

MACHINE LEARNING

Bộ môn Tin học – Khoa Toán Cơ Tin học

Cao Văn Chung

caovanchung@hus.edu.vn

Tổng quan môn học

Nội dung - Tổ chức môn học

Giới thiệu

Phân loại các mô hình Học máy

Học có giám sát

Học phi giám sát

Học bán giám sát

Học tăng cường

Đánh giá một mô hình học máy

Tài liệu tham khảo

Nội dung môn học

- Tổng quát về học máy (Machine Learning) và kĩ thuật nhận dạng mẫu (Pattern Recognition) sử dụng phương pháp thống kê.
- Các chủ đề chính gồm:
 - Học có giám sát (Supervised learning);
 - Học không có giám sát (Unsupervised learning);
 - Các vấn đề lý thuyết liên quan (Vapnik–Chervonenkis theory...);
 - Sự phát triển và các hướng nghiên cứu hiện tại của ngành học máy và ứng dụng.

Kiến thức tiên quyết

- Cấu trúc dữ liệu và thuật toán (MAT3514)
- Xác suất thống kê (MAT2323)
- Giải tích số (MAT2034)
- Đại số tuyến tính (MAT2400)
- Giải tích đa biến
- Lập trình với một ngôn ngữ bậc cao (Java, Python, C/C++ ...)

Tổ chức học tập

- 03 tín chỉ (2*15 giờ lý thuyết + 2*15 giờ thực hành & bài tập)
- Kiểm tra – đánh giá
 - Thường xuyên (20%):
 - Bài tập về nhà
 - Kiểm tra trắc nghiệm
 - Câu hỏi của giáo viên
 - Giữa kỳ (20%): Kiểm tra trắc nghiệm/lập trình và vấn đáp.
 - Hết môn (60%):
 - Bài tập lớn theo nhóm (tối đa 3 thành viên) với các đề tài do giáo viên cung cấp

Tài liệu

Tài liệu tham khảo

- Bài giảng do giáo viên cung cấp
- Tom Mitchell (1997), *Machine Learning*, McGraw-Hill
- Kevin P. Murphy (2012), *Machine Learning: A Probabilistic Perspective*, The MIT Press
- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2016) , *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition, Springer Series in Statistics
- Ian Goodfellow, Yoshua Bengio, Aaron and Courville (2016), *Deep Learning*, MIT Press
- Christopher Bishop (2006), *Pattern Recognition and Machine Learning*, Springer.
- Richard Duda, Peter Har, David Stork (2001), *Pattern Classification*, 2nd Edition, John Wiley & Sons.

Giới thiệu

Học máy là gì?

Các lĩnh vực liên quan và ứng dụng

Phân loại các dạng học máy

Các mô hình sinh và phân biệt

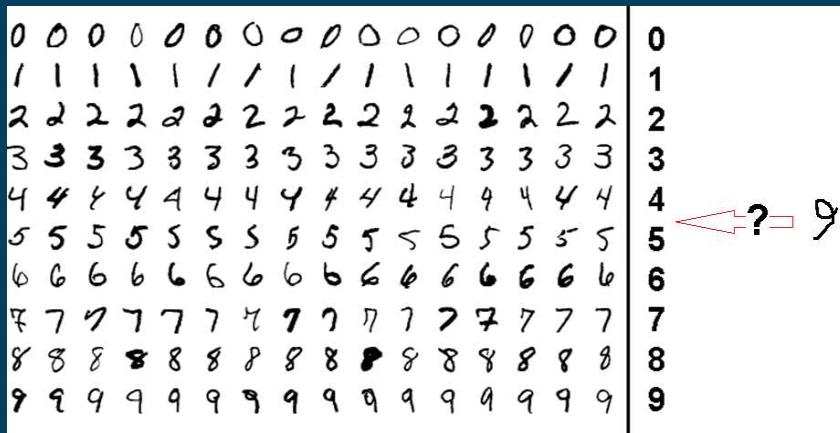
7

Machine learning - AI

- Machine Learning là lĩnh vực của Trí tuệ nhân tạo (Artificial Intelligence – AI):
 - AI: có mục đích làm cho máy tính có được trí thông minh
 - Machine Learning: làm cho chương trình máy tính có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể.
- AI – Mục tiêu & Machine learning – Phương tiện:
 - Nghiên cứu và xây dựng các kĩ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.
- Ví dụ Trí tuệ nhân tạo (Artificial Intelligence – AI) và Machine Learning

Machine learning - AI

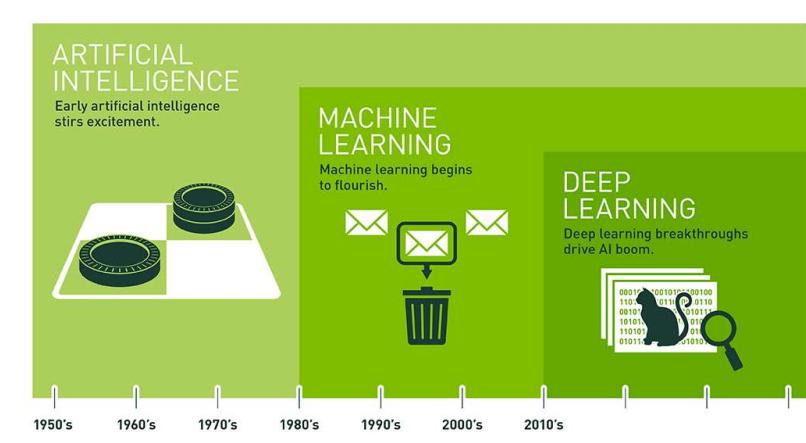
- Ví dụ: “Học” để nhận dạng chữ số viết tay



MACHINE LEARNING - 2024

9

Machine learning - AI



MACHINE LEARNING - 2024

10

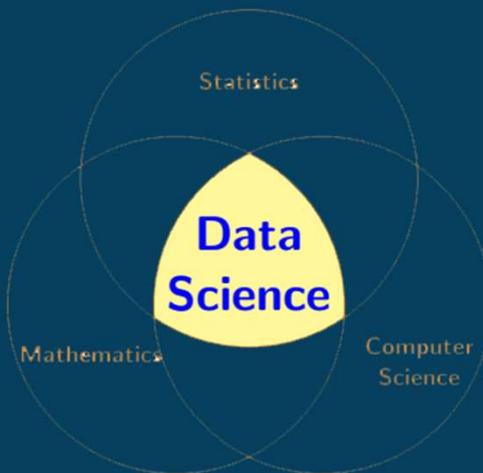
Machine learning - Data Science

- Machine Learning khai thác dữ liệu (thông tin) và tạo thành tri thức. ML – trung tâm của khoa học dữ liệu (Data Science)
- Data Science (Khoa học dữ liệu):
 - Một lĩnh vực liên ngành sử dụng các phương pháp khoa học, quy trình, thuật toán và hệ thống để chiết xuất kiến thức và hiểu biết sâu sắc từ dữ liệu ở nhiều dạng khác nhau, cả có cấu trúc và không có cấu trúc.
 - “The sexiest job of the 21st century” (Harvard Business Review, 2012)
 - Khoa học dữ liệu sử dụng các kỹ thuật và lý thuyết từ nhiều lĩnh vực.

