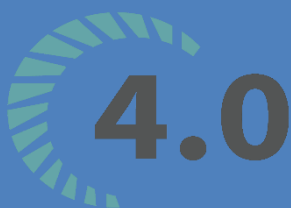


BỘ MÔN KHOA HỌC MÁY TÍNH – KHOA CÔNG NGHỆ THÔNG TIN
ĐẠI HỌC KHOA HỌC TỰ NHIÊN THÀNH PHỐ HỒ CHÍ MINH, ĐẠI HỌC QUỐC GIA TP HCM

NHẬP MÔN HỌC MÁY



Sinh viên thực hiện: 1712778 - Thống A Thảo
1712898 - Trần Việt Văn

GV phụ trách: Trần Trung Kiên

ĐỒ ÁN/BÀI TẬP MÔN HỌC - NHẬP MÔN HỌC MÁY
HỌC KỲ I – NĂM HỌC 2019-2020



BẢNG THÔNG TIN CHI TIẾT NHÓM

Mã nhóm:	53	
Số lượng:	2	
MSSV	Họ tên	Email
1712778	Thống A Thảo	Thaonthong9917@gmail.com
1712898	Trần Việt Văn	Van.iu.snsd@gmail.com

Bảng phân công & đánh giá hoàn thành công việc, quá trình thực hiện				
Công việc thực hiện	Thời gian	Người thực hiện	Mức độ hoàn thành	Đánh giá của nhóm
Tìm kiếm tài liệu, đọc lại cơ sở lí thuyết	11/8 – 15/8	1712898 – Trần Việt Văn	90%	9/10
Phân tích, code các model linear kernel, train và đánh giá kết quả	15/8 – 21/8	1712778 – Thống A Thảo	90%	9/10
Phân tích, code các model rbf kernel, train và đánh giá kết quả	15/8 – 21/8	1712898 – Trần Việt Văn	90%	9/10
Debug, xem lại kết quả, chọn phương pháp tối ưu, retrain model, chạy thực nghiệm	21/8 – 25/8	1712898 – Trần Việt Văn 1712778 – Thống A Thảo	80%	8/10
Làm report	25/8 – 26/8	1712778 – Thống A Thảo	90%	9/10

YÊU CẦU ĐỒ ÁN- BÀI TẬP

Loại bài tập	<input checked="" type="checkbox"/> Lý thuyết <input type="checkbox"/> Thực hành <input checked="" type="checkbox"/> Đồ án <input type="checkbox"/> Bài tập
Ngày bắt đầu	11/8/2020
Ngày kết thúc	25/8/2020

