Biên soạn: TS. ĐỖ THANH NGHỊ - TS. PHẠM NGUYÊN KHANG

GIÁO TRÌNH

NGUYÊN LÝ MÁY HỌC

NHÀ XUẤT BẢN ĐẠI HỌC CẦN THƠ
2012

BIÊN MỤC TRƯỚC XUẤT BẢN THỰC HIỆN BỞI
TRUNG TÂM HỌC LIỆU TRƯỜNG ĐẠI HỌC CẦN THƠ

Đỗ, Thanh Nghị
Giáo trình nguyên lý máy học / Đỗ Thanh Nghị, Phạm Nguyên Khang .- Cần Thơ : Nxb. Đại học Cần Thơ, 2012
154 tr. : minh họa ; 24 cm
Sách có danh mục tài liệu tham khảo

1. Machine learning 2. Image processing
3. Nguyên lý máy học 4. Giải thuật máy học

I. Nhan đề II. Phạm, Nguyên Khang

006.31 - DDC 22
Ngh300 MFN 175317

LỜI GIỚI THIỆU

Nhằm góp phần làm phong phú thêm nguồn tư liệu phục vụ nghiên cứu, học tập cho bạn đọc trong và ngoài ngành Công nghệ thông tin, Nhà Xuất bản Đại học Cần Thơ xin được phép ấn hành và giới thiệu cùng bạn đọc quyển sách giáo trình "Nguyên lý máy học" do TS. Đỗ Thanh Nghị và TS. Phạm Nguyên Khang biên soạn. Giáo trình bao gồm 6 chương với 154 trang; được tổ chức thành 4 phần một cách có hệ thống. Phần 1 giới thiệu khái quát về máy học và nguyên lý học thống kê. Phần 2 trình bày các giải thuật học có giám sát tiêu biểu. Phần 3 trình bày các giải thuật học không giám sát. Phần 4 giới thiệu các ứng dụng phổ biến của máy học. Nội dung các chương giới thiệu nguyên lý học thống kê, mạng nơ-ron nhân tạo, máy học véc-tơ hỗ trợ (SVM), phương pháp đánh giá mô hình phân lớp, các giải thuật gom nhóm dữ liệu như k-*means*, bản đồ tự tổ chức (SOM), cực đại hóa kỳ vọng (EM) và ứng dụng nhận dạng ký tự số viết tay, phân lớp tự động văn bản. Thêm vào đó, cuối mỗi chương còn có rất nhiều bài tập và tài liệu tham khảo hữu ích cho bạn đọc. Giáo trình là tài liệu tham khảo có giá trị cho sinh viên ngành Công nghệ thông tin và bạn đọc muốn tìm hiểu và nghiên cứu về máy học.

Nhà Xuất bản Đại học Cần Thơ chân thành cám ơn các Tác giả và sự đóng góp ý kiến của quý Thầy Cô trong Hội đồng thẩm định trường Đại học Cần Thơ để giáo trình "Nguyên lý máy học" được ra mắt bạn đọc.

Nhà Xuất bản Đại học Cần Thơ trân trọng giới thiệu đến giảng viên, sinh viên và bạn đọc giáo trình này.

Chân thành cám ơn!

NHÀ XUẤT BẢN ĐẠI HỌC CẦN THƠ

LỜI NÓI ĐẦU

Đã trải qua nhiều thập kỷ, chúng ta chứng kiến sự phát triển của ngành khoa học máy tính. Xuất phát điểm từ công nghệ chế tạo phần cứng, lập trình cấp thấp, đến công nghệ thông tin, tổ chức, quản lý và xử lý hiệu quả hệ thống thông tin. Hiện nay, ngành khoa học máy tính đang bước vào thời kỳ công nghệ tri thức, ở mức trừu tượng cao hơn, làm máy tính trở nên thông minh hơn, để có thể giúp con người giải quyết nhiều vấn đề phức tạp trong thực tiễn. Chẳng hạn như máy học là hướng tiếp cận nhằm phát triển các kỹ thuật cho phép các máy tính có thể học từ dữ liệu để nhận dạng các mẫu phức tạp và đưa ra các quyết định thông minh. Máy học có tính ứng dụng rất cao trong thực tế như chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và điều khiển robot. Chính vì vậy, nguyên lý máy học trở thành môn học chuyên ngành rất quan trọng của chương trình đào tạo cử nhân, kỹ sư và thạc sĩ công nghệ thông tin. Trong thời gian giảng dạy vừa qua, chúng tôi nhận thấy có quá ít tài liệu tham khảo trình bày một cách có hệ thống về nguyên lý máy học. Chính vì lý do đó, chúng tôi đã thực hiện biên soạn quyển giáo trình "Nguyên lý máy học" nhằm mục đích cung cấp thêm tài liệu tham khảo bằng tiếng Việt đến độc giả là sinh viên chuyên ngành công nghệ thông tin.