Machine learning - Data Science

- Data Science (Khoa học dữ liệu):
 - Data mining (Khai phá dữ liệu): tìm kiếm những thông tin, tri thức hoàn toàn mới tiềm năng có ích trong nguồn dữ liệu.
 - Machine Learning (Học máy): dự đoán một số thông tin của dữ liệu dựa trên những đặc tính đã biết.
- Big Data
- Machine Learning vs. Big Data:
 - Machine learning phát triển hơn nhờ sự gia tăng của khối lượng dữ liệu của Big Data;
 - Giá trị của Big Data phụ thuộc vào khả năng khai thác tri thức từ dữ liệu của machine learning.

Machine learning - Data Science



MACHINE LEARNING - 2024

13

Machine learning

- Tom Mitchell, Prof. Of. Carnegie Mellon University (CMU):

"Một chương trình máy tính CT được xem là học cách thực thi một lớp nhiệm vụ NV thông qua trải nghiệm KN, đối với thang đo năng lực NL nếu như dùng NL ta đo thấy năng lực thực thi của chương trình có tiến bộ sau khi trải qua KN" (máy đã học).

- Nó cũng kéo theo:
 - Machine learning làm cho chương trình máy tính có khả năng tự học hỏi dựa trên dữ liệu đưa vào mà không cần phải được lập trình cụ thể, tường minh.
 - Machine learning chuyển đổi dữ liệu, thông tin thành tri thức.

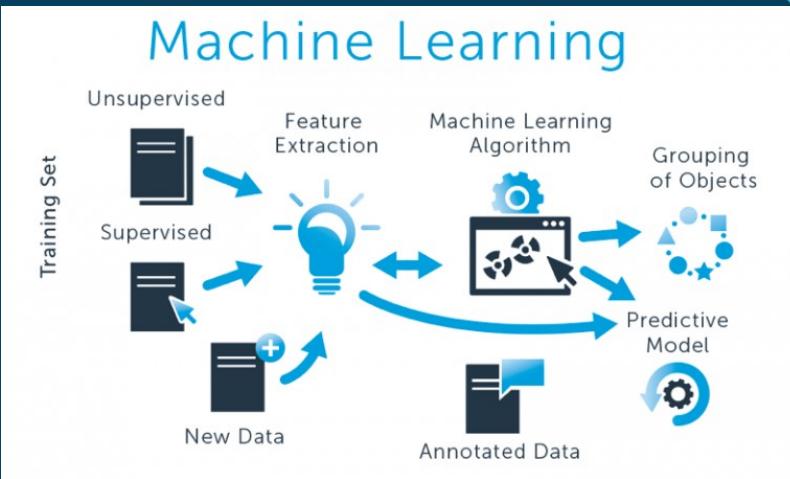
MACHINE LEARNING - 2024

14

Machine learning

- Mục tiêu của Machine Learning (trước mắt):
 - Làm cho máy tính có những khả năng nhận thức cơ bản của con người như nghe, nhìn, hiểu được ngôn ngữ, giải toán, ...
 - Hỗ trợ con người trong việc xử lý một khối lượng thông tin khổng lồ mà chúng ta phải đối mặt hàng ngày, hay còn gọi là Big Data.
- Machine Learning cho phép dự báo/dự đoán dựa trên thông tin đã có.

Machine learning



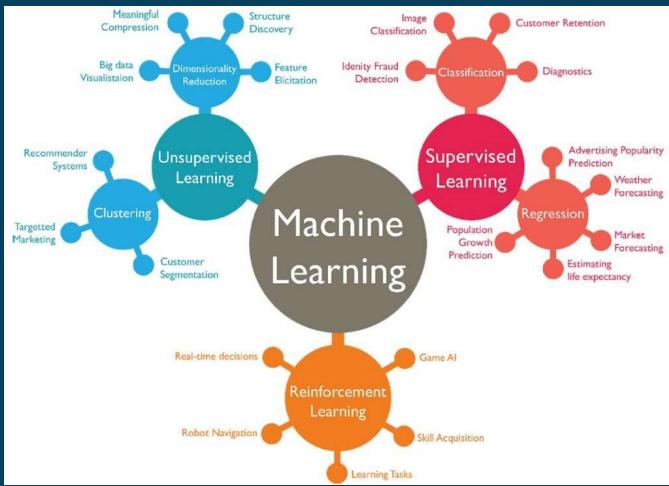
Các loại Machine learning

- Dựa vào phương pháp học
 - Supervised Learning (Học có giám sát): thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước.
 - Unsupervised Learning (Học không giám sát): Trong thuật toán này, chúng ta không biết được outcome hay nhãn mà chỉ có dữ liệu đầu vào.
 - Semi-Supervised Learning (Học bán giám sát): Các bài toán khi chúng ta có một lượng lớn dữ liệu nhưng chỉ một phần trong chúng được gán nhãn.
 - Reinforcement Learning (Học Cứng Cố - Tăng cường): Các thuật toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance).