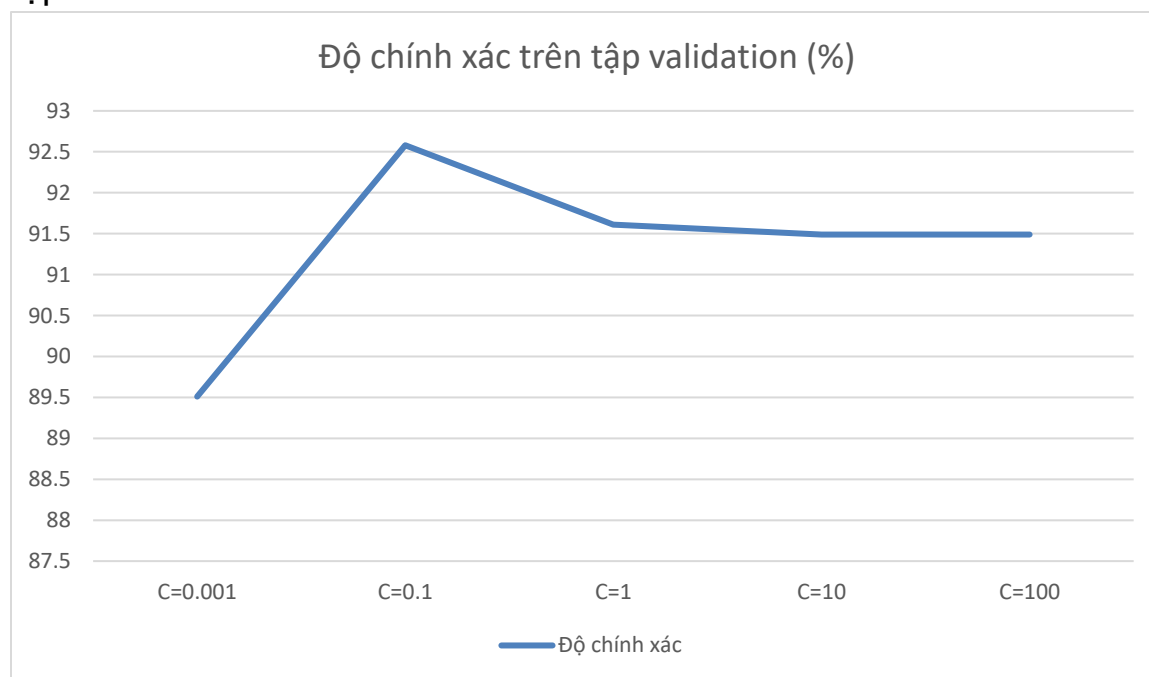
A. Yêu cầu của Đồ án/Bài tập

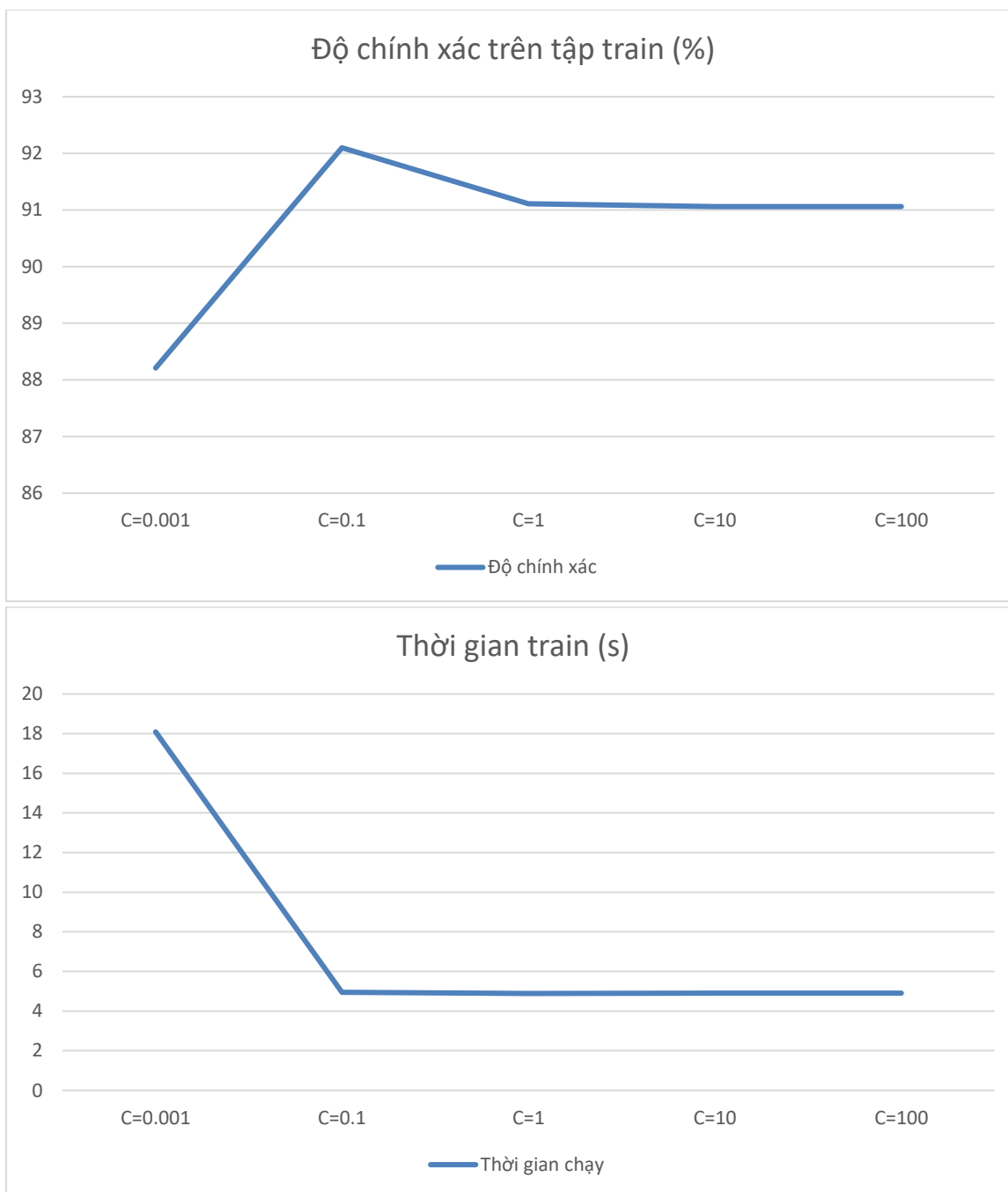
Tạo model nhận dạng chữ số viết tay từ bộ dataset có sẵn, sử dụng svm.

