

Trường Chính sách Công và Quản lý Fulbright

Niên khóa 2021-2023

Học kỳ Xuân

(14/02 – 04/06/2022)

Khoa học Dữ liệu

Nhóm giảng viên

Giảng viên: Huỳnh Nhật Nam (email: nam.huynh@fulbright.edu.vn)

Trợ giảng: Nguyễn Minh Hiếu (email: mpp23.hieunguyen@student.fulbright.edu.vn)

Giờ lên lớp

8:30 đến 11:45

Giờ tiếp học viên

Huỳnh Nhật Nam: 12:00 – 13:30 và 17:00 - 18:00 (từ Thứ 2 đến Thứ 6)

Nguyễn Minh Hiếu: 19:00 - 20:00 (Thứ 2 - Chủ nhật, từ ngày 19/3/2023 đến ngày 9/4/2023)

Giờ tiếp học viên dành cho các nhóm hoặc cá nhân có nhu cầu thảo luận mở rộng hoặc cần thêm sự trợ giúp, hướng dẫn, giải đáp các thắc mắc. Giờ tiếp học viên đôi khi sẽ thay đổi để đảm bảo hỗ trợ đầy đủ cho các học viên. Ngoài ra, khi cần học viên có thể đăng ký trước với nhóm giảng viên để có thể gặp vào các thời gian phù hợp khác, hoặc liên lạc với các giảng viên qua e-mail.

Mục Tiêu Của Môn Học

- Hiểu các nguyên tắc cơ bản của khoa học dữ liệu và thuật toán học máy, các ứng dụng tiềm năng của chúng trong chính sách công và hỗ trợ phân tích, ra quyết định kinh doanh.
- Có thể lập kế hoạch và đề xuất một nghiên cứu phân tích dữ liệu, bao gồm các thành phần như yêu cầu dữ liệu, đánh giá chất lượng dữ liệu, phương pháp phân tích phù hợp và đánh giá hiệu suất mô hình.
- Có thể đọc, hiểu và đánh giá các ấn phẩm học thuật và báo cáo ngành sử dụng các phương pháp phân tích khoa học dữ liệu để đánh giá chính sách và các ứng dụng kinh doanh.
- Nhận thức về những nhu cầu trong việc áp dụng khoa học dữ liệu vào các vấn đề chính sách và kinh doanh trong đời sống thực.
- Hiểu biết tổng quan về các chuyên đề liên quan đến các mô hình trí tuệ nhân tạo phục vụ cho đối thoại (conversational AI và Large Language Models) và an ninh mạng.

Mô Tả Môn Học

Khóa học sẽ giới thiệu các nguyên tắc và khái niệm cơ bản nền tảng của một số thuật toán phổ biến nhất trong học máy và các ứng dụng của nó trong phân tích chính sách và kinh doanh. Các bài giảng sẽ được thiết kế để cung cấp cho học viên cái nhìn tổng quan về lĩnh vực khoa học dữ liệu và giới thiệu về các công cụ cơ bản của bộ môn này để xử lý, khám phá, hiểu dữ liệu và rút ra cái nhìn sâu sắc từ dữ liệu, cũng như những cân nhắc về việc sử dụng các công cụ này trong các ứng dụng ngoài đời thực. Trong các buổi giảng khách mời (guest lectures), học viên sẽ được chia sẻ về các chuyên đề liên quan đến các nghiên cứu ứng dụng mới nhất của trí tuệ nhân tạo tập trung vào các mô hình ngôn ngữ sử dụng dữ liệu lớn (Large Language Models, ví dụ ChatGPT), ứng dụng của quản trị và khai thác dữ liệu trong chuyển đổi số của quản lý nhà nước, và chủ đề an ninh mạng.

Khóa học không đặt nặng yếu tố toán học. Những chi tiết về toán học sẽ được cung cấp vừa đủ để giúp học viên hiểu các khái niệm khoa học dữ liệu và các kỹ thuật liên quan. Ngôn ngữ lập trình Python sẽ được sử dụng nhưng chỉ để trình bày các khái niệm và kỹ thuật này. Tuy nhiên, các kiến thức cơ bản về xác suất và thống kê có thể giúp học viên nắm bắt một số khái niệm và nguyên lý trên nhanh và chắc chắn hơn.

