# Lời nói đầu

Quá trình số hóa của xã hội chúng ta đã nâng cao đáng kể khả năng tạo ra và thu thập dữ liệu từ các nguồn đa dạng. Một lượng dữ liệu khổng lồ đã tràn ngập hầu hết mọi khía cạnh của cuộc sống. Sự bùng nổ của dữ liệu, dù được lưu trữ hay chỉ là dữ liệu thoáng qua, đã tạo ra nhu cầu cấp bách về các kỹ thuật mới và công cụ tự động thông minh, giúp chúng ta chuyển đổi khối lượng lớn dữ liệu đó thành thông tin và tri thức hữu ích. Điều này đã dẫn đến sự ra đời của một lĩnh vực triển vọng và phát triển mạnh mẽ trong khoa học máy tính được gọi là *khai phá dữ liệu* cùng với các ứng dụng đa dạng của nó. Khai phá dữ liệu, còn được biết đến phổ biến với tên gọi *khám phá tri thức từ dữ liệu (knowledge discovery from data – KDD)*, là quá trình tự động hoặc thuận tiện nhằm trích xuất các mô hình biểu thị tri thức được ẩn chứa hoặc thu thập từ các cơ sở dữ liệu lớn, kho dữ liệu, Web, các kho lưu trữ thông tin khổng lồ khác, hoặc các luồng dữ liệu.

Cuốn sách này khám phá các khái niệm và kỹ thuật của việc *phát hiện tri thức* và *khai phá dữ liệu*. Là một lĩnh vực đa ngành, khai phá dữ liệu dựa vào các công trình từ nhiều lĩnh vực bao gồm thống kê, học máy, nhận dạng mẫu, công nghệ cơ sở dữ liệu, truy xuất thông tin, xử lý ngôn ngữ tự nhiên, khoa học mạng, hệ thống dựa trên tri thức, trí tuệ nhân tạo, tính toán hiệu năng cao và trực quan hóa dữ liệu. Chúng tôi tập trung vào các vấn đề liên quan đến tính khả thi, tính hữu ích, hiệu quả và khả năng mở rộng của các kỹ thuật nhằm phát hiện các mẫu ẩn trong các *tập dữ liệu lớn*. Do đó, cuốn sách này không nhằm mục đích giới thiệu về thống kê, học máy, hệ thống cơ sở dữ liệu hay các lĩnh vực tương tự, mặc dù chúng tôi cung cấp một số kiến thức nền để hỗ trợ đọc giả hiểu được vai trò của chúng trong khai phá dữ liệu. Thay vào đó, cuốn sách là một giới thiệu tổng hợp về khai phá dữ liệu. Nó hữu ích cho sinh viên ngành khoa học máy tính, các nhà phát triển ứng dụng và các chuyên gia kinh doanh, cũng như các nhà nghiên cứu liên quan đến bất kỳ ngành nào trong số các ngành đã liệt kê ở trên.

Khai phá dữ liệu xuất hiện vào cuối thập niên 1980, đạt được những bước tiến

vượt bậc trong thập niên 1990 và tiếp tục phát triển mạnh mẽ vào thế kỷ mới. Cuốn sách này trình bày một bức tranh tổng quan về lĩnh vực, giới thiệu các khái niệm và kỹ thuật khai phá dữ liệu thú vị, đồng thời bàn luận về các ứng dụng và hướng nghiên cứu. Một động lực quan trọng khi viết cuốn sách này là nhu cầu xây dựng một khuôn khổ có tổ chức cho việc nghiên cứu khai phá dữ liệu—một nhiệm vụ đầy thách thức do tính đa ngành rộng lớn của lĩnh vực đang phát triển nhanh này. Chúng tôi hy vọng cuốn sách sẽ khuyến khích những người có các nền tảng và kinh nghiệm khác nhau trao đổi quan điểm về khai phá dữ liệu, từ đó góp phần thúc đẩy và định hình thêm cho lĩnh vực thú vị và năng động này.