Giáo trình được soạn dựa trên kiến thức, kinh nghiệm học tập nghiên cứu, giảng dạy của chúng tôi trong suốt từ năm 2000 đến nay và các tài liệu tham khảo chính được trình bày trong mục tài liệu tham khảo của từng chương. Chúng tôi cung cấp các kiến thức cơ bản và các giải thích nhằm đơn giản việc trình bày các giải thuật học. Cuối mỗi chương có bài tập giúp độc giả ôn tập lại kiến thức của chương và trao dồi kỹ năng thực hành. Giáo trình được tổ chức thành sáu chương:

Chương 1 tập trung vào việc giới thiệu các kiến thức cơ bản của máy học. Nội dung bao gồm khái niệm học có giám sát, học không giám sát, nguyên lý học thống kê được xem là nền tảng của các kỹ thuật học từ dữ liệu.

Chương 2 giới thiệu kiến thức cơ bản về giải thuật học có giám sát mạng nơ-ron nhân tạo. Chúng tôi trình bày về mô hình mạng nơ-ron của McCulloch và Pitts, kiến trúc mạng nơ-ron, giải thuật huấn luyện mạng Perceptron, mạng nơ-ron đa tầng cho vấn đề phân lớp.

Chương 3 giới thiệu về giải thuật học có giám sát máy học véc-tơ hỗ trợ cho vấn đề phân lớp, hồi quy và phát hiện phần tử cá biệt. Nội dung bắt đầu từ việc mô hình hóa bài toán đến giải thuật máy học véc-tơ hỗ trợ. Chương được kết thúc với phần thảo luận về giải thuật máy học véc-tơ hỗ trợ.

Chương 4 cung cấp cho sinh viên về phương pháp đánh giá hiệu quả của giải thuật học có giám sát. Các nghi thức kiểm tra và tiêu chí để so sánh hiệu quả của giải thuật học.

Chương 5 tập trung trình bày các giải thuật học không giám sát như giải thuật gom nhóm k-means, bản đồ tự tổ chức và giải thuật cực đại hóa kỳ vọng. Chúng tôi giới thiệu kiến thức cơ bản từ việc mô hình hóa bài toán đến giải thuật học cho gom nhóm.

Chương 6 giới thiệu hai ứng dụng phổ biến trong thực tế của máy học là nhận dạng chữ viết tay và phân loại văn bản tự động với mạng nơ-ron, máy học véc-tơ hỗ trợ.

Nội dung giáo trình được giảng dạy cho sinh viên kỹ sư, thạc sĩ công nghệ thông tin trong thời lượng 30 tiết lý thuyết. Giảng viên có thể phân bố thời gian dạy lý thuyết kết hợp với các ứng dụng minh họa của máy học. Chương 1, 2, 3, 4 có thể được giảng dạy trong 18 tiết lý thuyết. Chương 5 có thể được trình bày trong 10 tiết lý thuyết, và thời gian còn lại là của chương 6. Nếu giảng dạy cho sinh viên kỹ sư, cần giảm bớt hai giải thuật bản đồ tự tổ chức và cực đại kỳ vọng cho gom nhóm.