Các loại Machine learning

- Dựa vào chức năng
 - Regression Algorithms
 - Classification Algorithms
 - Instance-based Algorithms
 - Regularization Algorithms
 - Bayesian Algorithms
 - Clustering Algorithms
 - Artificial Neural Network Algorithms
 - Dimensionality Reduction Algorithms
 - Ensemble Algorithms

Các loại Machine learning



MACHINE LEARNING - 2024

19

Học có giám sát

Khái niệm

Ví dụ

20

Supervised Learning

- Supervised learning là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước.
- Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Supervised Learning

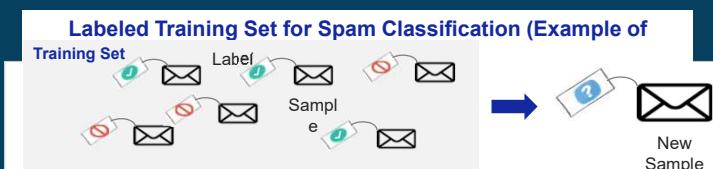
- Mô hình toán học
 - Cho tập dữ liệu vào: $X = \{x_1, x_2, x_3, \dots, x_N\}$ đã được gắn với nhãn $Y = \{y_1, y_2, y_3, \dots, y_N\}$
 - Các x_i, y_j đều là các vector
 - Các cặp tương ứng: $(x_i, y_i) \in X \times Y$ – gọi là dữ liệu huấn luyện (training data)
 - Dựa trên training data, cần điều chỉnh ánh xạ $f(x)$ từ X sang Y sao cho có xấp xỉ: $y_i \approx f(x_i)$ ($\forall i = 1, 2, \dots, N$)
- Mục đích: với dữ liệu mới x , ước lượng được y tương ứng phù hợp nhất qua $y = f(x)$.

Supervised Learning

- Ví dụ 1: Nhận dạng chữ số (latin) viết tay
 - $\{x_i\}$ là tập các ảnh của các chữ số viết tay đã được “gán nhãn”
 - $\{y_i\}$ là các “nhãn”, tức là tên các chữ số từ 0 đến 9, hoặc không phải chữ số.
 - Từ training data $\{(x_i, y_i)\}$, chương trình tìm ra liên hệ để với ảnh x đầu vào mới, chương trình tự xếp ứng với nhãn y_i nào (tức ký tự ở trong ảnh x là chữ số nào – hoặc không phải chữ số).
- Ví dụ 2: Nhận dạng khuôn mặt người
 - $\{x_i\}$ là tập các ảnh có khuôn mặt người (đã được “gán nhãn” – tức đã biết ứng với tên ai);
 - $\{y_i\}$ là các “nhãn”, tức là tên người ứng với khuôn mặt ở trong ảnh.
 - Từ training data $\{(x_i, y_i)\}$, chương trình tự xếp ứng với nhãn y_i nào (tức kết luận x chứa mặt người nào).

Supervised Learning

- **Classification** (Phân loại): các **label** của **input data** được chia thành một số hữu hạn nhóm. Ví dụ: email là spam hay không?; Khách hàng tín dụng có khả năng thanh toán nợ hay không.



- **Regression** (Hồi quy): Nếu label không được chia thành các nhóm mà là một giá trị thực (numeric) cụ thể.

- Từ training data (x_i, y_i) , tìm $f(x)$ phù hợp nhất: $y_i \approx f(x_i)$ với mọi i
 - Với đầu vào mới x , dự đoán/ước lượng đầu ra y qua phiến hàm: $y = f(x)$.

Supervised Learning

- Ví dụ 1 & Ví dụ 2 là dạng Classification (Phân loại)
- Ví dụ 3: Ước lượng tuổi của người qua ảnh khuôn mặt
 - $\{x_i\}$ là tập các ảnh có khuôn mặt người (đã được “gán nhãn” – tức đã biết ứng với bao nhiêu tuổi);
 - $\{y_i\}$ là các “nhãn”, tức là độ tuổi ứng với khuôn mặt ở trong các ảnh x_i .
 - Từ training data $\{(x_i, y_i)\}$, chương trình tự xếp một ảnh khuôn mặt người mới x ứng với nhãn y nào (tức kết luận người có khuôn mặt trong ảnh x sẽ là bao nhiêu tuổi).
 - Ví dụ 3 là dạng Regression (Hồi quy).

Các bài toán áp dụng học có giám sát

Bài toán hồi quy (Regression) &
Bài toán phân loại (Classification)

Các loại bài toán học máy có giám sát

- Bài toán hồi quy (Regression): Đầu ra là giá trị số
- Bài toán phân loại (Classification): Đầu ra là một nhãn phân loại

Classification (Class Label Prediction)	Regression (Numerical Prediction)
K-Nearest Neighbors	Linear Regression
Logistic Regression	Extended Regression Analysis (ex: Polynomial Regression, Nonlinear Regression, Penalized Regression, etc.)
Artificial Neural Network	Artificial Neural Network
Decision Tree	Decision Tree
Support Vector Machine	Support Vector Machine (Regression)
Naïve Bayes	PLS (Partial Least Squares)
Ensemble Method (Random Forest, etc.)	Ensemble Method (Random Forest, etc.)

MACHINE LEARNING - 2024

27

Học không giám sát

Khái niệm

Ví dụ

28

UnSupervised Learning

- UnSupervised learning:
 - Chỉ có tập dữ liệu vào X , không có tập nhãn.
 - Dựa vào tính chất/cấu trúc của X để thực hiện công việc
- Ví dụ:
 - Phân cụm dữ liệu (clustering);
 - Giảm số chiều của dữ liệu (dimension reduction)

UnSupervised Learning

- Các dạng phương pháp Unsupervised learning:
 - Clustering (phân nhóm):
 - Chia toàn bộ dữ liệu X thành các nhóm nhỏ dựa trên sự liên quan giữa các dữ liệu trong mỗi nhóm.
 - Clustering thường dựa vào tính “tương tự” của các phần tử x_i có trong X để đưa chúng vào các nhóm.
 - Association (tìm liên hệ):
 - Khám phá ra một quy luật dựa trên tập hợp nhiều dữ liệu cho trước.
 - Thường dựa trên liên hệ giữa các thuộc tính của các đối tượng dữ liệu để nghiên cứu

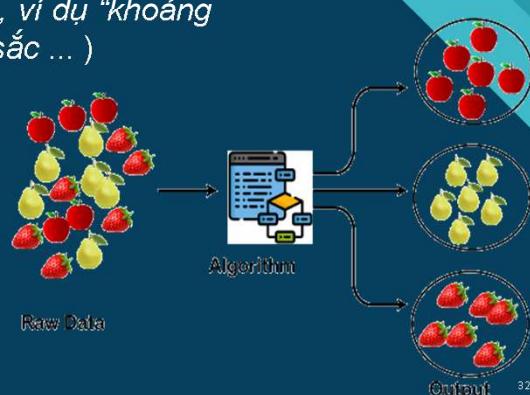
UnSupervised Learning

Cluster	Visualization and Dimension Reduction	Association Rule Learning
<ul style="list-style-type: none">• K-Means• DBSCAN• Hierarchical Clustering Analysis (HCA)• Anomaly Detection and Outlier Detection• One-Class SVM• Isolation Forest	<ul style="list-style-type: none">• Principal Component Analysis(PCA)• Kernel PCA• Local Linear Embedding• t-SNE	<ul style="list-style-type: none">• Apriori• Eclat

UnSupervised Learning

- Ví dụ 1: Clustering
 - $\{x_i\}$ là tập các đối tượng, với độ đo $SIM(x_i, x_j)$ – similarity $x_i \leftrightarrow x_j$
(SIM dựa vào cấu trúc/tính chất của $\{x_i\}$, ví dụ “khoảng cách”, chất liệu, hình dáng, giá trị, màu sắc ...)