## Cấu trúc của cuốn sách

Cho tới nay, lĩnh vực khai phá dữ liệu đã có những tiến bộ vượt bậc. Nhiều phương pháp, hệ thống và ứng dụng mới trong khai phá dữ liệu đã được phát triển, đặc biệt là để xử lý các loại dữ liệu mới, bao gồm mạng thông tin, đồ thị, các cấu trúc phức tạp, và luồng dữ liệu, cũng như văn bản, Web, đa phương tiện, chuỗi thời gian và dữ liệu không gian – thời gian. Sự phát triển nhanh chóng cùng với nội dung kỹ thuật phong phú và mới mẻ đã khiến việc bao quát toàn bộ lĩnh vực trong một cuốn sách trở nên khó khăn. Chúng tôi quyết định tập trung vào các nội dung cốt lõi với phạm vi và chiều sâu đủ lớn, đồng thời để việc xử lý các loại dữ liệu phức tạp và ứng dụng của chúng cho những cuốn sách chuyên về các chủ đề cụ thể đó.

Các chương của cuốn sách được mô tả ngắn gọn như sau, với sự nhấn mạnh vào xu thế mới:

**Chương 1** cung cấp *phần giới thiệu* về lĩnh vực đa ngành của khai phá dữ liệu. Chương này bàn về quá trình tiến hóa của công nghệ thông tin, dẫn đến nhu cầu về khai phá dữ liệu và tầm quan trọng của các ứng dụng của nó. Nó xem xét các loại dữ liệu cần được khai thác và trình bày một cách phân loại tổng quát các nhiệm vụ khai phá dữ liệu, dựa trên các loại tri thức cần được phát hiện, các loại công nghệ được sử dụng, và các loại ứng dụng hướng tới. Chương này cho thấy khai phá dữ liệu là sự giao thoa của nhiều ngành khác nhau với các ứng dụng rộng khắp. Cuối cùng, nó bàn luận về cách thức khai phá dữ liệu có thể tác động đến xã hội.

**Chương 2** giới thiệu về *dữ liệu, phép đo và tiền xử lý dữ liệu*. Ban đầu, chương thảo luận về các đối tượng dữ liệu và các loại thuộc tính, sau đó giới thiệu các phép đo điển hình cho việc mô tả dữ liệu thống kê cơ bản. Nó cũng giới thiệu các phương pháp đo lường mức độ tương đồng và khác biệt của các loại dữ liệu khác

nhau. Tiếp theo, chương chuyển sang giới thiệu các kỹ thuật tiền xử lý dữ liệu. Cụ thể, chương này giới thiệu khái niệm chất lượng dữ liệu cùng với các phương pháp làm sạch dữ liệu và tích hợp dữ liệu. Nó cũng thảo luận về các phương pháp khác nhau cho việc biến đổi dữ liệu và giảm chiều dữ liệu.

**Chương 3** cung cấp một giới thiệu toàn diện về *kho dữ liệu* và *xử lý phân tích trực tuyến (online analytical processing – OLAP)*. Chương bắt đầu với một định nghĩa được chấp nhận rộng rãi về kho dữ liệu, giới thiệu kiến trúc và khái niệm hồ dữ liệu (data lake). Sau đó, chương nghiên cứu thiết kế logic của kho dữ liệu dưới dạng mô hình dữ liệu nhiều chiều, đồng thời xem xét các thao tác OLAP và cách lập chỉ mục dữ liệu OLAP để phân tích hiệu quả. Chương này bao gồm một phần trình bày sâu sắc các kỹ thuật xây dựng dữ liệu khối (data cube) như một phương thức triển khai kho dữ liệu.