Trong thời gian biên soạn tài liệu, chúng tôi nhận được sự hỗ trợ quý báu về vật chất, tinh thần từ Khoa Công Nghệ Thông Tin và Truyền Thông, Trường Đại Học Cần Thơ. Chúng tôi đặc biệt gửi lời cám ơn đến Quý Thầy Cô thuộc Bộ Môn Khoa Học Máy Tính, Hội Đồng Thẩm Định Giáo Trình, đã nhiệt tình góp ý cho bản thảo.

Do thời gian có hạn và lần đầu tiên biên soạn, tài liệu không thể tránh khỏi những thiếu sót. Chúng tôi mong nhận được sự góp ý chân thành từ quý độc giả để quyển sách ngày được hoàn thiện hơn.

Cần Thơ, tháng 6 năm 2012

Đỗ Thanh Nghị - Phạm Nguyên Khang

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

THUẬT NGỮ

Machine learning: máy học

Supervised learning: học có giám sát

Unsupervised learning: học không giám sát

Classification: phân lớp (học có giám sát)

Regression: hồi quy

Clustering: gom cụm, gom nhóm (học không giám sát)

Minimizing a risk functional: cực tiểu rủi ro

Empirical risk minimization: cực tiểu rủi ro thực nghiệm

Empirical inference: suy diễn thực nghiệm

Vapnik-Chervonenkis dimension: chiều VC

Structural risk minimization induction principle: Nguyên lý quy nạp cực tiểu rủi ro cấu trúc

Neural network: mạng nơ-ron nhân tạo

MLP (multilayer perceptron): mạng perceptron đa tầng

Convolutional neural network: mạng nơ-ron tích chập

Overfitting: học vẹt

Underfitting: học không thuộc bài

SVM (Support vector machines): máy học véctơ hỗ trợ

Hyperplane: siêu phẳng

Kernel: hàm nhân

Feature space: không gian trung gian hay không gian đặc trưng

Density function estimation: ước lượng hàm mật độ xác suất

SOM (Self-organizing map): bản đồ tự tổ chức

MLE (maximum likelihood estimation): phương pháp hợp lý cực đại

EM (Expectation maximization): cực đại hóa kỳ vọng

Decision tree: cây quyết định

Naive Bayes: Bayes thơ ngây

kNN (k nearest neighbors): k láng giềng

Mẩu tin còn được gọi là điểm dữ liệu, phần tử. Thuộc tính dữ liệu cũng là trường dữ liệu, biến hay chiều dữ liệu. Một lớp của dữ liệu thường được gán một nhãn hay còn gọi là nhãn.

PHẦN I

GIỚI THIỆU

- *KHÁI NIỆM MÁY HỌC*
- *NGUYÊN LÝ HỌC THỐNG KÊ*

Chương 1

GIỚI THIỆU

Máy học, có tài liệu gọi là học máy (machine learning), là một lĩnh vực của trí tuệ nhân tạo, liên quan đến việc phát triển các kỹ thuật cho phép các máy tính có thể *học* từ dữ liệu. Máy học là tiếp cận để tạo ra các chương trình máy tính có thể học tự động từ dữ liệu để nhận dạng các mẫu phức tạp và đưa ra các quyết định thông minh. Vấn đề khó khăn có thể thấy ở đây là đầu vào của tập hành vi ứng xử là rất lớn, nhưng chúng ta lại có tập dữ liệu quan sát để máy học là hữu hạn, không thể phủ được toàn bộ tập hành vi ứng xử. Chương trình máy học phải tổng quát từ dữ liệu hữu hạn cho trước, có thể sinh ra hành vi ứng xử thích hợp cho những trường hợp (dữ liệu) mới đến. Máy học có tính ứng dụng rất cao trong thực tế như tìm kiếm thông tin, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA, nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và điều khiển robot.

Trong chương này, chúng tôi trình bày các khái niệm cơ bản trong máy học. Trước tiên, chúng tôi giới thiệu ngắn gọn về máy học. Tiếp đến, các vấn đề chính trong máy học bao gồm học có giám sát (supervised learning, ví dụ như phân lớp, hồi quy) và học không giám sát (unsupervised learning, ví dụ như ước lượng mật độ, gom nhóm dữ liệu). Chúng tôi sẽ trình bày nguyên lý học thống kê [Vapnik, 1995], [Vapnik, 1999], là nền tảng nguyên lý học máy, nguyên lý cực tiểu rủi ro thực nghiệm, yếu tố quan trọng trong học thống kê, chiều VC (Vapnik-Chervonenkis dimension), tại sao mạng nơ-ron và máy học SVM là giải thuật học tổng quát nhất.