- Dựa vào SIM để nhóm $\{x_i\}$ vào các tập con $S_j = \{x_{ij}\}$ với SIM của các phần tử trong mỗi nhóm đều “gần nhau”



UnSupervised Learning

• Ví dụ 2: Association

- $\{x_i\}$ là tập dữ liệu chứa thông tin mua sắm của khách hàng. Bản ghi ứng với mỗi lần mua hàng, chứa tên các mặt hàng;
- Bằng cách xét số lần xuất hiện của các cặp sản phẩm trong mỗi lần mua hàng => tìm được liên hệ các mặt hàng nào thường được mua cùng nhau.

• Ứng dụng:

- Gợi ý khách hàng mua sản phẩm dựa trên nhu cầu thường đi kèm với nhau
- Gợi ý từ tiếp theo cần dùng khi soạn thảo văn bản

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

market basket transactions

(Diapers, Beer) Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

MACHINE LEARNING - 2024

33

Supervised Learning vs. Unsupervised Learning

Supervised Learning

- **Training Data:** (x, y)
 $x \in X$ – Observation data;
 $y \in Y$ – Labels
- **Goal:**
 - Learn a function (mapping, model) f to map: $x \rightarrow y = f(x)$

UnSupervised Learning

- **Data:**
 - $x \in X$ – Observation data;
 - No labels – Just data
- **Goal:**
 - Learn some underlying hidden structure of data

MACHINE LEARNING - 2024

34

Semi-Supervised Learning (Học bán giám sát)

35

Semi-Supervised Learning

- Semi-Supervised Learning:
 - Có một lượng lớn dữ liệu X nhưng chỉ một phần trong chúng được gán nhãn.
 - Nhóm này nằm giữa hai nhóm đã được nêu trong phần trước.
- Ví dụ:
 - Chỉ có một phần ảnh hoặc văn bản được gán nhãn (ví dụ bức ảnh về người, động vật hoặc các văn bản khoa học, chính trị).
 - Phần lớn các bức ảnh/văn bản khác chưa được gán nhãn được thu thập từ internet (chi phí thấp).
 - Học từ dữ liệu hỗn hợp này \Rightarrow **Semi-Supervised Learning**

Reinforcement Learning (Học tăng cường)

37

Reinforcement Learning

- Là các thuật toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance).
 - Reinforcement learning (hiện nay) chủ yếu được áp dụng vào Lý Thuyết Trò Chơi.
 - Các thuật toán xác định bước đi tiếp theo để đạt được kết quả cao nhất.
- Ví dụ:
 - Máy tính AlphaGo chơi cờ vây thắng cả con người (số nước đi rất lớn $\sim 10^{761}$).
 - Máy sử dụng cả Supervised learning và Reinforcement learning.

Học dựa trên mô hình (model - based) và Học dựa trên mẫu (instance - based)

Phân loại thống kê

Mô hình sinh và mô hình phân biệt

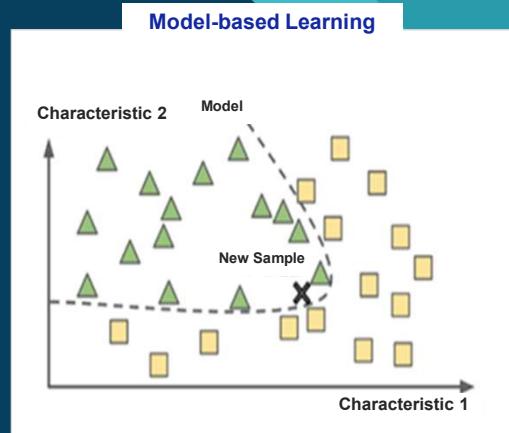
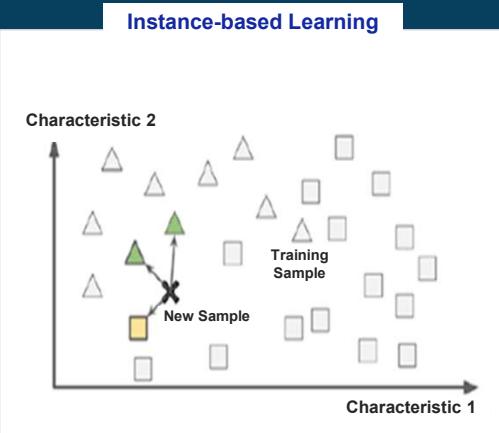
Một vài ví dụ

39

Instance based vs. Model based Learning

- **Instance based:** học bằng cách ghi nhớ các mẫu đào tạo. Phương pháp dựa trên sự tổng quát hóa bằng cách so sánh dữ liệu mới và các mẫu đã học bằng cách sử dụng các phép đo tương tự.
- **Model based:** Xây dựng một mô hình, học từ tập hợp mẫu và sử dụng nó để dự đoán.

Instance based vs. Model based Learning



Generative & Discriminative classifiers (Mô hình sinh và Mô hình phân biệt)

Phân loại thống kê

Mô hình sinh và mô hình phân biệt

Một vài ví dụ

Học máy - Phân loại thống kê

- Phân loại bằng thống kê: *Thủ tục để sắp xếp các cá thể vào từng nhóm dựa trên số lượng thông tin về một hay nhiều tính chất kế thừa của cá thể đó (ví dụ các điểm, các biến, các đặc điểm, v.v...) và dựa vào một tập huấn luyện của các cá thể đã được đánh nhãn sẵn.*
- Học máy và phân loại thống kê.
 - Supervised Learning: Phân loại dữ liệu x vào một trong các lớp nhãn y_i .
 - Unsupervised Learning: Phân nhóm dựa trên tính chất/tần suất...
 - Clustering \Leftrightarrow Phân cụm/phân nhóm.
 - Association – Phân theo khả năng xuất hiện (tần suất).