Chương 4 và 5 trình bày các phương pháp *khai phá các mẫu thường xuyên, các luật kết hợp và mối tương quan* trong các tập dữ liệu lớn. **Chương 4** giới thiệu các khái niệm cơ bản, như phân tích giỏ hàng (market basket analysis), với nhiều kỹ thuật khai phá tập mục thường xuyên được trình bày một cách có hệ thống. Các kỹ thuật này bao gồm từ thuật toán Apriori cơ bản và các biến thể của nó đến các phương pháp tiên tiến hơn nhằm cải thiện hiệu quả, bao gồm phương pháp tăng trưởng mẫu thường xuyên (frequent pattern growth), khai phá mẫu thường xuyên với định dạng dữ liệu theo chiều dọc, và khai phá các tập mục thường xuyên đóng và cực đại. Chương cũng bàn về các phương pháp đánh giá mẫu và giới thiệu các chỉ số đo lường để khai phá các mẫu có tương quan. **Chương 5** tập trung vào các phương pháp khai phá mẫu tiên tiến. Nó bàn về các phương pháp khai phá mẫu trong không gian đa cấp và nhiều chiều, khai phá luật kết hợp định lượng, khai phá dữ liệu có chiều lớn, khai phá các mẫu hiếm và tiêu cực, khai thác mẫu nén hay xấp xỉ, và khai phá mẫu dựa trên ràng buộc. Sau đó, chương chuyển sang các phương pháp tiên tiến cho khai phá mẫu tuần tự và mẫu đồ thị con. Ngoài ra, chương cũng trình bày các ứng dụng của khai thác mẫu, bao gồm khai phá cụm từ trong dữ liệu văn bản và khai phá các lỗi copy – paste trong chương trình phần mềm.

Chương 6 và 7 mô tả phương pháp *phân loại dữ liệu*. Do tầm quan trọng và sự đa dạng của các phương pháp phân loại, nội dung được chia thành hai chương riêng biệt. **Chương 6** giới thiệu các khái niệm và phương pháp cơ bản cho việc phân loại, bao gồm việc xây dựng cây quyết định, phân loại Bayes, bộ phân loại $k$NN ($k$ – nearest neighbor, $k$ – láng giềng gần nhất) và các bộ phân loại tuyến tính. Nó cũng bàn về các phương pháp đánh giá và lựa chọn mô hình cũng như các phương pháp cải thiện độ chính xác của phân loại, bao gồm các phương pháp

tổng hợp (ensemble) và cách xử lý dữ liệu mất cân bằng. **Chương 7** trình bày các phương pháp phân loại nâng cao, bao gồm lựa chọn đặc trưng, mạng tin cậy Bayes, SVM (support vector machine, máy véctơ hỗ trợ), phân loại dựa trên quy tắc và dựa trên mẫu. Các chủ đề bổ sung bao gồm phân loại với giám sát yếu, phân loại với kiểu dữ liệu phong phú, phân loại đa lớp, học khoảng cách từ xa, diễn giải kết quả phân loại, thuật toán di truyền và học tăng cường.

*Phân tích cụm* là chủ đề của Chương 8 và 9. **Chương 8** giới thiệu các khái niệm và phương pháp cơ bản trong phân cụm dữ liệu, bao gồm tổng quan về các phương pháp phân tích cụm cơ bản, các phương pháp phân vùng, phương pháp phân cấp, cũng như các phương pháp dựa trên mật độ và dựa trên lưới. Nó cũng giới thiệu các phương pháp đánh giá hiệu quả của việc phân cụm. **Chương 9** bàn về các phương pháp phân cụm nâng cao, bao gồm phân cụm dựa trên mô hình xác suất, phân cụm dữ liệu có chiều lớn, phân cụm dữ liệu đồ thị và mạng, cũng như phân cụm bán giám sát.

