NHẬP MÔN HỌC MÁY

Giảng viên: Trần Trung Kiên

BÁO CÁO CUỐI KÌ

PHÂN LỚP ẢNH CHỮ SỐ VIẾT TAY BẰNG SVM

1712616 - Đinh Văn Ngọc
1712615 - Nguyễn Trọng Nghĩa

Khoa Công nghệ Thông tin
Đại học Khoa học Tự nhiên
TP HCM Tháng 8/2019.

1. Kế hoạch thực hiện:

STT	Công việc	Mức độ hoàn thành	Người thực hiện	Kết quả
1	Tìm hiểu và tổng hợp lý thuyết mô hình SVM	100%	Đinh Văn Ngọc và Nguyễn Trọng Nghĩa	Nắm được kiến thức và cách hoạt động của SVM
2	Chuẩn bị đọc dữ liệu vận hành trên colab	100%	Nguyễn Trọng Nghĩa	
3	Thực hiện huấn luyện dữ liệu với kernel linear	100%	Đinh Văn Ngọc	Lựa chọn các bộ tham số phù hợp và chọn được tham số được xem là tốt nhất cho bài toán.
4	Thực hiện huấn luyện dữ liệu với kernel rbf	100%	Nguyễn Trọng Nghĩa	
5	Trực quan hóa và đánh giá kết quả	100%	Nguyễn Trọng Nghĩa và Đinh Văn Ngọc	Có thể quan sát rõ được sự thay đổi kết quả của mô hình với các tham số khác nhau.
6	Viết báo cáo	100%	Đinh Văn Ngọc và Nguyễn Trọng Nghĩa	

2. Huấn luyện SVM:

Bài toán được thực trên môi trường google colab.
Quan sát phân bố nhãn của tập train.

Kernel Linear:

- Thử nghiệm với các giá trị C từ 0+ đến C vô cùng (Để thu hẹp vùng giá trị tốt của C). Số sau gấp 10 lần số trước bắt đầu từ 0.001.
  C = [0.001,0.01,0.1,1,10,100,1000,10000]

Kết quả: chỉ huấn luyện đến 100, C = 1000, 10000 mô hình chạy quá lâu hơn 5 tiếng không thu được kết quả.

- Thử nghiệm với tham số C từ 0.1 đến 1 bước nhảy 0.1 và từ 0.01 đến 0.1 bước nhảy là 0.01
  C = [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9] và
  C = [0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.09]

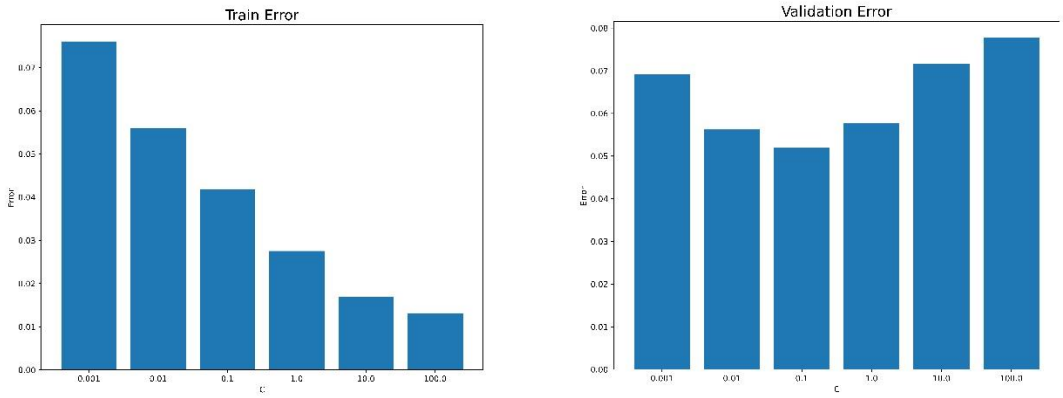
Kernel Rbf:

- Thử nghiệm với một vài giá trị gamma [0.001,0.0001,0.01,0.1,0.5,1] nhằm xác đinh khoảng gamma cho giá trị tốt và thời gian phù hợp và C =1.
  Kết quả: với gamma = 0.5 và 1 thì mô hình chạy rất lâu, thu được kết quả có độ lỗi lớn.