Học máy - Phân loại thống kê

- Mô hình toán:
 - Cho sẵn một tập huấn luyện
$$\{ (x_1, y^1), \dots, (x_i, y^i), \dots, (x_n, y^n) \}; \text{ với } x_i \in X, y^i \in Y$$
 - Xây dựng ánh xạ phân loại: $h: X \rightarrow Y$ để ánh xạ
$$\forall x \in X: y = h(x) \in Y - \text{nhãn tương ứng}$$
- Các hệ thống nhận dạng mẫu (*Pattern Recognition*) thường sử dụng phân loại thống kê

Mô hình sinh và mô hình phân biệt

- Hai cách tiếp cận chính của Phân loại thống kê:
 - Mô hình phân biệt (Discriminative Classifiers):
Dựa vào training data (x_i, y_i) để giải quyết phân lớp theo một trong hai cách:
 - Học mô hình (hàm) $f: x \rightarrow y = f(x)$
 - Với dữ liệu x , tính khả năng x có nhãn y (xác suất có điều kiện $P(y|x)$)
 - Mô hình tạo sinh (Generative Classifiers):
 - Dựa vào/mô tả phân phối dữ liệu trong bản thân tập dữ liệu để xác định một dữ liệu mới thuộc một nhãn sẽ như thế nào.
 - Bằng cách lấy mẫu từ mô hình này, ta có thể tạo ra dữ liệu mới.
 - Mô hình này cho biết khả năng (xác suất) xảy ra của một mẫu dữ liệu mới.

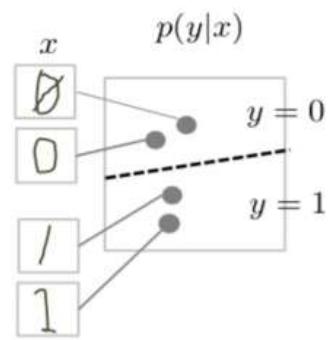
Mô hình sinh và mô hình phân biệt

Mô hình toán:

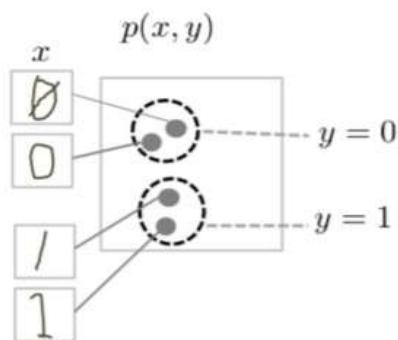
- Tập dữ liệu (observable variable) X
& biến đích (target variable) Y
- Mô hình phân biệt (Discriminative Classifiers):
Dựa vào training data (x_i, y_i) mô hình một trong hai cách:
 - Học mô hình (hàm) $f: x \rightarrow y = f(x)$
 - Mô hình xác suất có điều kiện của Y với điều kiện X : $P(Y|X = x)$
- Mô hình tạo sinh (Generative Classifiers): mô hình thống kê của phân phối xác suất đồng thời (joint probability distribution) trên $X \times Y$: $P(X, Y)$

Mô hình sinh và mô hình phân biệt

- Discriminative Model



- Generative Model



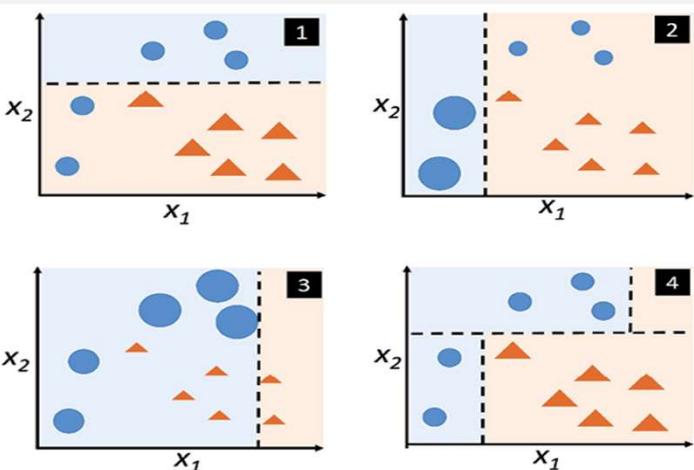
Statistical Classification

Giới thiệu mô hình phân loại thống kê
Một số mô hình phân biệt

Bài toán phân loại

- Khái niệm: phân loại thống kê là cách tiếp cận học tập có giám sát nhằm huấn luyện một chương trình để phân loại thông tin mới, chưa được phân lớp (gắn nhãn) dựa trên mức độ liên quan của nó với dữ liệu được phân lớp (gắn nhãn), đã biết.
- Các khái niệm liên quan:
 - Dữ liệu (biến quan sát - **observable variable**) X & biến đích (tập nhãn - **target variable**) Y
 - Dữ liệu huấn luyện $\{x_i, y_i\} \subset X \times Y$, dữ liệu đã được gán nhãn
- Các thuật toán sắp xếp dữ liệu chưa được gắn nhãn thành các lớp có nhãn hoặc các loại thông tin được gọi là **bộ phân loại (classifiers)**.

Bài toán phân loại



Phân loại tuyến tính (Linear classifiers)

- Bộ phân loại tuyến tính phân loại dữ liệu thành các nhóm (gán nhãn) dựa trên mối liên hệ tuyến tính của các đặc trưng đầu vào.
- Các phân loại tuyến tính phân tách dữ liệu bằng cách sử dụng các dạng siêu phẳng (đường thẳng/mặt phẳng):

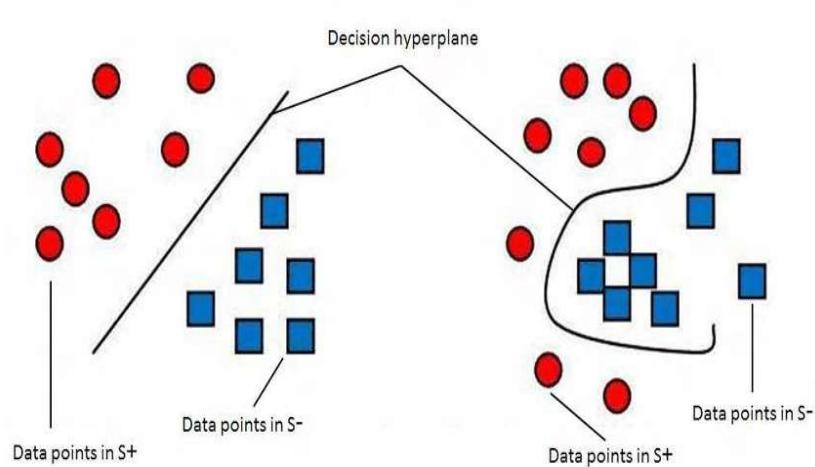
$$y = f(x) = Ax + b \quad (y, b \in Y \subseteq R^n; x \in X \subset R^m)$$

- Chỉ có thể được sử dụng để phân loại dữ liệu có thể phân tách tuyến tính.
- Có thể được sửa đổi để phân loại dữ liệu có thể phân tách không tuyến tính.
- Ví dụ: Perceptron; Logistic Regression; SVM...