**Chương 10** giới thiệu về *học sâu*, một họ các kỹ thuật mạnh mẽ dựa trên mạng nơron nhân tạo với nhiều ứng dụng rộng rãi trong thị giác máy tính, xử lý ngôn ngữ tự nhiên, dịch máy, phân tích mạng xã hội, và nhiều lĩnh vực khác. Chúng tôi bắt đầu với các khái niệm cơ bản và một kỹ thuật nền tảng gọi là thuật toán lan truyền ngược (backpropagation). Sau đó, chương giới thiệu các kỹ thuật nhằm cải thiện quá trình huấn luyện các mô hình học sâu, bao gồm các hàm kích hoạt đáp ứng, tốc độ học thích nghi, phương pháp dropout, tiền huấn luyện, hàm mất mát entropy chéo và tự mã hóa (autoencoder). Ngoài ra, chương cũng trình bày một số kiến trúc học sâu được sử dụng phổ biến, từ mạng nơron truyền thẳng (feed−forward neural network), mạng nơron tích chập (convolutional neural network), mạng nơron hồi tiếp (recurrent neural network) đến mạng nơron đồ thị (graph neural network).

**Chương 11** dành riêng cho việc *phát hiện ngoại lệ*. Chương giới thiệu các khái niệm cơ bản về ngoại lệ và phân tích ngoại lệ, đồng thời bàn về các phương pháp phát hiện ngoại lệ theo góc nhìn về mức độ giám sát (tức là các phương pháp có giám sát, bán giám sát và không giám sát), cũng như theo góc nhìn về các phương pháp tiếp cận (bao gồm các phương pháp thống kê, phương pháp dựa trên lân cận, phương pháp dựa trên tái tạo, phương pháp dựa trên phân cụm và phương pháp dựa trên phân loại). Chương cũng thảo luận về các phương pháp khai phá ngoại lệ theo ngữ cảnh và theo tập hợp, cũng như các phương pháp phát hiện điểm bất thường trong dữ liệu có chiều lớn.

Cuối cùng, trong **Chương 12**, chúng tôi bàn luận về các *xu hướng tương lai* và *hướng nghiên cứu mới* trong lĩnh vực khai phá dữ liệu. Chúng tôi bắt đầu với phần

giới thiệu ngắn gọn về việc khai phá các loại dữ liệu phức tạp, bao gồm dữ liệu văn bản, đồ thị và mạng, cũng như dữ liệu không gian – thời gian. Sau đó, chúng tôi giới thiệu một số ứng dụng của khai phá dữ liệu, từ phân tích cảm xúc và ý kiến, phát hiện sự thật và thông tin sai lệch, lan truyền thông tin và dịch bệnh, đến năng suất và khoa học nhóm. Tiếp theo chương đề cập đến các phương pháp khai phá dữ liệu khác, bao gồm cấu trúc hóa dữ liệu phi cấu trúc, tăng cường dữ liệu, phân tích nhân quả, mạng dưới dạng ngữ cảnh và tự động hóa học máy (auto – ML). Cuối cùng, chương thảo luận về tác động xã hội của khai phá dữ liệu, bao gồm khai phá dữ liệu bảo vệ quyền riêng tư, tương tác giữa con người và thuật toán, tính công bằng, khả năng giải thích và độ bền vững, cũng như khai phá dữ liệu vì lợi ích xã hội.

Ngoài ra, một tập hợp các phụ lục giới thiệu ngắn gọn các kiến thức toán học cần thiết để hiểu nội dung của cuốn sách.

Trong toàn bộ văn bản, font chữ *in nghiêng* được sử dụng để nhấn mạnh các thuật ngữ đã được định nghĩa, và font chữ **in đậm** được sử dụng để làm nổi bật hoặc tóm tắt những ý chính. Font chữ ***in đậm nghiêng*** được sử dụng để biểu thị các đại lượng nhiều chiều.

