

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)
Phan Văn Hoàng Sơn (email: sonpvh@vng.com.vn)
Trợ giảng: Võ Tuấn Kiệt (email: mpp22.kietvo@student.fulbright.edu.vn)

Giờ lên lớp 13:30 đến 16:45

Giờ tiếp học viên

Huỳnh Nhật Nam: 17:00 – 18:00 (Thứ 2 đến Thứ 6)
Phan Văn Hoàng Sơn: 17:00 – 18:00 (Thứ 2 đến Thứ 6)
Võ Tuấn Kiệt: Online trên Teams: 20:00 - 21:00 (từ 18/03 - 01/04 và 12/04 - 26/04)
Trực tiếp tại Trường: 17:00 - 19:00 (từ 02/04 - 09/04)

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 phê bình 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 / hoặc 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 về nhu cầu an ninh mạng, thách thức và cơ hội.

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. Học viên sẽ được chia sẻ về chủ đề an ninh mạng để hiểu được lĩnh vực này là gì, các chiến lược an ninh mạng, các giá trị kinh doanh của an ninh mạng và xây dựng tư duy an ninh mạng (đặc biệt là tư duy không tin cậy – “zero-trust”) cho các nhà lãnh đạo.

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.

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 đánh giá cuối cùng sẽ ở dạng dự án nhóm, trong đó các nhóm sẽ được yêu cầu trình bày hiểu biết của mình về các chủ đề cụ thể được thảo luận trong bài giảng. Các nhóm chọn các chủ đề muốn tìm hiểu sâu dựa trên sở thích chung, trong đó có thể là mong muốn tóm tắt thông tin được đề cập trong các bài giảng. Nhưng các nhóm được khuyến khích hãy tìm kiếm các nguồn và ví dụ bên ngoài để chứng minh sự hiểu biết của mình về các chủ đề đã chọn. Các nhóm cũng có thể (và được khuyến khích) chọn phát triển một đề xuất cho một nghiên cứu phân tích để giải quyết một câu hỏi quan trọng về chính sách hoặc một vấn đề kinh doanh.

Bài đọc yêu cầu

- Provost, F., and Fawcett, T. 2013. *Data Science for Business*. O'Reilly Media, Inc. [PF, 2013]

Bài đọc tùy chọn

- James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. *An Introduction to Statistical Learning – with Applications in R*. Springer. [JWHT, 2013]
- Các bài đọc ngắn bổ sung về các chủ đề chuyên ngành.

Đánh Giá

1. Tham gia trong lớp [15%]

Học viên phải tham dự tất cả các bài giảng và đóng góp tích cực vào quá trình thảo luận các vấn đề được đưa ra trong lớp học.

2. Tóm tắt bài đọc yêu cầu [25%]

Học viên được yêu cầu tóm tắt các tài liệu cần đọc. 7 buổi giảng sẽ có 7 bài tóm tắt dưới dạng bài tập cá nhân, với yêu cầu độ dài bài viết không quá 1000 từ ($\pm 10\%$) và phải nộp bằng file word qua Teams trước 8 giờ 20 sáng của ngày diễn ra bài giảng.

3. Báo cáo/Đề xuất và bài thuyết trình nhóm [60%]

Học viên tự lập nhóm và không nên có quá 3 thành viên. Học viên phải thông báo việc lập nhóm với Ban Giảng Viên vào cuối Bài giảng 1 (07/04/2022). Học viên được yêu cầu làm việc theo nhóm và trình bày hiểu biết của mình dưới hình thức trình bày nhóm với 1 trong 2 phương án sau:

Phương án 1: Báo cáo Tổng hợp thảo luận trên lớp:

- Trong mỗi buổi giảng, học viên đều sẽ nhận được một nhóm các câu hỏi gợi ý để thảo luận, chia sẻ kinh nghiệm và quan điểm. 6 buổi giảng sẽ có 6 bài thảo luận trên lớp.
- Nhóm học viên lựa chọn phương án này được yêu cầu lập báo cáo tổng hợp nội dung của 6 bài thảo luận (mỗi bài có độ dài khoảng 1000 từ, $\pm 10\%$).
- Học viên lựa chọn 1 trong 6 bài thảo luận này để trình bày cho Ban giảng viên vào ngày diễn ra buổi học thứ 8 (09/04/2022). Nhóm học viên được khuyến khích tìm kiếm thông tin và ví dụ từ các nguồn bên ngoài để chứng minh sự hiểu biết của mình đối với báo cáo trình bày.

Phương án 2: Đề xuất nghiên cứu phân tích dữ liệu

- Các nhóm cũng có thể tập trung vào việc phát triển một đề xuất nghiên cứu phân tích dữ liệu để giải quyết một câu hỏi quan trọng về chính sách hoặc một vấn đề kinh doanh. Nhóm học viên sẽ phải (1) trình bày vấn đề, (2) xác định và thảo luận các yêu cầu dữ liệu, (3) đề xuất một kế hoạch đánh giá chất lượng dữ liệu, (4) lựa chọn phương pháp phân tích và (5) đánh giá mô hình phân tích kết quả.
- Nhóm học viên lựa chọn phương án này được yêu cầu lập một đề xuất nghiên cứu với các nội dung chính như trên. Đề xuất có độ dài khoảng 2500 từ ($\pm 10\%$).
- Các nhóm thực hiện phương án này nên tham vấn BGV để nhận được hỗ trợ về các câu hỏi gợi ý.