Khóa học sẽ bao gồm các bài giảng và các phiên thảo luận, trong đó học viên có thể đặt câu hỏi hoặc thảo luận về các chủ đề cụ thể mà họ quan tâm. Hiểu được những thách thức của việc giảng dạy phân tích dữ liệu mà không có công thức (toán học), Ban Giảng Viên đã chọn một cách tiếp cận có sắc thái cung cấp sự hài hòa, dựa trên tường thuật, sự kết hợp của lý thuyết, thảo luận trong lớp và các ứng dụng máy tính.

Bài đọc yêu cầu

- Cao Hào Thi. Tài liệu môn “*Thống kê và Phân tích Dữ liệu*” [CHT]
- James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. *An Introduction to Statistical Learning – with Applications in R*. Springer. [JWHT]
- Kelleher, J., and Tierney, B. 2018. *Data Science*. The MIT Press. [KT]
- Provost, F., and Fawcett, T. 2013. *Data Science for Business*. O’Reilly Media, Inc. [PF]
- Ronald K. Pearson, *Exploratory Data Analysis Using R*, CRC Press, 2018 [RKP]
- Shmueli, G., Bruce, P., and Patel, P. 2016. *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner, 3rd Edition*. Wiley [SBP]

Bài đọc tùy chọn

- Các bài đọc ngắn bổ sung về các chủ đề chuyên ngành.

Đánh Giá

1. *Chuyên cần [10%]*

2. *Tham gia trong lớp [10%]*

Học viên phải tham gia đóng góp tích cực và có tính xây dựng vào quá trình thảo luận các vấn đề được đưa ra trong lớp học.

2. Tóm tắt bài đọc yêu cầu và trình bày nhóm [30%]

Học viên được yêu cầu đọc, hiểu, tóm tắt các tài liệu cần đọc và trình bày nhóm các chủ đề chính của từng buổi giảng. Các chủ đề chính của mỗi buổi của 6 buổi giảng chính (mà học viên sẽ được yêu cầu trình bày theo nhóm) sẽ được thông báo cho học viên 1 tuần trước khi bắt đầu môn học. Học viên tự lập nhóm từ 2-3 thành viên. Các nhóm được yêu cầu giữ nguyên thành viên trong suốt thời gian của môn học. Các nhóm được yêu cầu thông báo về danh sách thành viên mỗi nhóm 1 tuần trước khi bắt đầu môn học. Bài trình bày không bắt buộc sử dụng trình chiếu. Nếu có sử dụng bài trình chiếu, các nhóm cần nộp nội dung trình chiếu cho BGV ít nhất 60 phút trước giờ vào lớp.

Bài trình bày sẽ được ban giảng viên chấm theo nhóm. Mỗi nhóm phải thông báo tỷ lệ đóng góp của từng thành viên sau mỗi bài trình bày. Điểm cuối cùng của mỗi cá nhân cho phần trình bày nhóm được tính dựa trên tỷ lệ đóng góp này.

3. Bài tập nhóm [20%]

Học viên được yêu cầu hoàn tất 3 bài tập nhóm. Các bài tập sẽ được thông báo cho học viên khi bắt đầu môn học. Học viên được yêu cầu nộp cả 3 bài tập trước 8g20 sáng của ngày 3/4.

3. Bài đánh giá cuối môn [30%]

Học viên được yêu cầu thực hiện một video trình bày về một trong những đề tài đã học. Học viên chọn đề tài từ danh sách mà BGV cung cấp vào buổi cuối cùng của môn học. Học viên cần trình bày được những khái niệm quan trọng trong đề tài và khả năng ứng dụng đề tài này để giải quyết một vấn đề trong công việc của học viên (hoặc một vấn đề bất kỳ trong thực tiễn).

Học viên nộp video cho BGV trước 8g20 sáng của ngày 16/04/2023 (tức 2 tuần sau khi kết thúc môn học).

