



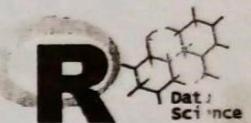
R for Data Science

Bài 16: Tổng quan Machine Learning

Ngành LT & Mạng

https://csc.edu.vn/lap-trinh-va-cSDL/R-Programming-Language-for-Data-Science_190

2020



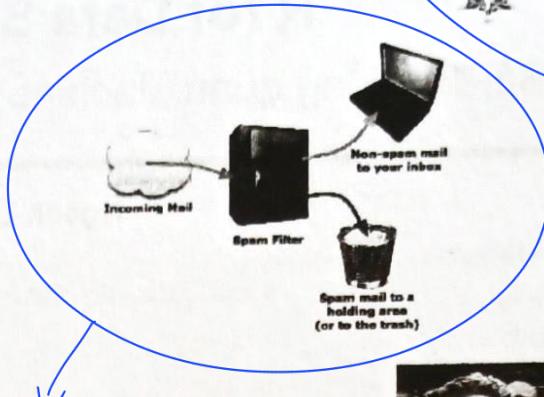
Nội dung

1. Giới thiệu
2. Phân loại
3. Thuật ngữ
4. Làm việc với dự án Machine Learning



Giới thiệu

Machine Learning ???



lý ứng của ML vào detect
spam trong gmail



Is Robot

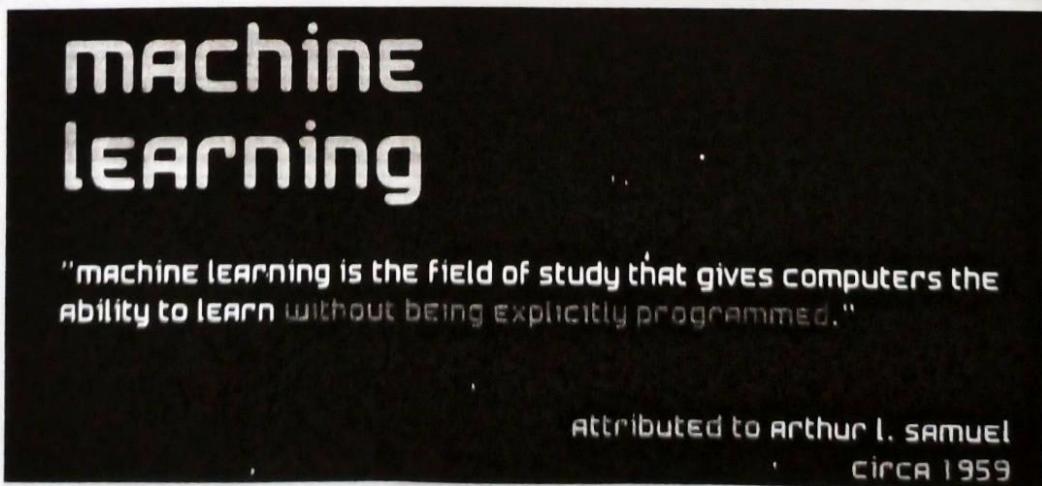
Nhà dân tộc
mèo



Giới thiệu

□ Machine learning

- Là khoa học (và cả nghệ thuật) của việc lập trình máy tính để máy tính có thể “học hỏi” từ dữ liệu



☐ Machine learning

- Là một nhánh của Trí tuệ nhân tạo (AI - artificial intelligence), là một lĩnh vực trong khoa học máy tính sử dụng các kỹ thuật thống kê để cung cấp cho máy tính khả năng “học”, với dữ liệu mà không cần được lập trình một cách rõ ràng.

Theo https://en.wikipedia.org/wiki/Machine_learning •

R programming language for Data Science

5

Giới thiệu



Artificial Intelligence

Any technique which enables computers to mimic human behavior.

Machine Learning

Subset of AI techniques which use statistical methods to enable machines to improve with experiences. → Sử dụng các kỹ thuật thống kê

Deep Learning

Subset of ML which make the computation of multi-layer neural networks feasible. → Tính toán trên đa lớp

Giới thiệu

□ Machine learning

- Có liên quan chặt chẽ tới **thống kê tính toán** và cũng tập trung vào việc dự đoán, ra quyết định thông qua việc sử dụng máy tính.
- Có quan hệ chặt chẽ với **tối ưu toán học**, trong đó cung cấp các phương thức, lý thuyết và các miền ứng dụng cho lĩnh vực tương ứng.
- Đôi khi được cho vào cùng nhóm với **data mining**, khi nó tập trung nhiều hơn vào phân tích dữ liệu khám phá, còn gọi là **unsupervised learning**.