- Thử nghiệm với bộ tham số gamma trong khoảng (0.001-0.1) và bộ tham số C (0.01- 1000).
  Gamma = [0.001,0.005,0.01,0.05,0.1]
  C = [0.01,0.1,1,10,100,1000]

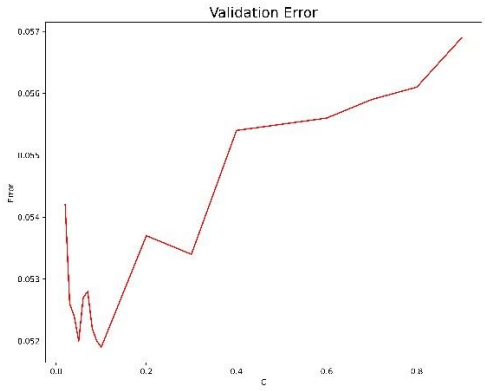
3. Đánh giá kết quả huấn luyện:

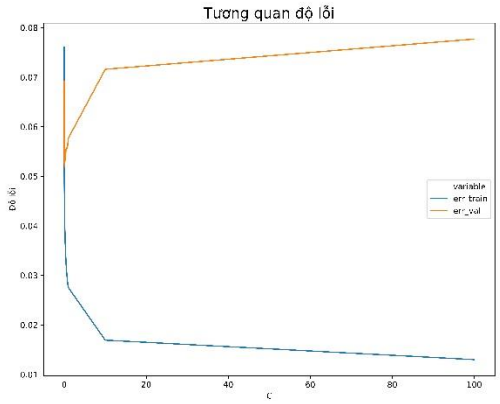
Mô hình kernel linear:



Nhận xét:

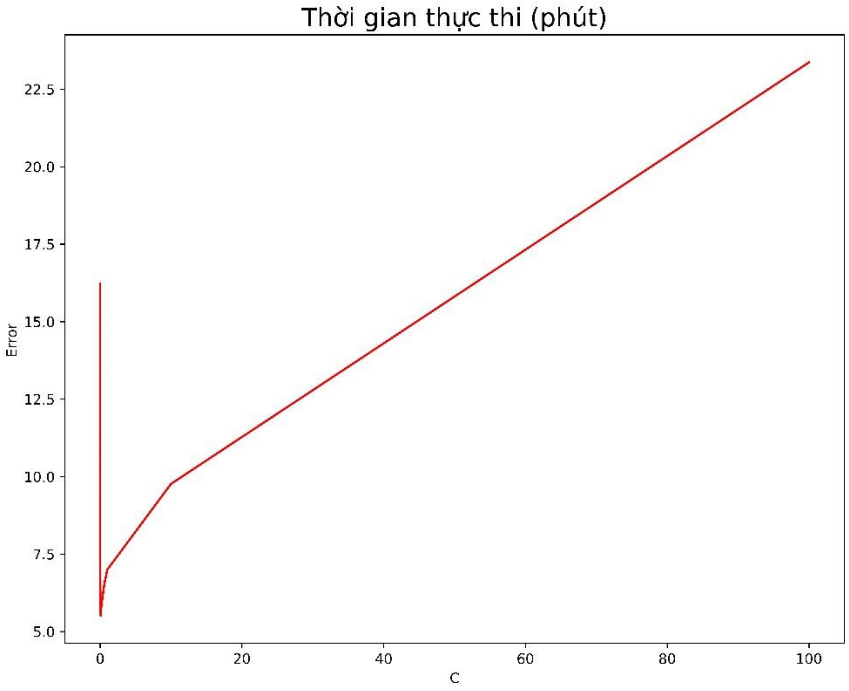
- Độ lỗi tập train giảm dần cho thấy sự phù hợp với ý nghĩa của tham số C.
- C càng lớn thì độ lỗi trên tập train càng nhỏ và C càng nhỏ thì độ lỗi càng lớn.
- Độ lỗi tập validation giảm rồi tăng cho ta thấy khi C = 0.1 trở lên thì mô hình bị overfitting.