Lịch Học

Bài giảng 1. Phân tích khám phá dữ liệu

[HNN]

- Tổng quan khoa học dữ liệu - lịch sử, khái niệm, ứng dụng
- Quy trình khai thác dữ liệu CRISP-DM
- Các kỹ thuật cơ bản trong phân tích khám phá dữ liệu
 - Phân loại dữ liệu
 - Các bước phân tích khám phá dữ liệu
 - Ước lượng và kiểm định thống kê biến 1 chiều

Tài liệu đọc. Chương 1 và Chương 2 [KT]

Mục 3.1 và 3.2, Chương 3 [RKP]

“Phần 1: Thống kê Ứng dụng” tài liệu môn “Thống kê và Phân tích Dữ liệu”

Bài giảng 2. Phân tích khám phá dữ liệu (tiếp theo)

[HNN]

- Các kỹ thuật cơ bản trong phân tích khám phá dữ liệu

- Ước lượng và kiểm định thống kê biến 2 chiều
- Các kỹ thuật xác định các điểm dữ liệu bất thường (outliers và inliers)
- Dữ liệu bị mất và các phương pháp thay thế

- Phương pháp Bootstrapping và các ứng dụng

Tài liệu đọc. Mục 3.3, Chương 3 [RKP]

Mục 5.2, Chương 5 [JWHT]

“Phần 1: Thống kê Ứng dụng” [CHT]

Bài giảng 3. Học có giám sát (supervised learning)

[HNN]

- Các khái niệm
 - Học có giám sát và học không giám sát
 - Mô hình dự báo và mô hình suy luận
 - Mô hình có thông số và mô hình không có thông số
 - Các loại lỗi (trong quá trình học có giám sát)
- Đánh đổi giữa độ lệch - phương sai
- Diễn giải các kết quả của mô hình tuyến tính từ quan điểm khoa học dữ liệu

Tài liệu đọc. Chương 2 và Chương 3 [JWHT]

“Phần 2: Phân tích Dữ liệu” [CHT]

Bài giảng 4. Học có giám sát (supervised learning) (tiếp theo)

[HNN]

- Các phương pháp kiểm định chéo (cross validation)
- Overfitting và underfitting
- Rò rỉ dữ liệu (data leakage) và thẩm định chéo lồng ghép (nested cross validation)
- Cải thiện mô hình tuyến tính
 - Các phương pháp chọn biến
 - Các phương pháp quy tắc hóa (regularisation) - Ridge and Lasso
 - Các phương pháp giảm chiều và phương pháp phân tích thành phần chính (principal component analysis)
- Các cân nhắc khi sử dụng dữ liệu nhiều chiều trong mô hình tuyến tính

Tài liệu đọc. Mục 5.1, Chương 5 và Chương 6 [JWHT]

Bài giảng 5. Học có giám sát (supervised learning) (tiếp theo)

[HNN]

- Các mô hình phân tích dự báo không thông số
 - Các mô hình cây quyết định: Random Forests, Bagging, Boosting
 - Support Vector Machine
- Khái lược các kỹ thuật phân loại (classification) phổ biến
- Đo lường hiệu suất phân loại

*Tài liệu đọc. Chương 4, Chương 8, và Chương 9 [JWHT]
Chương 7 và Chương 8 [PF]*

Bài giảng 6. Học không giám sát (unsupervised learning)	[HNN]
--	--------------

- Phân cụm (clustering)
 - Phương pháp K-means và phương pháp kết tụ (agglomerative)
 - Các thang đo đánh giá kết quả phân cụm
 - Xác định các biến đặc trưng của kết quả phân cụm
 - Dùng các mô hình học có giám sát để đánh giá kết quả phân cụm
- Khai thác thông tin về mối liên hệ giữa các quan sát
 - Association rules
 - Collaborative filterings

*Tài liệu đọc. Chương 10 [JWHT]
Chương 14 [SBP]*

Bài giảng 7. Bài giảng khách mời	[HNN/Khách mời]
---	------------------------

- Bài giảng khách mời 1. Các tiến bộ gần đây, ứng dụng và các hạn chế của các mô hình trí tuệ nhân tạo đàm thoại (conversational AI)
- Bài giảng khách mời 2. Chuyển đổi số trong quản lý nhà nước - vai trò của dữ liệu

Bài giảng 8. Bài giảng khách mời	[HNN/Khách mời]
---	------------------------

- Bài giảng khách mời 3. Giới thiệu về an ninh mạng, chiến lược, giá trị kinh doanh và tư duy “zero-trust”
- Tổng kết môn học