# Phân lớp ảnh chữ số viết tay bằng SVM

Đồ án vấn đáp môn Máy học

Hồng Thanh Hoài
*Khoa Công nghệ thông tin*
*Đại học Khoa học Tự nhiên*
Thành phố Hồ Chí Minh, Việt Nam
hthoai1006@gmail.com

Huỳnh Minh Huấn
*Khoa Công nghệ thông tin*
*Đại học Khoa học Tự nhiên*
Thành phố Hồ Chí Minh, Việt Nam
minhhuanhuynh289@gmail.com

***Tóm tắt nội dung*—Support Vector Machine (SVM) là một phương pháp học có giám sát trong các mô hình nhận dạng mẫu với cơ sở toán học chặt chẽ. Nó không chỉ hoạt động tốt với các dữ liệu được phân tách tuyến tính mà còn tốt với cả dữ liệu phân tách phi tuyến. Với nhiều bài toán, SVM mang lại kết quả tốt như mạng neuron với hiệu quả sử dụng tài nguyên tốt hơn hẳn. Ở đồ án này, nhóm tiến hành thực nghiệm SVM bằng thư viện scikit-learn trên bộ dữ liệu MNIST (phân lớp ảnh chữ số viết tay).**

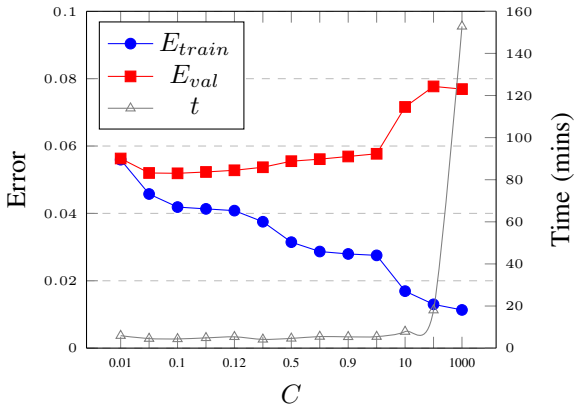
## I. Quá trình thực hiện

- *10/06/2019 - 15/06/2019*: Tìm hiểu và nắm vững kiến thức lý thuyết.
- *20/06/2019 - 25/06/2019*: Tìm hiểu thư viện scikit-learn và tiến hành train để chọn được các tham số tối ưu nhất với mỗi mô hình.
- *26/06/2019 - 28/06/2019*: Tổng hợp kết quả và viết báo cáo.
- *29/06/2019*: Ôn lại phần lý thuyết chuẩn bị cho vấn đáp.

## II. Huấn luyện SVM

### A. Dùng linear kernel

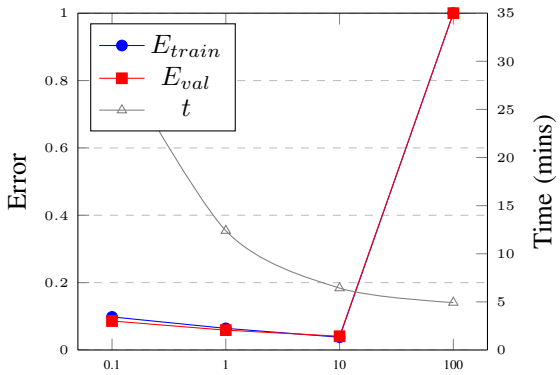
Độ lỗi trên tập *training* và *validation* khi thay đổi $C$.



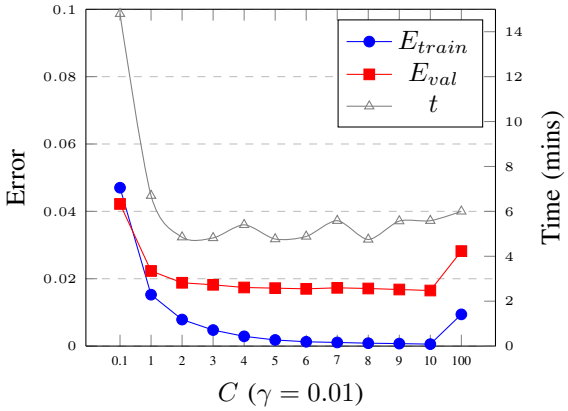

**Nhận xét**: Theo lý thuyết, $C$ lớn sẽ dẫn đến trường hợp "lower bias, higher variance", tức *overfitting*. $C$ nhỏ sẽ dẫn đến trường hợp "higher bias, lower variance", tức *underfitting*. Thật vậy, thực nghiệm cho thấy khi $C$ càng lớn thì $E_{train}$ càng giảm, nhưng $E_{val}$ lại tăng (overfitting). $C$ quá nhỏ thì $E_{val}$ có độ lỗi còn lớn (underfitting).

### B. Dùng RBF kernel

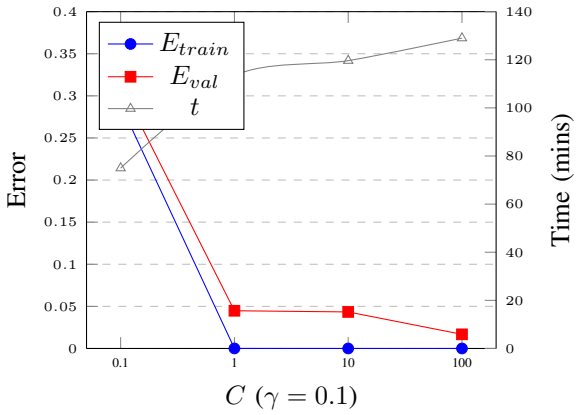
(1) Độ lỗi trên tập *training* và *validation* khi thay đổi $C$.




(2) Độ lỗi trên tập *training* và *validation* khi thay đổi $C$.



(3) Độ lỗi trên tập *training* và *validation* khi thay đổi $C$.




**Nhận xét**: Cũng tương tự như tham số $C$, $\gamma$ lớn sẽ dẫn đến trường hợp "lower bias, higher variance", tức *overfitting*. $\gamma$ nhỏ sẽ dẫn đến trường hợp "higher bias, lower variance", tức *underfitting*. Thật vậy, thực nghiệm cho thấy khi $\gamma$ nhỏ (biểu đồ 1) thì $E_{train}$ và $E_{val}$ đều rất lớn. Khi $\gamma$ lớn thì lại dẫn đến *overfitting*, $E_{train}$ có trường hợp bằng 0 nhưng $E_{val}$ lại lớn.

$\rightarrow$ Vậy, ta chọn hàm dự đoán cuối cùng với $E_{val}$ nhỏ nhất là khi dùng RBF kernel, với $C = 10$, $\gamma = 0.01$, cho $E_{val} = 0.0165$ ($Accuracy = 0.9835$).

## III. Đánh giá SVM

Với RBF kernel, $C = 10$, $\gamma = 0.01$, ta có kết quả sau:

| | |
|---|---|
| Training Score | 0.99942 |
| Training Error | 0.00058 |
| Testing Score | 0.982 |
| Testing Error | 0.018 |
| Time | 4min 47s |


## Lời cảm ơn

Nhóm chúng em xin chân thành cảm ơn thầy Trần Trung Kiên đã giảng giải rất kỹ lưỡng để chúng em có thể hiểu rõ và hoàn thành được đồ án này.


## Tài liệu


[1] Documentation of scikit-learn 0.21.2.
[2] Support Vector Machine, *Machine Learning cơ bản*.