Cuốn sách này có một số đặc điểm mạnh mẽ giúp nó nổi bật so với các giáo trình khác về khai phá dữ liệu. Nó trình bày một phạm vi rất rộng nhưng cũng sâu sắc về các nguyên lý của khai phá dữ liệu. Các chương được viết sao cho độc lập nhất có thể, cho phép đọc giả lựa chọn đọc theo thứ tự phù hợp với sở thích cá nhân. Các chương nâng cao cung cấp một cái nhìn tổng quan quy mô lớn hơn và có thể được xem là phần đọc thêm đối với những đọc giả quan tâm. Tất cả các phương pháp chính của khai phá dữ liệu đều được trình bày. Cuốn sách giới thiệu các chủ đề quan trọng về khai phá dữ liệu liên quan đến phân tích OLAP đa chiều, một chủ đề thường bị bỏ qua hoặc chỉ được đề cập tối thiểu trong các sách khác về khai phá dữ liệu. Ngoài ra, cuốn sách còn duy trì trang web với nhiều tài nguyên trực tuyến nhằm hỗ trợ giảng viên, sinh viên và các chuyên gia trong lĩnh vực. Các tài nguyên này sẽ được trình bày chi tiết hơn trong mục tiếp theo.

## Đối với giảng viên

Cuốn sách này được thiết kế nhằm cung cấp một cái nhìn tổng quát nhưng chi tiết về lĩnh vực khai phá dữ liệu. Trước tiên, nó có thể được sử dụng để giảng dạy một khóa học giới thiệu về khai phá dữ liệu ở bậc đại học nâng cao hoặc ở bậc sau

đại học năm nhất. Hơn nữa, cuốn sách còn cung cấp các tài liệu thiết yếu cho một khóa học sau đại học nâng cao về khai phá dữ liệu.

Tùy thuộc vào thời gian giảng dạy, nền tảng của sinh viên và sở thích của bạn, bạn có thể lựa chọn các chương khác nhau để giảng dạy theo thứ tự liên tiếp khác nhau. Ví dụ, một khóa học giới thiệu có thể bao gồm các chương sau:

- Chương 1: Giới thiệu
- Chương 2: Dữ liệu, phép đo và tiền xử lý dữ liệu
- Chương 3: Kho dữ liệu và xử lý phân tích trực tuyến
- Chương 4: Khai thác mẫu: các khái niệm và phương pháp cơ bản
- Chương 6: Phân loại: các khái niệm cơ bản
- Chương 8: Phân tích cụm: các khái niệm và phương pháp cơ bản

Nếu có thời gian, có thể lựa chọn thêm một số nội dung về học sâu (Chương 10) hoặc phát hiện điểm bất thường (Chương 11). Trong mỗi chương, các khái niệm cơ bản nên được trình bày, trong khi một số phần về các chủ đề nâng cao có thể được tùy chọn để giảng dạy.

Lấy một ví dụ khác, đối với những nơi có khóa học học máy tập trung sâu vào học có giám sát, khóa học khai phá dữ liệu có thể tập trung đào sâu vào phân cụm. Một khóa học như vậy có thể dựa trên các chương sau:

- Chương 1: Giới thiệu
- Chương 2: Dữ liệu, phép đo và tiền xử lý dữ liệu
- Chương 3: Kho dữ liệu và xử lý phân tích trực tuyến
- Chương 4: Khai thác mẫu: các khái niệm và phương pháp cơ bản
- Chương 8: Phân tích cụm: các khái niệm và phương pháp cơ bản
- Chương 9: Phân tích cụm: các phương pháp nâng cao
- Chương 11: Phát hiện điểm bất thường

Một giảng viên dạy khóa học khai phá dữ liệu nâng cao có thể thấy Chương 12 đặc biệt hữu ích, vì chương này bàn về một phổ rộng các chủ đề mới trong khai phá dữ liệu, đang phát triển năng động và nhanh chóng.

Ngoài ra, bạn cũng có thể lựa chọn giảng dạy toàn bộ cuốn sách theo chuỗi hai khóa học, bao quát tất cả các chương trong sách, và khi có thời gian, bổ sung một số chủ đề nâng cao như khai thác đồ thị và mạng. Các nội dung cho các chủ đề nâng cao đó có thể được lựa chọn từ các chương bổ trợ có sẵn trên trang web của cuốn sách, kèm theo một tập hợp các bài báo nghiên cứu được lựa chọn.