Giới thiệu

□ Machine learning là:

- **Học từ dữ liệu**
- **Học từ chính nó**
- **Khám phá các mẫu ẩn (hidden pattern)**
- **Đưa ra các quyết định theo hướng dữ liệu**



Giới thiệu

☐ Tại sao sử dụng Machine Learning?

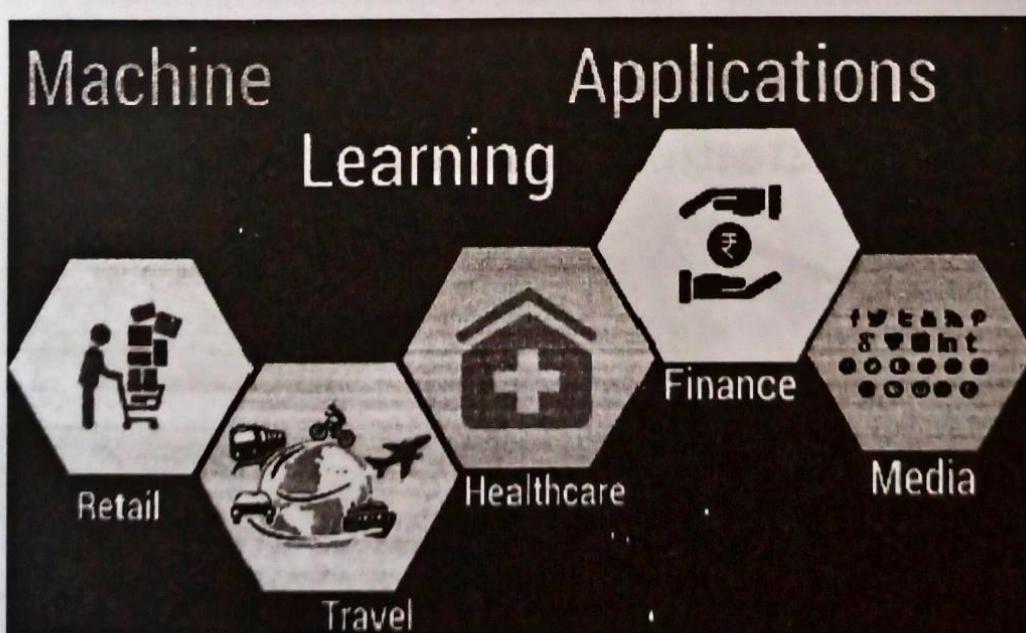
- Các vấn đề mà các giải pháp tồn tại yêu cầu rất nhiều thực hiện thủ công hoặc danh sách các quy tắc => một thuật toán Machine Learning có thể đơn giản hóa code và hoạt động tốt hơn.
- Các vấn đề phức tạp mà chưa có giải pháp tốt nào khi sử dụng phương pháp tiếp cận truyền thống: các kỹ thuật Machine Learning tốt có thể tìm ra giải pháp.
- Khi môi trường biến động: một hệ thống Machine Learning có thể thích nghi với dữ liệu mới.
- Có thể nhận thông tin chi tiết về các vấn đề phức tạp và **lượng dữ liệu lớn**

R programming language for Data Science

9

Giới thiệu

☐ Một số ứng dụng của Machine Learning



Giới thiệu

❑ Một số ứng dụng của Machine Learning

Nhận dạng giọng nói



Speech Recognition

(nhân deep learning)