1.1 KHÁI NIỆM CƠ BẢN MÁY HỌC

Máy học là chương trình máy tính cho phép học tự động từ dữ liệu để nhận dạng các mẫu phức tạp, tạo ra hành vi ứng xử thông minh với trường hợp mới đến. [Mitchell, 1997] định nghĩa về học máy như sau: một chương trình máy tính được gọi là học từ kinh nghiệm E với một vài lớp của vấn đề T và độ đo hiệu quả P, nếu hiệu năng của vấn đề trong T, đánh giá theo tiêu chí P, được cải thiện từ kinh nghiệm E.

Kinh nghiệm E chính là dữ liệu thực nghiệm (hay còn gọi là ví dụ). Lớp vấn đề trong máy học có thể phân thành hai nhóm chính là học có giám sát (supervised learning) và học không giám sát (unsupervised learning).

Chúng ta xét một ví dụ sau đây về máy học có giám sát. Giả sử người ta cần xây dựng mô hình học tự động từ dữ liệu, để có thể dự báo chơi golf hay không. Theo đó, dữ liệu weather [Mitchell, 1997], dùng cho máy học (như bảng 1.1), có các thuộc tính outlook (quang cảnh), temperature (nhiệt độ), humidity (độ ẩm), windy (gió) và Play (lớp dự báo: chơi golf hay không chơi cho từng trường hợp cụ thể). Một giải thuật học A dùng tập dữ liệu weather để học mô

hình dự báo *H*, sao cho một ngày nào đó có thời tiết *[outlook=rainy, temp=73, humidity=90, windy=FALSE]* thì mô hình *H* có thể dự báo được có chơi golf hay không. Hơn nữa, giải thuật *A* có thể cải tiến mô hình *H* dựa trên tập dữ liệu học để có thể cho kết quả dự báo tốt.

Bảng 1.1: Tập dữ liệu weather

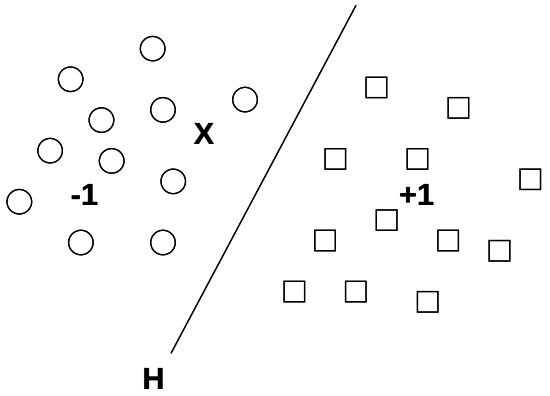
Outlook	Temp.	Humidity	Windy	Play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	Yes
sunny	75	70	TRUE	Yes
overcast	72	90	TRUE	Yes
overcast	81	75	FALSE	Yes
rainy	71	91	TRUE	no

Trong giải thích các khái niệm máy học tiếp theo sau đây, chúng ta thống nhất dùng thuật ngữ tập dữ liệu có *m* phần tử $\{x_1, x_2, ..., x_m\}$, với $x_i \in R^n$, là các véc-tơ *n* biến (thuộc tính, chiều) ngẫu nhiên, độc lập, phân phối đồng nhất (independent and identically distributed, viết tắt là i.i.d.).

Học có giám sát: thuật toán học tạo ra một hàm ánh xạ dữ liệu đầu vào tới kết quả đích mong muốn (nhãn, lớp, giá trị cần dự báo). Trong học có giám sát, tập dữ liệu dùng để huấn luyện phải được gán nhãn, lớp hay giá trị cần dự báo.

