CSC14005 – MACHINE LEARNING

FINAL PROJECT

Note: trong quá trình làm đồ án này, nếu có thắc mắc mail về ptnghia@fit.hcmus.edu.vn

Đây là bài tập nhóm, mỗi **nhóm 2 người**.

1. Nội dung đồ án

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

Tài liệu:

- Các video bài giảng: 14 – SVM, 15 – Kernel Methods, 16 – RBFs. 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? (SVM muốn tìm "siêu phẳng" phân lớp như thế nào?)
  - "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 nàđến việc học?
  - Siêu tham số γ trong Gaussian/RBF kernel ảnh hưởng như thế nàđến việc học?

Phân K lớp (K > 2):

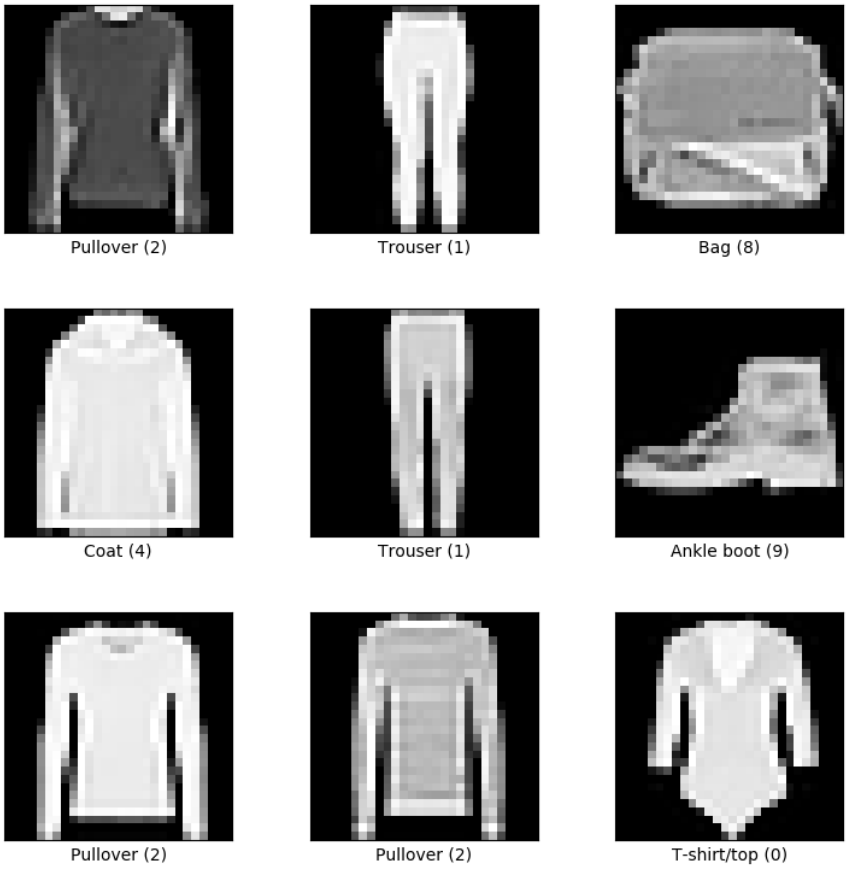
- Từ SVM phân 2 lớp, làm thế nàđể phân được K lớp? Gợi ý: "one-against-one" là phương pháp thường được sử dụng trong SVM (xem thêm)

1.2. Huấn luyện SVM để phân lớp ảnh thời trang

Mô tả dữ liệu

Bộ dữ liệu được sử dụng là bộ Fashion MNIST. Thông tin chi tiết về bộ dataset này có thể xem ở đây.

- Các ảnh trong Fashion-MNIST tương ứng với các lớp: áo phông, quần dài, áo thun, váy, áo khoác, dép, áo sơ-mi, giày thể thao, túi và giày cao gót.
- Dataset này gồm 60000 mẫu huấn luyện, và 10000 mẫu thử nghiệm.
- Từng mẫu là ảnh grayscale có kích thước 28×28 (như vậy, véc-tơ input sẽ có số chiều là 28×28=784), "correct ouput" thuộc 1 trong 10 lớp trên cho biết loại đối tượng thời trang. Dưới đây là một số mẫu trong bộ Fashion MNIST



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; rất 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
- 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 độ lỗi trên tập test. Thử so sánh với một số kết quả bằng phương pháp khác được ghi nhận trên bộ dataset này.

2. Nộp bài trên moodle

Sinh viên cần nộp:

- **Báo cáo**: 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), các phân tích, nhận xét.
- **Source code**: tất cả file source code gồm file notebook (.ipynb) hay python (.py)
- **Video** trình bày đồ án trong khoảng 15-20min. Upload lên youtube dưới dạng unlist và dẫn link trong file báo cáo hay file readme đính kèm (Không submit file video lên moodle)