Nhận xét:

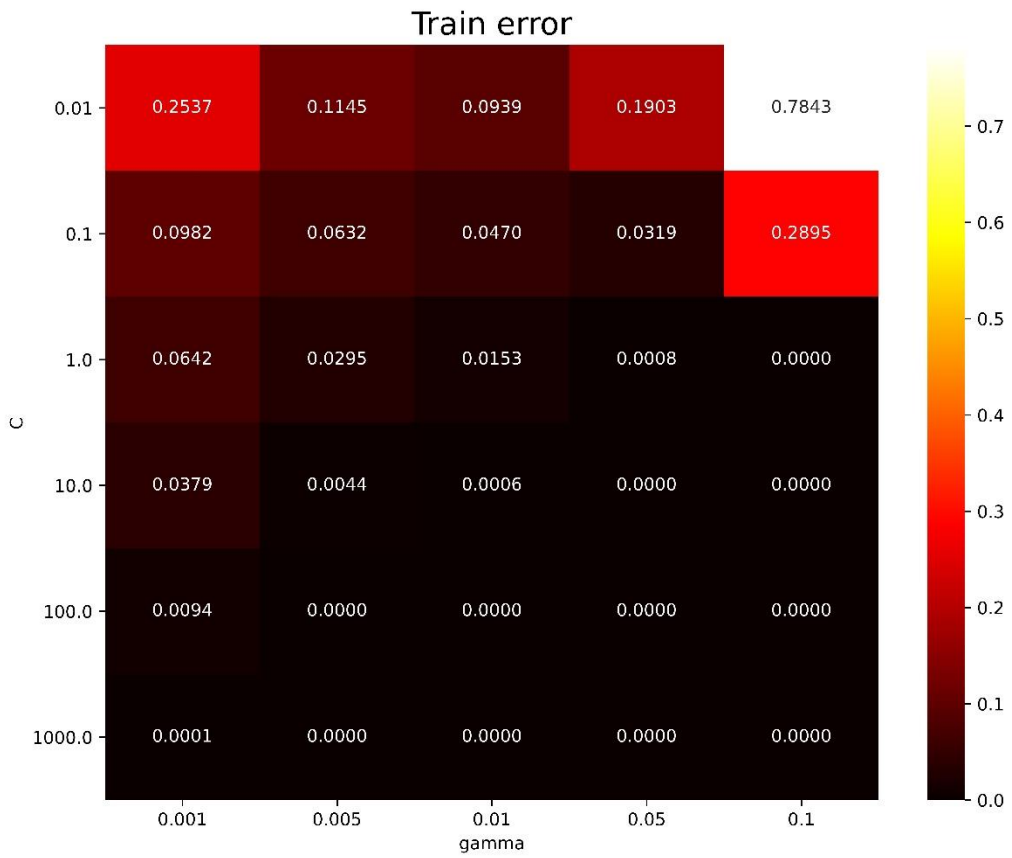
- Mô hình đạt độ lỗi nhỏ nhất tại C = 0.1



Nhận xét:

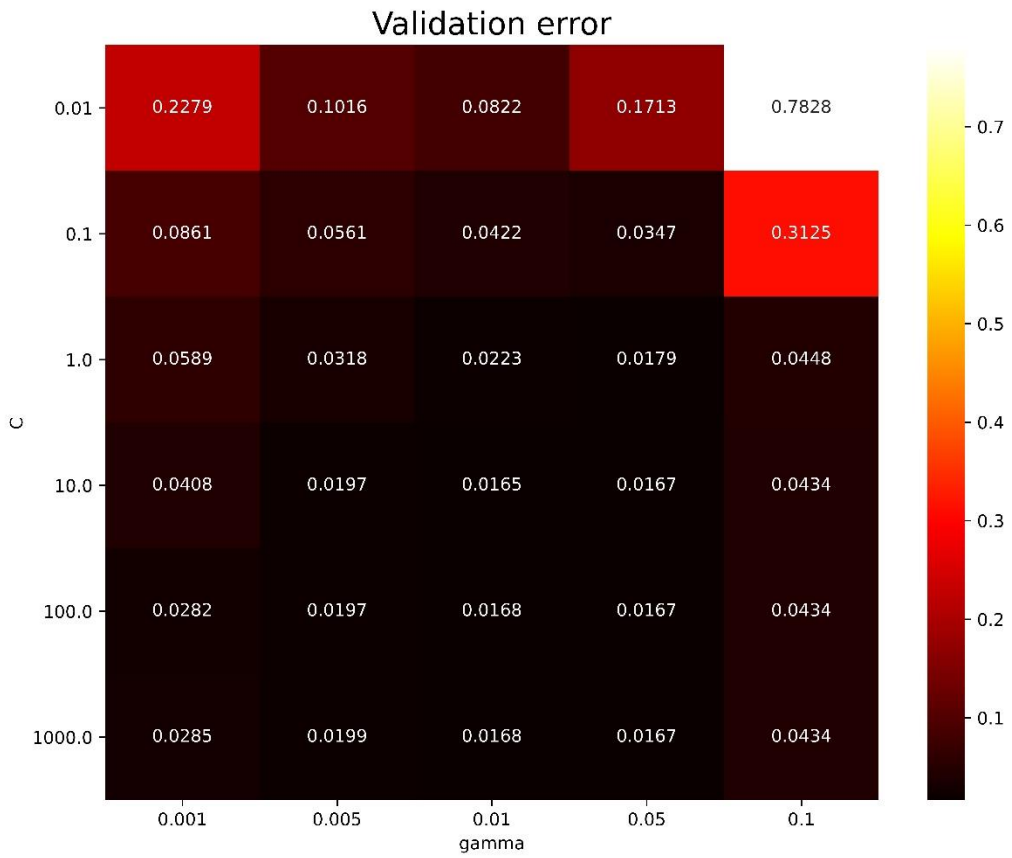
- C tăng dần thì thời gian thực thi tăng (vì mô hình đang cố gắng fit dữ liệu với số lượng dữ liệu lỗi thấp nhất).
- C nhỏ dần thì thời gian thực thi cũng cao (vì mô hình chấp nhận độ lỗi lớn dẫn đến số lượng support vector lớn khiến thời gian tính toán tăng).

Mô hình kernel rbf:



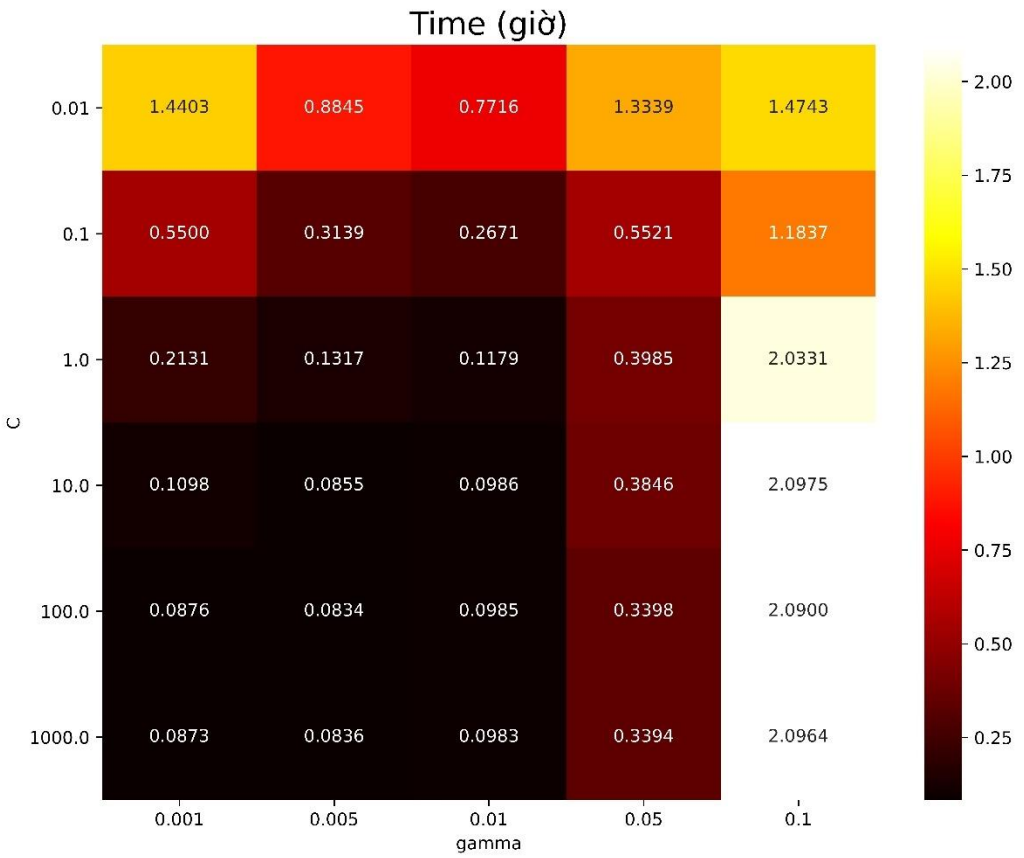
Nhận xét:

- Nhìn chung C và gamma càng lớn thì độ lỗi trên tập train càng nhỏ.
  (Phù hợp với lý thuyết khi C lớn và gamma lớn thì khả năng fit của mô hình vô cùng tốt).
- C lớn hơn 1 thì khi gamma tăng độ lỗi trên tập sẽ giảm liên tục.
- Dùng kernel khả năng fit vô cùng tốt (độ lỗi trên tập train bằng 0).



Nhận xét:

- Độ lỗi nhỏ nhất là 0.0165 ứng với gamma =0.01 và C = 10
- Ứng mỗi gamma thì C tăng thì độ lỗi tập validation luôn giảm dần.
- Có thể cảm nhận được khi gamma lớn 0.1 thì mô hình bị overfitting.

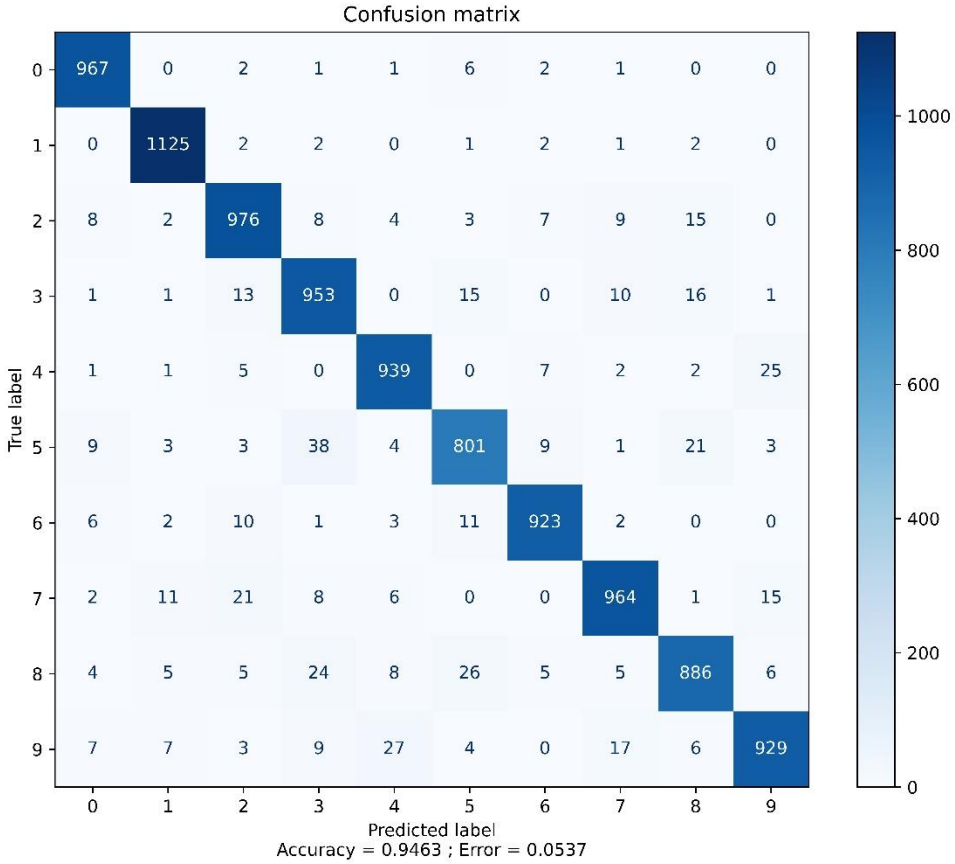


Nhận xét:

- Nhìn chung khi gamma tăng thì thời gian huấn luyện cũng tăng theo (Vì gamma cao sẽ cố gắng fit mạnh vào dữ liệu).
- Khi C nhỏ thời gian huấn luyện cũng cao như đã đánh giá ở mô hình linear.