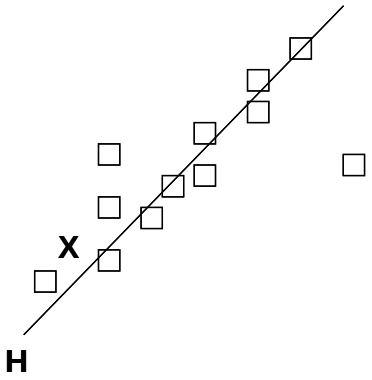
Xét bài toán phân lớp (classification) xây dựng mô hình *H* được huấn luyện từ tập dữ liệu $\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ (với $x_i \in R^n$, có nhãn (lớp) tương ứng y_i rời rạc, ví dụ $y_i \in \{\pm 1\}$), tách lớp, để khi phần tử *x* mới đến, mô hình *H* có thể gán nhãn (phân lớp) *x* vào một trong số lớp cho trước (ví dụ như nhãn của *x* được *H* gán là *-1*, xem hình 1.1).

Vấn đề học từ tập dữ liệu weather để dự báo chơi golf hay không là một ví dụ về bài toán phân lớp (trong trường hợp này, y_i là rời rạc, 2 lớp là play = yes và play = no).



***Hình 1.1**: Bài toán phân lớp*

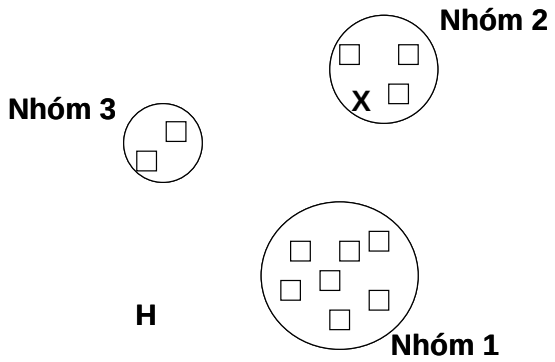
Bài toán hồi quy (regression), tương tự như bài toán phân lớp, xây dựng mô hình H được huấn luyện từ tập dữ liệu $\{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ (với $x_i \in R^n$, tương ứng y_i là giá trị dự báo là liên tục, $y_i \in R^1$), đi qua dữ liệu sao cho sai khác là nhỏ nhất, để khi phần tử x mới đến, mô hình H có thể dự báo x là một giá trị (ví dụ như x được H giá trị dự báo là khoảng cách từ x đến H như xem hình 1.2).



***Hình 1.2**: Bài toán hồi quy*

Để minh họa cho bài toán hồi quy, chúng ta có thể xét vấn đề dự báo lượng mưa của tiểu vùng sông Mê kông. Với tập dữ liệu thu thập bao gồm các biến như nhiệt độ, bức xạ mặt trời, độ ẩm, hướng gió, tốc độ gió, lượng mưa. Giải thuật học A cần học mô hình dự báo H để có thể dự báo lượng mưa (trong trường hợp này, y_i là số liên tục do lượng mưa cần dự báo là bao nhiêu mm) của ngày nào đó khi biết được nhiệt độ, bức xạ mặt trời, độ ẩm, hướng gió và tốc độ gió của ngày đó.

Học không giám sát: thuật toán học thực hiện mô hình hóa một tập dữ liệu đầu vào, không được gán nhãn (lớp, giá trị cần dự báo). Xét bài toán gom nhóm dữ liệu, xây dựng mô hình *H* được huấn luyện từ tập dữ liệu $\{x_1, x_2, ..., x_m\}$ (với $x_i \in R^n$), sao cho hai phần tử dữ liệu của cùng nhóm phải có tính chất rất giống nhau hơn là hai phần tử thuộc hai nhóm khác nhau. Ví dụ như hình 1.3, xây dựng mô hình *H* gom nhóm dữ liệu thành ba nhóm.



***Hình 1.3**: Bài toán gom nhóm dữ liệu*

Ví dụ như chúng ta cần gom nhóm sinh viên có các đặc trưng tương tự nhau như sở thích, kết quả học tập, trình độ ngoại ngữ, tính năng động.

Ngoài ra, còn có lớp bài toán như:

Học bán giám sát (semi-supervised learning): kết hợp dữ liệu có gán nhãn và không gán nhãn để tạo một bộ phân lớp.

