Đồ án vấn đáp môn Máy học
Phân lớp ảnh chữ số viết tay bằng SVM

1 Nội dung đồ án

1.1 Tìm hiểu về lý thuyết mô hình SVM (Support Vector Machine)

Các tài liệu

- Các video bài giảng: 14 – SVM (đã học trên lớp), 15 – Kernel Methods, 16 – RBF (trong đó, video 14 và 15 là hai video chính về SVM; video 16 xem thêm để hiểu hơn về Gaussian/RBF kernel)
- Tài liệu (dễ đọc) về việc chuyển từ "primal form" sang "dual form": Mục 5 – "Lagrange duality" trong file "Lagrange.pdf" đính kèm

Các ý chính cần nắm

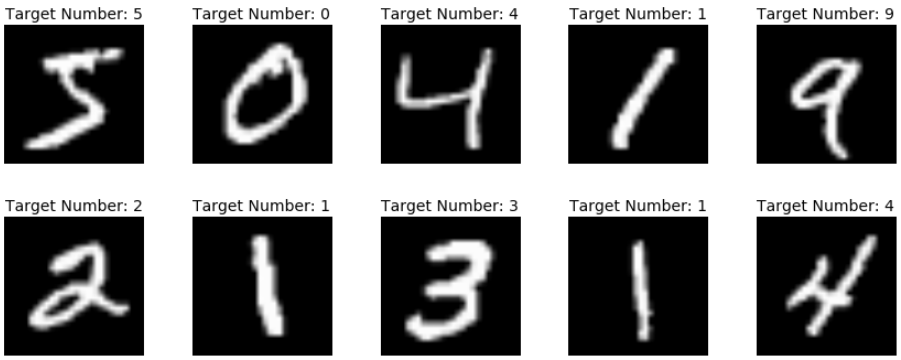
- Phân 2 lớp
  - Dữ liệu khả tách tuyến tính
    - Tập hypothesis của SVM?
    - Thuật toán học của SVM? [Ý chính cần nắm là SVM muốn tìm "siêu phẳng" phân lớp như thế nào, và điều này được thể hiện bằng cách tối ưu hóa hàm mục tiêu nào? Ở đây, bạn không cần phải hiểu về thuật toán giải bài toán tối ưu hóa này (Quadratic Programming / SMO) mà xem như là một hộp đen có input là bài toán tối ưu hóa của mình và output là lời giải]
    - "Support vector" là gì? "Support vector" liên quan như thế nào đến khả năng tổng quát hóa của SVM?
  - Dữ liệu không khả tách tuyến tính
    - SVM dùng soft-margin và kernel để giải quyết trường hợp dữ liệu không khả tách tuyến tính như thế nào?
    - Siêu tham số C trong soft-margin ảnh hưởng như thế nào đến việc học?
    - Siêu tham số γ trong Gaussian/RBF kernel ảnh hưởng như thế nào đến việc học?
- Phân K lớp (K > 2)
  - Từ SVM phân 2 lớp, làm thế nào để phân được K lớp? Gợi ý: "one-versus-one" là phương pháp thường được sử dụng trong SVM (thử đọc ở đây); ngoài ra, còn có phương pháp "one-versus-all".

1.2 Huấn luyện SVM để phân lớp ảnh chữ số viết tay

Mô tả dữ liệu

Bộ dữ liệu được sử dụng là bộ MNIST. Mỗi mẫu (example) trong bộ MNIST gồm: input là ảnh chữ số viết tay grayscale có kích thước 28×28 (như vậy, véc-tơ input sẽ có số chiều là 28×28=784), "correct ouput"

∈ {0, 1, …, 9} cho biết chữ số tương ứng của ảnh (như vậy, sẽ có tất cả 10 lớp). Dưới đây là một số mẫu trong bộ MNIST:



Bạn download file dữ liệu "mnist.pkl.gz" đính kèm. Trong file dữ liệu này, các giá trị pixel đã được scale về [0, 1] bằng cách chia cho 255. Hơn nữa, người ta cũng đã chia cho bạn 3 tập:

- Tập training: gồm 50.000 mẫu
- Tập validation: gồm 10.000 mẫu
- Tập test: gồm 10.000 mẫu

Bạn xem đoạn code đọc file dữ liệu này trong file "ReadMNIST.ipynb" đính kèm.

Cài đặt SVM

Với level hiện tại, bạn không nên cài đặt SVM từ A đến Z. Bạn sẽ sử dụng SVM đã được cài đặt sẵn trong thư viện scikit-learn (bạn đọc document để xem cách sử dụng; khá dễ). Thư viện này đã được cài đặt cho bạn khi bạn cài đặt gói Anaconda.

Huấn luyện SVM

- Dùng linear kernel (hay nói cách khác là không dùng kernel)
    - Thử nghiệm với các giá trị khác nhau của siêu tham số C; với mỗi giá trị C, ghi nhận lại: độ lỗi trên tập training, độ lỗi trên tập validation, thời gian huấn luyện
    - Bình luận về kết quả [Gợi ý: Theo lý thuyết thì giá trị C ảnh hưởng như thế nào đến quá trình học (C quá nhỏ thì sao, C quá lớn thì sao)? Kết quả thí nghiệm có phù hợp với lý thuyết không?]
- Dùng Gaussian/RBF kernel
    - Thử nghiệm với các giá trị khác nhau của siêu tham số C và γ; với mỗi giá trị C và γ, ghi nhận lại: độ lỗi trên tập training, độ lỗi trên tập validation, thời gian huấn luyện
    - Bình luận về kết quả.
- Chọn hàm dự đoán có độ lỗi nhỏ nhất trên tập validation là hàm dự đoán cuối cùng.

Đánh giá SVM

Với hàm dự đoán cuối cùng ở trên, bạn đánh giá hàm dự đoán này bằng cách đo độ lỗi trên tập test. Bạn có thể xem các kết quả của người ta ở đây.

2 Vấn đáp và nộp bài trên moodle

2.1 Vấn đáp

- Thời gian: dự kiến **30/6/2019** và **1/7/2019** (lịch cụ thể sẽ được thông báo sau)
- Địa điểm vấn đáp: **I81**
- Như đã nói trên moodle, yêu cầu làm việc nhóm là: **tất cả** các thành viên trong nhóm đều phải **hiểu rõ** về đồ án của nhóm mình
- Khi đi vấn đáp, mỗi nhóm cần nộp cho mình bản in của file báo cáo (không cần mang theo laptop để demo)
    - Ở đầu file báo cáo, ghi MSSV và họ tên của các thành viên trong nhóm. Đồng thời ghi lại quá trình làm việc nhóm (không phải là cuối cùng bạn mới viết phần này; khi làm việc nhóm thì cần phải lên kế hoạch, trong quá trình làm thì có thể có điều chỉnh kế hoạch, bản kế hoạch khi kết thúc làm việc nhóm chính là phần này)
    - Nội dung file báo cáo (ứng với phần "huấn luyện SVM" và "đánh giá SVM" trong mục 1.2 ở trên): ghi nhận lại các kết quả (nên dùng bảng biểu, đồ thị), bình luận về các kết quả
    - Báo cáo nên trình bày rõ ràng, ngắn gọn (khoảng 3-4 trang)

2.2 Nộp bài trên moodle

Bạn sẽ nộp bài (gồm báo cáo + code) trên moodle sau buổi vấn đáp.