Trong cả hai phương án:

- Mỗi nhóm đều phải ghi rõ đóng góp và tỷ lệ đóng góp của từng thành viên (nhóm tự quyết định) tại cuối bài Báo cáo/Đề xuất.
- Báo cáo/Đề xuất phải được nộp phải bằng file word qua Teams trước 8 giờ 20 sáng ngày 09/04/2022.
- Vào ngày thuyết trình nhóm, mỗi nhóm có 20 phút để trình bày và 10 phút nhận phản biện/góp ý chỉnh sửa. File trình bày các nhóm cần gửi trước 13 giờ 20 ngày 09/04/2022 cho trợ giảng để được hỗ trợ chuẩn bị cho phần thuyết trình.
- Sau khi trình bày, mỗi nhóm có 1 tuần để hiệu chỉnh Báo cáo/Đề xuất và nộp lại trước 8 giờ 20 ngày 16/04/2022

Lịch Học

Bài giảng 1. Giới Thiệu

[HNN/PVHS]

Introduction

- Lịch sử khoa học dữ liệu, quy trình khai thác dữ liệu CRISP-DM và các biến thể
History of data science, the data mining process CRISP-DM and its variations
- Chọn đúng vấn đề khoa học dữ liệu
Picking the right data science problems

Tài liệu đọc. Chapter 1 and Chapter 2 [PF, 2013]

Bài giảng 2. Phân Tích Khám Phá Dữ Liệu

[HNN/PVHS]

Exploratory Data Analysis

- Các kỹ thuật cơ bản trong phân tích khám phá dữ liệu
Fundamental techniques in exploratory data analysis
- Phân tích khám phá dữ liệu trong các ứng dụng thực tế
Exploratory data analysis in real-life applications

Tài liệu đọc. EDA for Feature Selection in Machine Learning [Google Cloud]

Tham khảo thêm “Phần 1: Thống kê Ứng dụng” từ môn “Thống kê và Phân tích Dữ liệu” [CHT]

Bài giảng 3. Sơ Đồ Học Tập - Hồi Quy Tuyến Tính

[HNN/PVHS]

Learning Diagram, Linear Regression Revisited

- Sơ đồ học máy, các loại lỗi và sự đánh đổi giữa độ lệch - phương sai
The (machine) learning diagram, types of errors, and bias-variance tradeoff
- Xem xét lại hồi quy tuyến tính từ các quan điểm khoa học dữ liệu và đánh giá
Revisiting linear regression from data science perspectives and evaluation.

Tài liệu đọc. Chapter 2 and Chapter 3 [JWHT, 2013]

Tham khảo thêm “Phần 2: Phân tích Dữ liệu” từ môn “Thống kê và Phân tích Dữ liệu” [CHT]

Bài giảng 4. Phân Loại và Lấy Mẫu Lại

[HNN]

Classification and Resampling

- Các kỹ thuật phân loại phổ biến bao gồm hồi quy logistics, cây quyết định
Popular classification techniques including logistics regression, decision trees
- Đo lường hiệu suất phân loại
Measuring the classification performance
- Quá khớp, thăm định chéo, bootstrapping, rò rỉ dữ liệu
Overfitting, cross validation, bootstrapping, data leakage

Tài liệu đọc. Chapters 4, 5, 7, 8 [PF, 2013], Chapters 4, 5, 8 and 9 [JWHT, 2013]

Bài giảng 5. Quy Tắc Hóa

[HNN/PVHS]

Regularisation

- Lựa chọn đặc tính, Ridge và Lasso, các phương pháp giảm kích thước
Feature selection, Ridge and Lasso, dimension reduction methods

Tài liệu đọc. Chapter 6 [JWHT, 2013]

Bài giảng 6. Phân Cụm và Đánh Giá Mô Hình - Đạo Đức và Quy Định Trong Khoa Học Dữ Liệu

[HNN/PVHS]

Clustering and Model Evaluation - Ethics and Regulations

- Các kỹ thuật phổ biến để phân cụm và đánh giá kết quả phân nhóm
Popular techniques for clustering and evaluating clustering results
- Đánh giá các mô hình phân tích dữ liệu trong bối cảnh kinh doanh
Evaluating data analytic models in business context
- Các mối quan tâm về đạo đức và quy định trong khoa học dữ liệu
Ethics and regulation concerns in data science

Tài liệu đọc. Chapters 6 [PF, 2013], Chapter 10 [JWHT, 2013]

Tham khảo thêm Chapter 2, Chapter 5 từ quyển Evaluating ML Models [AZ]

Bài giảng 7. An Ninh Mạng Cho Các Nhà Lãnh Đạo

[HNN/PVHS/Khách mời]

Cybersecurity for Leaders

- Bài giảng khách mời: Giới thiệu về an ninh mạng, chiến lược, giá trị kinh doanh và tư duy “zero-trust”
Guest lecture: cybersecurity introduction, strategies, business values and zero-trust mindset
- Đánh giá và phản hồi khóa học
Course review and feedback

Tài liệu đọc. TBA

Bài giảng 8. Thuyết Trình Nhóm

[HNN/PVHS]

Group presentations