Phân loại phi tuyến (Nonlinear classifiers)

- Bộ phân loại tuyến tính phân loại dữ liệu thành các nhóm (gán nhãn) dựa trên mối liên hệ phi tuyến tính.
- Nếu một trong bài toán phân loại, các ranh giới lớp dữ liệu của nó không thể được xác định với siêu phẳng tuyến tính, thì các bộ phân loại phi tuyến thường chính xác hơn các bộ phân loại tuyến tính.
- Nếu bài toán là tuyến tính, tốt nhất là sử dụng một bộ phân loại tuyến tính đơn giản hơn.
- Ví dụ: ANN, Random Forest, K-NN...

Linear vs. Nonlinear classifiers



MACHINE LEARNING - 2021

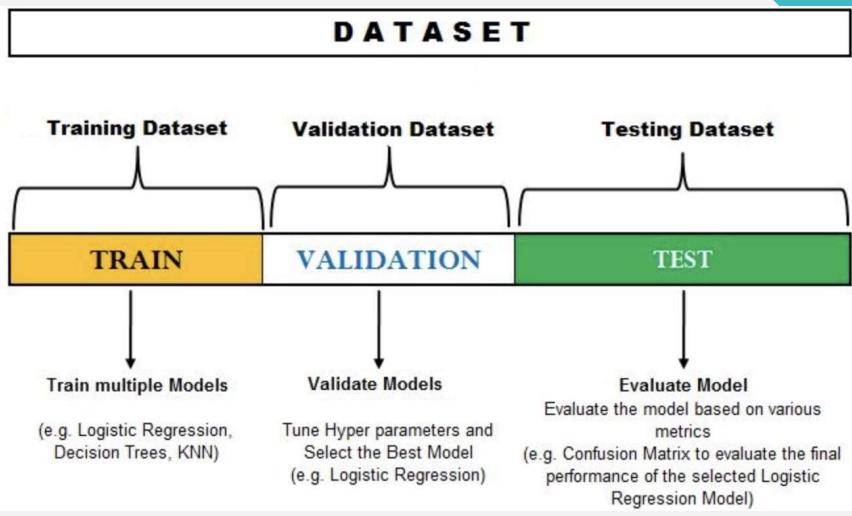
53

Đánh giá mô hình

Khái niệm liên quan
Các tiêu chí đánh giá
Một số độ đo

54

Một số khái niệm - Dataset



MACHINE LEARNING - 2021

55

Một số khái niệm - MLE

- Maximum Likelihood Estimation – Ước lượng hợp lý cực đại:
 - Dùng để ước lượng giá trị tham số của một mô hình xác suất /thống kê dựa trên những dữ liệu quan sát được.
 - Dựa vào việc cực đại hóa Likelihood function \Rightarrow Bộ tham số - phương pháp được coi là “Hợp lý cực đại”.
 - Theo Suy diễn Bayes: MLE - trường hợp đặc biệt của Maximum A Posteriori estimation (MAP), đưa ra giả thiết về phân phối đều của các tham số.
 - Theo suy diễn tần suất: MLE không khẳng định về xác suất của các tham số mà chỉ khẳng định về xác suất của các ước lượng

MACHINE LEARNING - 2021

56

Maximum Likelihood Estimation

- Mô hình toán MLE – theo quan điểm thống kê:
 - Dataset $X = \{x_1, x_2, \dots, x_n\}$ là mẫu quan sát ngẫu nhiên từ quần thể;
 - Mục đích: tìm quy luật về quần thể mà có thể nhất (hợp lý cực đại) sinh ra mẫu X , đặc biệt là phân phối xác suất chung của các biến ngẫu nhiên không nhất thiết độc lập hay cùng phân phối.
 - Mô hình f phụ thuộc tham số $\theta = \{\theta_1, \theta_2, \dots, \theta_d\}^T : f \in \{f(\cdot, \theta) | \theta \in \Theta\}$ với Θ - không gian tham số.
 - Cho trước mô hình xác suất - một họ các phân phối $\{f(\cdot, \theta) | \theta \in \Theta\}$ ($\theta \in \Theta$ có thể ở dạng dữ liệu nhiều chiều).
 - MLE tìm $\theta \in \Theta$ để “Likelihood function” $\mathcal{L}(\theta; x) = p_\theta(x) = P_\theta(X = x)$

Maximum Likelihood Estimation

- $\mathcal{L}(\theta; x)$ đạt cực đại \Leftrightarrow đi tìm cách giải thích hợp lý cho các dữ liệu quan sát được:
 - Định nghĩa Maximum Likelihood Estimates như sau:
 $\hat{\theta} \in \{\operatorname{argmax}_{\theta \in \Theta} \mathcal{L}(\theta; x)\}$ - nếu $\max \mathcal{L}(\theta; x)$ tồn tại
 - Có thể dùng log-likelihood function để thay thế:
$$\ell(\theta; x) = \ln(\mathcal{L}(\theta; x))$$
 - hoặc dùng hàm log-likelihood trung bình
$$\bar{\ell}(\theta; x) = \frac{1}{N} \ln(\mathcal{L}(\theta; x))$$
 - N là số phần tử mẫu
 - Estimator $\hat{\theta}$ xấp xỉ log-likelihood kỳ vọng của một quan sát duy nhất trong mô hình.