Các chương riêng lẻ trong cuốn sách này cũng có thể được sử dụng cho các buổi hướng dẫn hoặc cho các chủ đề đặc biệt trong các khóa học liên quan, như học máy, nhận dạng mẫu, kho dữ liệu và phân tích dữ liệu thông minh.

Mỗi chương kết thúc với một bộ bài tập, phù hợp để giao làm bài tập về nhà. Các bài tập này có thể là các câu hỏi ngắn nhằm kiểm tra việc nắm vững các khái niệm cơ bản, các câu hỏi dài đòi hỏi tư duy phân tích, hoặc các dự án triển khai. Một số bài tập cũng có thể được sử dụng làm chủ đề thảo luận nghiên cứu. Các ghi chú tài liệu tham khảo ở cuối mỗi chương có thể được dùng để tìm kiếm các tài liệu nghiên cứu chứa nguồn gốc của các khái niệm và phương pháp được trình bày, những cách xử lý sâu hơn về các chủ đề liên quan, cũng như các hướng mở rộng có thể có.

## Đối với sinh viên

Chúng tôi hy vọng cuốn giáo trình này sẽ khơi gợi sự quan tâm của bạn đối với lĩnh vực khai phá dữ liệu, một lĩnh vực trẻ nhưng phát triển nhanh chóng. Chúng tôi đã cố gắng trình bày nội dung một cách rõ ràng, với những giải thích cẩn thận về các chủ đề được đề cập. Mỗi chương đều kết thúc với một bản tóm tắt nêu ra những điểm chính. Chúng tôi đã đưa nhiều hình vẽ và minh họa trong suốt cuốn sách để làm cho nội dung trở nên thú vị và thân thiện với người đọc. Mặc dù cuốn sách này được thiết kế như một giáo trình, nhưng chúng tôi cũng cố gắng tổ chức nó sao cho nó có thể hữu ích như một cuốn tài liệu tham khảo hoặc cẩm nang, nếu sau này bạn quyết định nghiên cứu sâu hơn trong các lĩnh vực liên quan hoặc theo đuổi sự nghiệp trong khai phá dữ liệu.

Những kiến thức cần có để đọc cuốn sách này:

- Bạn nên có một số kiến thức cơ bản về các khái niệm và thuật ngữ liên quan đến thống kê, hệ thống cơ sở dữ liệu và học máy. Tuy nhiên, chúng tôi cố

gắng cung cấp đủ nền tảng cơ bản, để nếu bạn không quen thuộc lắm với các lĩnh vực này hoặc có chút quên, bạn sẽ không gặp khó khăn khi theo dõi các nội dung trong sách.

- Bạn nên có một số kinh nghiệm lập trình. Cụ thể, bạn cần có khả năng đọc giả mã và hiểu các cấu trúc dữ liệu đơn giản như mảng nhiều chiều và cấu trúc dữ liệu khác.

## Đối với chuyên gia

Cuốn sách này được thiết kế để bao quát một loạt các chủ đề trong lĩnh vực khai phá dữ liệu. Do đó, nó là một cẩm nang xuất sắc về chủ đề này. Vì mỗi chương được thiết kế sao cho độc lập nhất có thể, bạn có thể tập trung vào những chủ đề mà bạn quan tâm nhất. Cuốn sách có thể được sử dụng bởi các lập trình viên ứng dụng, các nhà khoa học dữ liệu và các quản lý dịch vụ thông tin muốn tự mình tìm hiểu những ý tưởng chủ chốt của khai phá dữ liệu. Cuốn sách cũng hữu ích cho các nhân viên phân tích dữ liệu kỹ thuật trong các ngành ngân hàng, bảo hiểm, y tế và bán lẻ, những người có hứng thú trong việc áp dụng các giải pháp khai phá dữ liệu vào doanh nghiệp của họ. Hơn nữa, cuốn sách có thể được xem như một tổng quan toàn diện về lĩnh vực khai phá dữ liệu, qua đó cũng mang lại lợi ích cho các nhà nghiên cứu muốn nâng cao trình độ hiện tại và mở rộng phạm vi ứng dụng của khai phá dữ liệu.