B. Kết quả thực nghiệm

I. Linear kernel

Biểu đồ giá trị chính xác (đặc trưng độ lỗi), và bảng thời gian chạy khi dùng linear model với các C khác nhau trên 5000 mẫu đầu của tập test





Các kết quả phù hợp với phần lí thuyết (thời gian thực nghiệm chạy tất cả 5 mô hình là gần 1 tiếng, nhưng lại hiển thị ít giây). Khi C quá nhỏ, mô hình sẽ có khoảng phân tách nhỏ, độ chính xác thấp. Nếu C quá lớn, mô hình sẽ bị lỗi trên 1 số điểm dữ liệu. Vì vậy C gần nhất ở đây là 1. Độ chính xác trên tập Validation là 91.61%

II. RBF kernel

Bảng giá trị chính xác (đặc trưng độ lỗi), và bảng thời gian chạy khi dùng rbf model với các C và gamma khác nhau trên 5000 mẫu đầu của tập test

Gamma	C				
	0.001	0.1	1	10	100
0.001	11.356%	73.764%	89.788%	92.514%	92.946%
0.1	11.356%	25.212%	83.914%	84.786%	84.786%
1	11.356%	11.356%	24.816%	25.102%	25.102%
10	11.356%	11.356%	20.23%	20.23%	20.23%
100	11.356%	11.356%	20.23%	20.23%	20.23%

Bảng 1: Độ chính xác trên tập train

Gamma	C				
	0.001	0.1	1	10	100
0.001	10.64%	76.49%	91.1%	93.26%	93.23%
0.1	10.64%	25.49%	85.17%	86.05%	86.05%
1	10.64%	10.64%	17.59%	17.85%	17.85%
10	10.64%	10.64%	10.64%	10.64%	10.64%
100	10.64%	10.64%	10.64%	10.64%	10.64%

Bảng 2: Độ chính xác trên tập validation

Gamma	C				
	0.001	0.1	1	10	100
0.001	45.86	36.47	14.09	6.60	5.28
0.1	45.78	43.68	46.28	45.5	46.76
1	44.51	46.73	47.85	49.90	49.23
10	50.68	48.26	49.76	49.29	50.76
100	49.81	48.60	49.54	50.42	49.75

Bảng 3: Thời gian train (s)

Khi thay đổi giá trị C và gamma khác nhau ta thu được độ lỗi khác nhau trên tập train giá trị gamma tăng thì độ lỗi tăng, tương tự với C, ta chọn C và gamma phù hợp để độ lỗi trên tập Validation nhỏ

nhất (độ chính xác cao nhất). Ở đây ta thấy với $C=10$ và γ bằng 0.001 thì độ chính xác đạt được là 93.26% nên được chọn

III. Retrain và chạy trên tập test

Sau khi thử các mô hình, ta chọn được rbf kernel với $C=10$ và $\gamma=0.001$

Tiến hành chạy trên tập test ta thu được kết quả:

- Độ chính xác: 95.49%
- Thời gian train: khoảng 300s

IV. Ghi chú

-Việc chọn model được thực hiện trên 5000 mẫu đầu của tập train để giảm bớt thời gian chạy.

-Kết quả có thể không chính xác vì có thể do các tác nhân (máy tính, chọn C và γ không phù hợp).

-Kết quả đã được thực nghiệm, có thể dùng model trong live predict để ứng dụng kiểm chứng.