Maximum Likelihood Estimation

- Ví dụ: $X = \{x_1, x_2, \dots, x_n\}$ - lấy từ phân bố chuẩn $\mathcal{N}(\mu, 1)$. Tìm μ
 - μ là MLE của phân phối:

$$\mathcal{L}(\mu; x) = P(X|\mu) = \prod_{k=1}^n \left(\frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (x_k - \mu)^2 \right\} \right)$$

- Lấy log-likelihood: $L(\mu) = \ln P(X|\mu) = -\frac{1}{2} \sum_{k=1}^n (x_k - \mu)^2$
- Lấy đạo hàm - giải pt $L'(\mu) = 0 \Rightarrow$ thu được MLE của μ :

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{k=1}^n x_k = \bar{x}$$

- Tương tự, nếu $X \sim \mathcal{N}(\mu, \sigma^2)$ thì MLE của σ^2 là

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$$

Các tiêu chí đánh giá

- Một bài toán Machine Learning có thể được giải quyết bởi nhiều phương pháp/những mô hình khác nhau.
- Cần phải đánh giá “hiệu năng” của mô hình trên dữ liệu mới (evaluate model performance on unseen data).
- Đánh giá mô hình giúp ta trả lời những câu hỏi sau:
 - Mô hình đã được huấn luyện thành công hay chưa?
 - Mức độ thành công của mô hình tốt đến đâu?
 - Khi nào nên dừng quá trình huấn luyện?
 - Khi nào nên cập nhật mô hình?

Các tiêu chí đánh giá

- 1. Tính chính xác (Accuracy)
 - Mức độ dự đoán (phân lớp) chính xác của hệ thống (đã được huấn luyện) đối với các ví dụ kiểm chứng (test instances)
- 2. Tính hiệu quả (Efficiency)
 - Chi phí về thời gian và tài nguyên (bộ nhớ) cần thiết cho việc huấn luyện và kiểm thử hệ thống
- 3. Khả năng xử lý nhiễu (Robustness)
 - Khả năng xử lý (chịu được) của hệ thống đối với các ví dụ nhiễu (lỗi) hoặc thiếu giá trị

Các tiêu chí đánh giá

- 4. Khả năng mở rộng (Scalability)
 - Hiệu năng của hệ thống (vd: tốc độ học/phân loại) thay đổi như thế nào đối với kích thước của tập dữ liệu
- 5. Khả năng diễn giải (Interpretability)
 - Mức độ dễ hiểu (đối với người sử dụng) của các kết quả và hoạt động của hệ thống

Độ đo đánh giá



- Tùy theo loại phương pháp sẽ cần công cụ/độ đo đánh giá khác nhau
- Trong phần tiếp theo chủ yếu xem xét các công cụ đánh giá độ chính xác cho phương pháp phân loại (classification)

Độ đo tính chính xác (Accuracy)

- Accuracy (độ chính xác) chỉ đơn giản đánh giá mô hình thường xuyên dự đoán đúng đến mức nào.
- Độ chính xác là tỉ lệ giữa số điểm dữ liệu được dự đoán đúng và tổng số điểm dữ liệu:

$$\text{Accuracy} = \frac{1}{|data_test|} \sum_{x \in data_test} \text{Identical}(o(x), c(x))$$
$$\text{Identical}(a, b) = \begin{cases} 1, & \text{nếu } (a = b) \\ 0 & \text{trong các TH còn lại} \end{cases}$$

- Trong đó:
 - x: là một ví dụ trong tập kiểm thử datatest
 - o(x): giá trị đầu ra (phân lớp) của hệ thống với dữ liệu x
 - c(y): Phân lớp đúng với ví dụ x

Độ đo tính chính xác (Accuracy)

- Nhược điểm của Accuracy là chỉ cho ta biết độ chính xác khi dự báo của mô hình- nhưng không thể hiện mô hình đang dự đoán sai như thế nào.
- Một mô hình có độ chính xác (Accuracy) cao chưa hẳn đã tốt.
- Accuracy lộ rõ hạn chế khi được sử dụng trên bộ dữ liệu không cân bằng (imbalanced dataset).
- Ví dụ:
 - Ngân hàng có 10000 giao dịch mỗi ngày, với xấp xỉ 10 giao dịch là bất thường
 - Mô hình dự đoán với Accuracy 99% cho giao dịch bình thường \Rightarrow Vô nghĩa!

Confusion Matrix

- Confusion matrix là một kỹ thuật đánh giá hiệu năng của mô hình cho các bài toán phân lớp.
- Confusion matrix là một ma trận thể hiện số lượng điểm dữ liệu thuộc vào một class và được dự đoán thuộc vào class.

		True Class		
		Apple	Orange	Mango
Predicted Class	Apple	7	8	9
	Orange	1	2	3
	Mango	3	2	1

Confusion Matrix

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
	Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$	

Confusion Matrix

- Đối với bài toán phân loại, chúng ta thường quan tâm đối tượng cần phân loại có thuộc lớp X, hay có thuộc tính X hay không.
- Ví dụ:
 - Với bài toán phân loại thư rác, ta quan tâm một email có phải là thư rác hay không, tức X = spam.
 - Lớp X là Positive class, lớp còn lại là Negative class.
- Bài toán phân loại có 02 loại ở đầu ra như trên gọi là phân loại nhị phân (Binary Classification).
- Đầu ra có thể là True/False; 1/0; 1/-1 ... ứng với positive (+) và negative (-)

Confusion Matrix

- Xét bài toán phân loại nhị phân:

- True Positive (TP): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng)
- True Negative (TN): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng)
- False Positive (FP): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai) – Type I Error
- False Negative (FN): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai) – Type II Error

Confusion Matrix

- Cần dự đoán kết quả xét nghiệm COVID-19 của 1000 người nghi nhiễm.
- Kết quả dự đoán của mô hình:
 - Mô hình dự đoán có 30 ca dương tính, trong khi thực tế có 13 người nhiễm COVID-19
 - Mô hình dự đoán có 970 ca âm tính, nhưng thực tế trong 970 ca đó có 20 ca dương tính
- Chúng ta có thể biểu diễn kết quả dự đoán của mô hình bằng confusion matrix như sau:

Confusion Matrix

		Predicted class	
		Positive	Negative
Actual class	Positive	13 (True Positive)	20 (False Negative)
	Negative	17 (False Positive)	950 (True Negative)