Học tăng cường (reinforcement learning): thuật toán học một ứng xử dựa vào quan sát về thế giới. Mỗi hành động đều có tác động tới môi trường và môi trường cung cấp thông tin phản hồi để hướng dẫn cho thuật toán trong quá trình học.

1.2 NGUYÊN LÝ HỌC THỐNG KÊ

Để nghiên cứu nền tảng lý thuyết của nguyên lý máy học, chúng tôi trình bày nguyên lý học thống kê [Vapnik, 1995], [Vapnik, 1999]. Trước tiên, chúng ta sẽ làm quen với mô hình ước lượng hàm, cực tiểu hóa rủi ro, đây là khung tổng quát của các bài toán học máy như phân lớp, hồi quy, ước lượng mật độ dữ liệu.

1.2.1 Mô hình ước lượng hàm

Mô hình máy học từ dữ liệu có thể được mô tả bởi ba bộ phận như sau.

1. Bộ sinh các véc-tơ ngẫu nhiên x, độc lập, cùng phân phối xác suất cố định nhưng chưa được biết trước $P(x)$.
2. Một giám sát (supervisor) trả về ở đầu ra một véc-tơ y khi nhận ở đầu vào véc-tơ x, theo hàm phân phối xác suất có điều kiện $P(y|x)$ cố định nhưng chưa được biết trước.
3. Một máy học có thể cài đặt một tập hàm $f(x,\alpha)$, với $\alpha \in \Lambda$

   (Λ là tập hợp các tham số, α không nhất thiết là một véc-tơ, chính vì vậy $f(x,\alpha)$ được dùng ở đây là tập hàm bất kỳ).

Vấn đề học máy chính là tìm kiếm (chọn lựa) từ tập hàm cho trước $f(x,\alpha)$, $\alpha \in \Lambda$, một hàm mà sao cho kết quả dự báo thu được gần nhất với kết quả của giám sát. Sự lựa chọn hàm dựa vào tập dữ liệu huấn luyện có m phần tử:

$$(x_1, y_1), (x_2, y_2), ..., (x_m, y_m) \tag{1.1}$$

với $x_i \in R^n$, là các véc-tơ n biến (thuộc tính, chiều), có nhãn (lớp) tương ứng y_i, và $P(x,y) = P(x)P(y|x)$.

1.2.2 Cực tiểu hóa rủi ro (minimizing a risk functional)

Xem xét vấn đề lựa chọn hàm xấp xỉ tốt nhất cho dự báo của giám sát, cũng chính là độ đo sự sai khác (loss) $L(y, f(x,\alpha))$ giữa giá trị thực, cần được dự báo y của giám sát (kết quả mong đợi, đích) và giá trị dự báo $f(x,\alpha)$ thu được từ mô hình máy học khi nhận đầu vào x. Sự sai khác được tính thông qua rủi ro $R(\alpha)$ như sau:

$$R[f] = \int L(y, f(x,\alpha)) dP(x,y) \tag{1.2}$$

Mục tiêu của máy học là tìm hàm $f(x,\alpha^*)$ từ lớp hàm $f(x,\alpha)$, $\alpha \in \Lambda$, để cực tiểu rủi ro $R[f]$ như công thức 1.2 trong khi chưa biết về $P(x,y)$, chỉ có được tập dữ liệu trong (1.1). Nguyên lý cực tiểu hóa rủi ro dựa vào hàm sai khác (loss) là khung tổng quát cho cả ba bài toán phân lớp, hồi quy và ước lượng mật độ dữ liệu.

Bài toán phân lớp:

Tập dữ liệu $\{x, y\} = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ (với $x_i \in R^n$, có nhãn (lớp) tương ứng $y_i \in \{\pm 1\}$), lớp hàm $f(x,\alpha)$ với $\alpha \in \Lambda$, hàm sai khác (loss):

$$L(y, f(x,\alpha)) = \begin{cases} 0 & y = f(x,\alpha) \\ 1 & y \neq f(x,\alpha) \end{cases} \tag{1.3}$$