Report Guides for Data Mining

Quách Đình Hoàng

2022/04/05

Chú ý: Việc tuân theo hướng dẫn này không nhất thiết giúp các nhóm được đánh giá cao (vì nó còn phụ thuộc vào chất lượng công việc được thực hiện) nhưng nó sẽ giúp các nhóm tổ chức báo cáo và hoàn thành project.

1. Tóm tắt (abstract) - 1 đoạn, khoảng 200-300 từ

Phần này là một đoạn mô tả động lực (motivation) của bài toán được đề cập trong project, các phương pháp nhóm sử dụng, và kết quả.

2. Giới thiệu (introduction) - Khoảng 1/2 đến 1 trang

Phần này giải thích bài toán và tại sao nó quan trọng. Nhóm nên cung cấp một số thông tin về bối cảnh liên quan đến bài toán. Những gì là input và output của bài toán cần được làm rõ. Chẳng hạn: "Input của bài toán là tập các {bức ảnh, tài liệu, hồ sơ cho vay, lịch sử giá của một cổ phiếu, ...}. Chúng tôi sử dụng các thuật toán {linear regression, decision tree, SVM, neural network, random forest, ...} để dự đoán {thể loại ảnh, thể loại văn bản, khả năng hồ sơ được duyệt cho vay, giá cổ phiếu sau một tháng, ...}". Mỗi nhóm có thể có input/output rất khác nhau. Việc xác định rõ input/output giúp người đọc dễ theo dõi phần còn lại của báo cáo.

3. Dữ liệu (data) - khoảng 1/2 đến 1 trang

Phần này mô tả tập dữ liệu nhóm sử dụng (nó đến từ đâu, được thu thập như thế nào, các đối tượng là gì, các đặc trưng/thuộc tính là gì, ...). Phần này không nên quá 1 trang. Tập dữ liệu có bao nhiêu training/validation/test examples? In ra vài training example trong tập dữ liệu nếu cần thiết. Nhóm có thực hiện tiền xử lý không? Nếu có thì như thế nào? (ví dụ: cách nhóm xử lý missing values, chuẩn hóa các thuộc tính, ...).

4. Phương pháp (method) - khoảng 1-3 trang

Nhóm cần tóm tắt ngắn gọn ý tưởng của các thuật toán mà nhóm sử dụng. Nhóm nên sử dụng các công thức hoặc hình ảnh để hỗ trợ phần giải thích ý tưởng của các thuật toán. Ví dụ: nếu nhóm sử dụng thuật toán SVM thì nhóm nên mô tả hàm mục tiêu và các ràng buộc của bài toán tối ưu trong thuật toán này; nếu nhóm sử dụng thuật toán logistic regression thì nhóm nên mô tả định nghĩa của hàm softmax. Nhóm cần thể hiện là mình hiểu cách thuật toán hoạt động. Mặc dù GV có thể biết/hiểu những thuật toán nhóm sử dụng, những người đọc khác có thể không. Do đó, với mỗi thuật toán, nhóm cần dành một đoạn văn mô tả cách nó hoạt động. Điều này đặt biệt cần thiết khi nhóm dùng các thuật toán mới và đặc thù, chưa được trình bày trong môn học (chẳng hạn: LSTM, GAN, ...), cho bài toán của mình.

5. Thực nghiệm, kết quả, và thảo luận (experiments, results, and discussions) - khoảng 1-3 trang

Nhóm nên mô tả chi tiết về các siêu tham số (hyper-parameters) được sử dụng cho mỗi thuật toán (ví dụ: learning rate, mini-batch size là bao nhiêu) và cách nhóm chọn được chúng. Nhóm có dùng cross-validation,

nếu có thì nhóm chia làm bao nhiêu phần (fold)? Nhóm cũng cần mô tả, giải thích (thêm công thức nếu cần thiết) các độ đo (measures) được sử dụng (chẳng hạn: accuracy, precision, F-measure, AUC, ...). Để trình bày kết quả, nhóm nên dùng kết hợp các bảng và biểu đồ. Nếu bài toán là phân loại thì nhóm cần mô tả confusion matrix hoặc AUC curves ngoài các độ đo precision, recall, và accuracy. Nếu bài toán là hồi quy, nhóm cần mô tả RMSE, R-square. Nhóm cũng cần trình bày các biểu đồ trực quan về kết quả. Ngoài ra, nhóm nên giải thích xem thuật toán có bị quá khớp với dữ liệu hay không và nhóm đã làm gì để giảm thiểu điều đó, nếu có. Nhóm cần đưa ra các nhận xét, bình luận về các số liệu, bảng biểu, biểu đồ. Các bảng, biểu đồ của nhóm phải được đánh số, đặt tiêu về, thêm các chú giải (legends, axis labels), và có kích thước, phông chữ dễ đọc khi in.

6. Kết luận (conclusion) - khoảng 1-2 đoạn

Phần này tóm tắt lại những kết quả chính của nhóm. Thuật toán nào có hiệu suất cao nhất? Tại sao nhóm nghĩ rằng một số thuật toán hoạt động tốt hơn những thuật toán khác? Nếu nhóm có nhiều thời gian hơn hoặc nhiều tài nguyên tính toán hơn, nhóm sẽ làm thêm cái gì?

7. Phụ lục (appendices) - tối đa 2 trang

Phần này là tùy chọn. Nó bao gồm các mô tả chi tiết hơn về các thuật toán hoặc các vấn đề kỹ thuật mà nhóm không có đủ không gian để trình bày ở các mục trên.

8. Đóng góp (contributions) - không có giới hạn

Phần này mô tả những gì mà mỗi thành viên trong nhóm đóng góp vào project này.

9. Tham khảo (references) - không có giới hạn

Phần này mô tả các tài liệu nhóm đã tham khảo để hoàn thành báo cáo này. Các tài liệu này có thể là (1) các bài báo, sách, blog mà nhóm đã tham khảo, (2) code, library nhóm đã sử dụng (ví dụ: dplyr, ggplot2, ...). Tài liệu tham khảo cần được viết theo đúng một trong các định dạng MLA, APA, hoặc IEEE. Nếu nhóm không biết các định dạng này thì mỗi tài liệu tham khảo phải bao gồm (theo thứ tự): các tác giả (cách nhau bằng dấu phẩy), tựa đề bài báo, tên conference/journal, nhà xuất bản, năm.