4. Tổng kết:

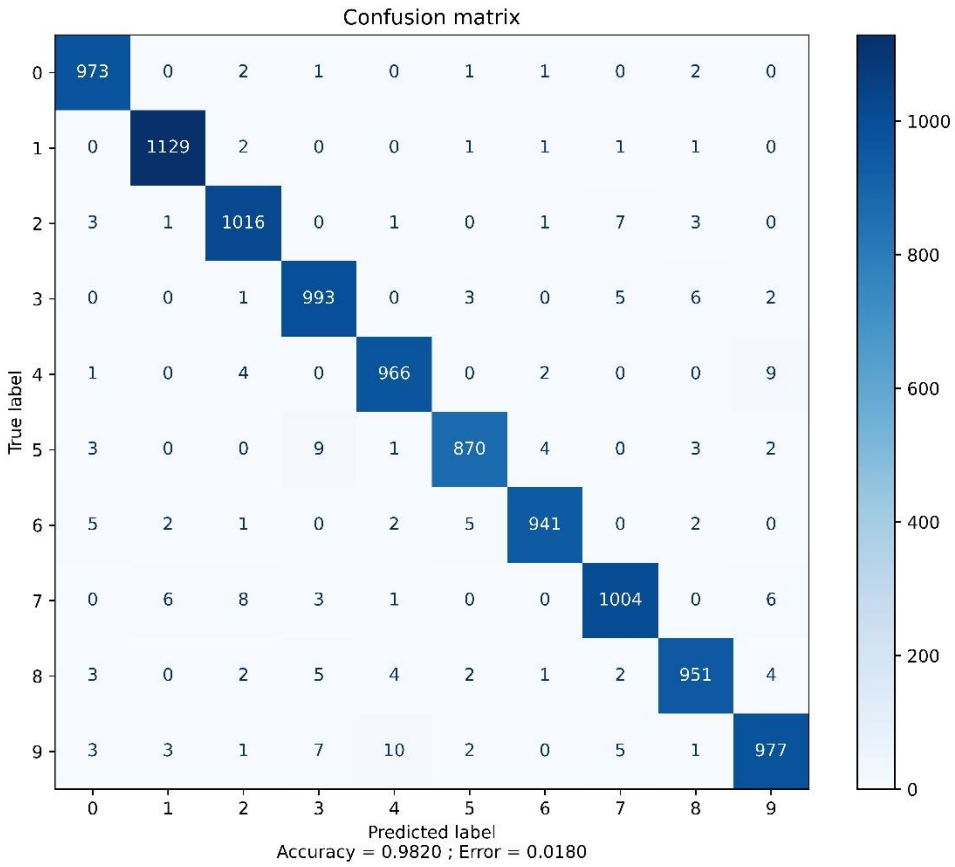
Kernel Linear: Với C = 0.1



Nhận xét:

- Phần lớn các ảnh đều được phân lớp đúng (Độ chính xác : 0.9463)
- Mô hình có sự nhầm lẫn phần lớn giữa số 4 và số 9 (27 mẫu số 9 được phân lớp là số 4 và 25 mẫu số 4 mà phân lớp số 9).
- Một số sự nhầm lẫn khác tương đối cao như giữa (8-3),(5-3),(7-1),(2-8)
- Nhìn chung những ảnh phân lớp sai đều có sự tương đồng về các đường nét tương đối cao.

Kernel Rbf: Với Gamma = 0.01, C = 10



Nhận xét:

- Nhìn chung sự nhầm lẫn ở các lớp số ở linear vẫn giữ nguyên. Nhưng được cải thiện hơn.
- Độ chính xác của mô hình đạt 0.982
- Lớp có số ảnh phân lớp sai nhiều nhất là lớp số 9.
- Lớp có độ chính xác cao nhất là lớp số 0 (Có thể cảm nhận vì số 0 rất dễ phân biệt).

Kết luận:

- Tập validation mô phỏng rất tốt cho tập test (Độ chính rất gần và nhỏ hơn một chút so với độ chính xác tập validation).