Confusion Matrix

- **Kết quả:**

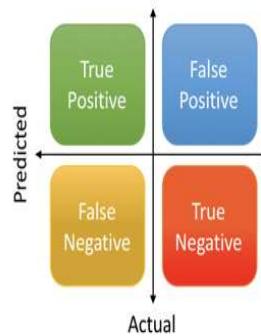
- True Positive TP = 13: có 13 người nhiễm COVID-19 được mô hình dự đoán đúng
- False Positive FP = 17: có 17 người âm tính với COVID-19, nhưng được mô hình dự đoán dương tính
- True Negative TN = 950: 950 trường hợp âm tính được mô hình phân loại chính xác
- False Negative FN = 20: có 20 trường hợp dương tính với COVID-19 nhưng bị mô hình phân loại sai

Precision và Recall

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$



Precision và Recall

Hãy trở lại với ví dụ về xét nghiệm COVID-19, precision và recall của mô hình trong ví dụ này là:

$$\text{precision} = \frac{TP}{TP + FP} = \frac{13}{13 + 17} = 0.43$$

$$\text{recall} = \frac{TP}{TP + FN} = \frac{13}{13 + 20} = 0.39$$

Có thể thấy rằng precision và recall của mô hình này còn thấp, tức độ chính xác của mô hình chưa cao.

F1-Score

• F-Score

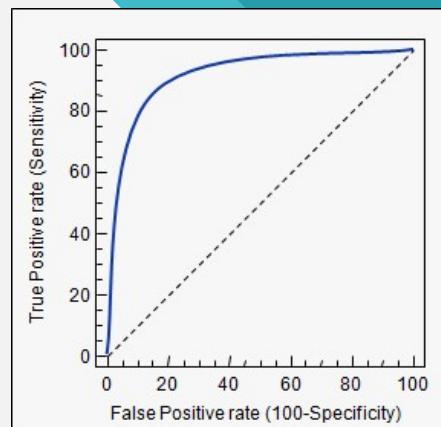
$$F_{\beta} = (1 + \beta^2) \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

Tham số β quyết định mức độ coi trọng giữa Precision và Recall

- $\beta > 1$: Recall được coi trọng hơn Precision
- $\beta < 1$: Precision được coi trọng hơn Recall
- $\beta = 1$: Precision và Recall được coi trọng ngang nhau

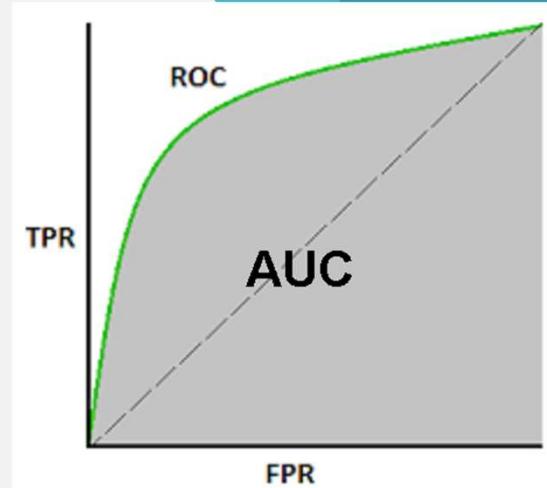
Một số đánh giá khác (đọc thêm)

- ROC (Receiver Operating Characteristic) curve dựa trên các tỷ lệ:
 - True Positive rate: $TPR = \frac{True\ Positive}{Total\ Positive}$
 - False Positive rate: $FPR = \frac{False\ Positive}{Total\ Negative}$
 - Vẽ đường biểu diễn liên hệ giữa TPR (trục tung) theo (FPR) trực hoành \Rightarrow thu được đường cong ROC.
 - Hình bên minh họa ROC curve tính theo %.
- Đường cong đi càng gần góc trên trái, độ chính xác tổng thể của mô hình càng cao (tự giải thích?).



Một số đánh giá khác (đọc thêm)

- AUC – Area under curve: Là phần diện tích dưới đường cong ROC.
- Giá trị AUC càng gần 1.0 thì khả năng phân loại của mô hình càng chính xác.
- Confidence interval (Khoảng tin cậy) – Tương tự khái niệm khoảng tin cậy trong ước lượng đại lượng thống kê.



MACHINE LEARNING - 2021

77

Một số đánh giá khác (đọc thêm)

- Gain chart & Lift chart - Dựa vào các tỉ lệ:

- Sensitive rate:

$$Se = Recall = \frac{\text{True Positive}}{\text{Total Positive} + \text{False Negative}}$$

- Support rate:

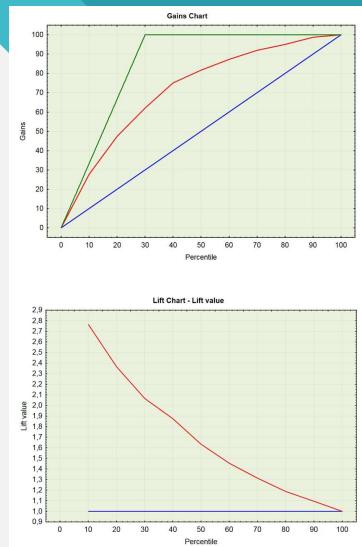
$$Su = \frac{\text{True Positive} + \text{False Positive}}{\text{Total Observations}}$$

- Gain chart biểu diễn liên hệ

$$\{(x, y)\} = \{(Su, Se) | Se \in [0,1]\}$$

- Lift chart biểu diễn liên hệ

$$\{(x, y)\} = \left\{ \left(Su, \frac{Se}{Su} \right) | Se \in [0,1] \right\}$$

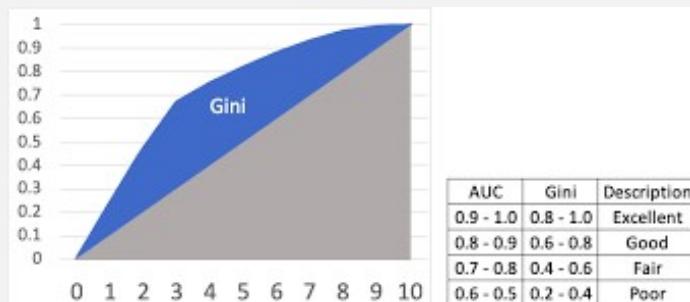


MACHINE LEARNING - 2021

78

Một số đánh giá khác (đọc thêm)

- Tự giải thích ý nghĩa biểu đồ Gain – Biểu đồ Lift.
- Hệ số Gini: $Gini_{Coef} = 2 \times AUC - 1$
- Hệ số Gini > 0.6 thì mô hình được coi là tốt (tự giải thích ý nghĩa của hệ số Gini).



The end of the 1st class