồ cho thấy khi C tăng thì E_{train} giảm (tốt cho tập huấn luyện), tuy nhiên E_{val} tăng \rightarrow overfitting.

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

. Bảng kết quả cho thấy C quá nhỏ thì E_{val} có độ lỗi lớn \rightarrow underfitting. Thậm chí là không tốt cho tập train (E_{train} cũng có độ lỗi rất lớn).

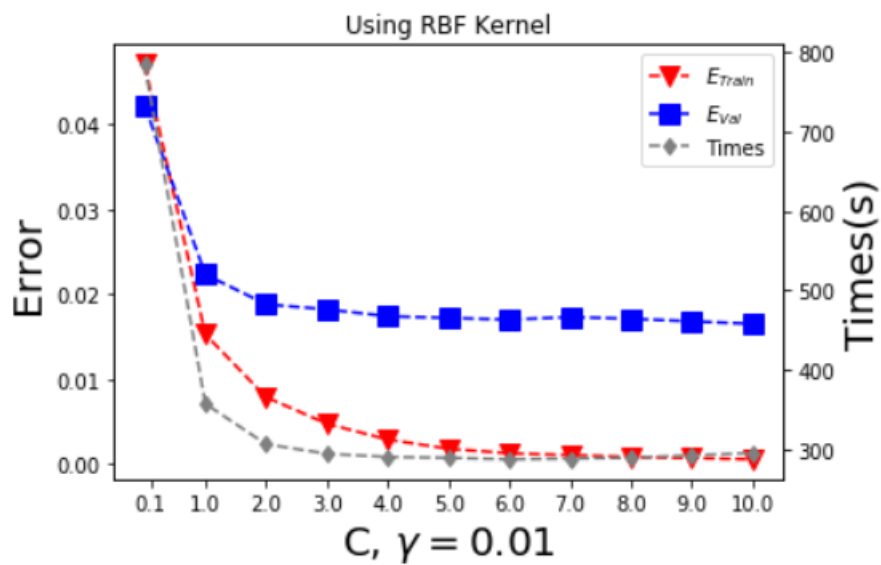
\Rightarrow Vậy dựa vào bảng kết quả ta chọn hàm dự đoán cuối cùng với E_{val} nhỏ nhất là $C = 0.1$

• 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	27min33s
	0.01	0.04702	0.0422	0.95298	0.9578	13min5s
	0.1	0.28952	0.3125	0.71048	0.6875	1h8min31s
1	0.001	6.422e-02	0.0589	0.93578	0.9411	10min56s
	0.01	1.526e-02	0.0223	0.98474	0.9777	5min57s
	0.1	4.000e-05	0.448	0.99996	0.9552	1h46min21s
2	0.01	0.00788	0.0188	0.99212	0.9812	5min6s
3	0.01	0.00476	0.0182	0.99524	0.9818	4min54s
4	0.01	0.0029	0.0174	0.9971	0.9826	4min50s
5	0.01	0.00182	0.0172	0.99818	0.9828	4min49s
6	0.01	0.00128	0.017	0.99872	0.983	4min47s
7	0.01	0.00108	0.0173	0.99892	0.9827	4min48s
8	0.01	0.00086	0.0171	0.99914	0.9829	4min49s
9	0.01	0.00074	0.0168	0.99926	0.9832	4min52s
10	0.001	0.03794	0.0408	0.96206	0.9592	5min51s
	0.01	0.00058	0.0165	0.99942	0.9835	4min55s
	0.1	0	0.0434	1	0.9566	1h49min12s

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 \rightarrow overfitting.

Bảng kết quả và biểu đồ cho thấy khi C và γ càng lớn thì Etrain nhỏ và Eval lớn \rightarrow 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 \rightarrow underfitting.

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

\Rightarrow 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 47s

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 59s