Các kỹ thuật và thuật toán được trình bày trong cuốn sách có tính ứng dụng cao. Thay vì lựa chọn các thuật toán chỉ hoạt động tốt trên các tập dữ liệu nhỏ mang tính "thử nghiệm", các thuật toán được mô tả trong sách nhằm mục đích khám phá các mẫu và tri thức ẩn chứa trong các tập dữ liệu lớn và thực tế. Các thuật toán được trình bày dưới dạng giả mã giả. Giả mã này tương tự với ngôn ngữ lập trình C, tuy nhiên được thiết kế sao cho dễ dàng theo dõi đối với những lập trình viên không quen thuộc với C hay C++. Nếu bạn muốn triển khai bất kỳ thuật toán nào, việc chuyển mã giả sang ngôn ngữ lập trình mà bạn lựa chọn sẽ là một công việc khá đơn giản.

## Trang web của sách cùng với tài nguyên

Cuốn sách có trang web tại địa chỉ

Đang cập nhật!

Trang web này chứa nhiều tài liệu bổ sung dành cho đọc giả cuốn sách hoặc bất kỳ ai có hứng thú với khai phá dữ liệu. Các tài nguyên bao gồm:

- **Slide bài giảng cho mỗi chương**. Ghi chú bài giảng dưới dạng slide Microsoft PowerPoint có sẵn cho mỗi chương.
- **Sổ tay cho giảng viên**. Bộ đáp án hoàn chỉnh cho các bài tập trong sách chỉ dành riêng cho giảng viên và được cung cấp trên trang web của nhà xuất bản.
- **Hình minh họa từ sách**. Điều này giúp bạn tạo slide riêng cho việc giảng dạy.
- **Mục lục** của sách ở định dạng PDF.
- **Danh sách sửa lỗi của sách**. Chúng tôi khuyến khích bạn chỉ ra bất kỳ lỗi nào trong cuốn sách. Khi lỗi được xác nhận, chúng tôi sẽ cập nhật danh sách sửa lỗi và ghi nhận sự đóng góp của bạn.

Ngoài ra, đọc giả quan tâm có thể truy cập các **trang web giảng dạy của các tác giả**. Tất cả các tác giả đều là giảng viên đại học. Hãy kiểm tra các trang web khóa học khai phá dữ liệu tương ứng của họ, nơi có thể chứa tài liệu cho các khóa học đại học và/hoặc sau đại học về khai phá dữ liệu, bao gồm slide bài giảng, giáo trình, bài tập về nhà, bài tập lập trình, dự án nghiên cứu, danh sách sửa lỗi và các thông tin liên quan khác.

# Lời cảm ơn

## Ấn bản thứ nhất của cuốn sách

# Về tác giả

**Nguyễn Đức Thịnh** hiện là Thạc sĩ ngành Toán ứng dụng thuộc Bộ môn Toán ứng dụng, Khoa Công nghệ Thông tin, tại Trường Đại học Xây dựng Hà Nội. Ông có kinh nghiệm giảng dạy trong các khóa học sau:

1) Xác suất thống kê
2) Toán học tính toán (Phương pháp số)
3) Toán rời rạc
4) Đại số hiện đại ứng dụng
5) Toán kinh tế
6) Nguyên lý ngôn ngữ lập trình
7) Học máy
8) Khai phá dữ liệu
9) Đại số tuyến tính (Hướng dẫn sinh viên thi Olympic toán học toàn quốc)