Nhận dạng khuôn mặt



Facial Recognition



Optical Character Recognition

MACHINE
LEARNING



Autonomous Driving

Handwritten digits to read

Demande de carte à 12075366 ->12075366

Demande, Dorisien

Le vous ai envoyé le 9 Mars 2011

Nhận dạng chữ viết tay

Lái xe tự động

R programming language for Data Science



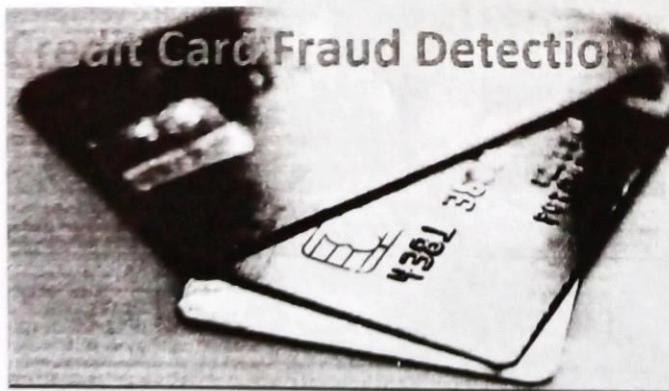
Image Recognition

Nhận dạng hình ảnh

11

Giới thiệu

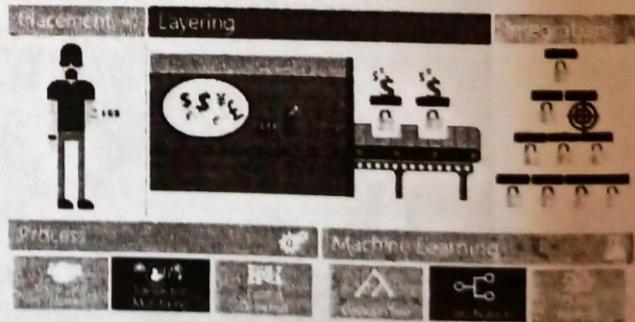
❑ Một số ứng dụng của Machine Learning



Phát hiện gian lận thẻ tín dụng

Fraud Detection

Money Laundering Prevention



Phòng chống rửa tiền

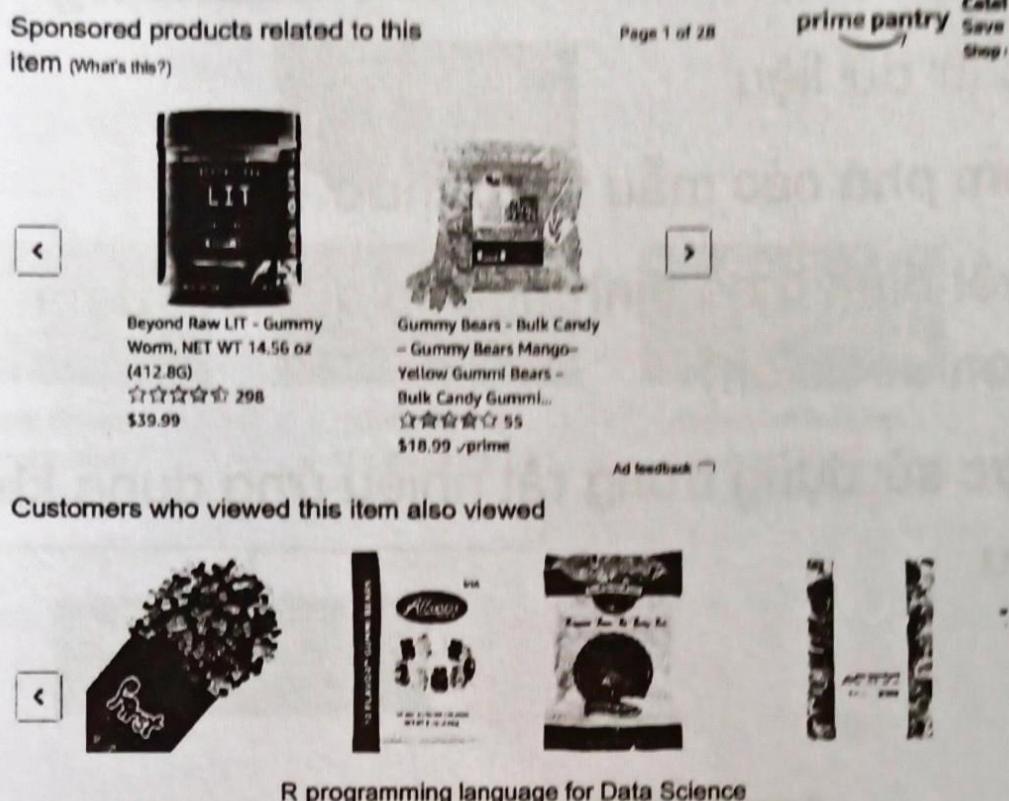
Giới thiệu

- Đề xuất trên các trang web

Sponsored products related to this item (What's this?)

Page 1 of 28

prime pantry Save
Shop



Beyond Raw LIT - Gummy Worm, NET WT 14.56 oz (412.8G)
★★★★★ 298 \$39.99

Gummy Bears - Bulk Candy - Gummy Bears Mango- Yellow Gumm...
Bulk Candy Gummi...
★★★★★ 55 \$18.99 /prime

Ad feedback

Customers who viewed this item also viewed



R programming language for Data Science

13

Giới thiệu *(Áp dụng của machine learning)*

- Dịch vụ truyền thông xã hội (Social Media Services)
 - People You May Know
 - Face Recognition
 - Similar Pins
- Hỗ trợ khách hàng online (Online Customer Support)
- Công cụ tìm kiếm, lọc dữ liệu (Search Engine Result Refining)
- Lọc thư rác và phần mềm độc hại (Email Spam & Malware Filtering)
- Giám sát bằng video (Videos Surveillance)
- Trợ lý ảo (Virtual Personal Assistants)

Giới thiệu

❑ Các mô hình của Machine Learning

- Học từ dữ liệu
- Khám phá các mẫu và xu hướng
- Quyết định theo định hướng dữ liệu (data-driven decision)
- Được sử dụng trong rất nhiều ứng dụng khác nhau

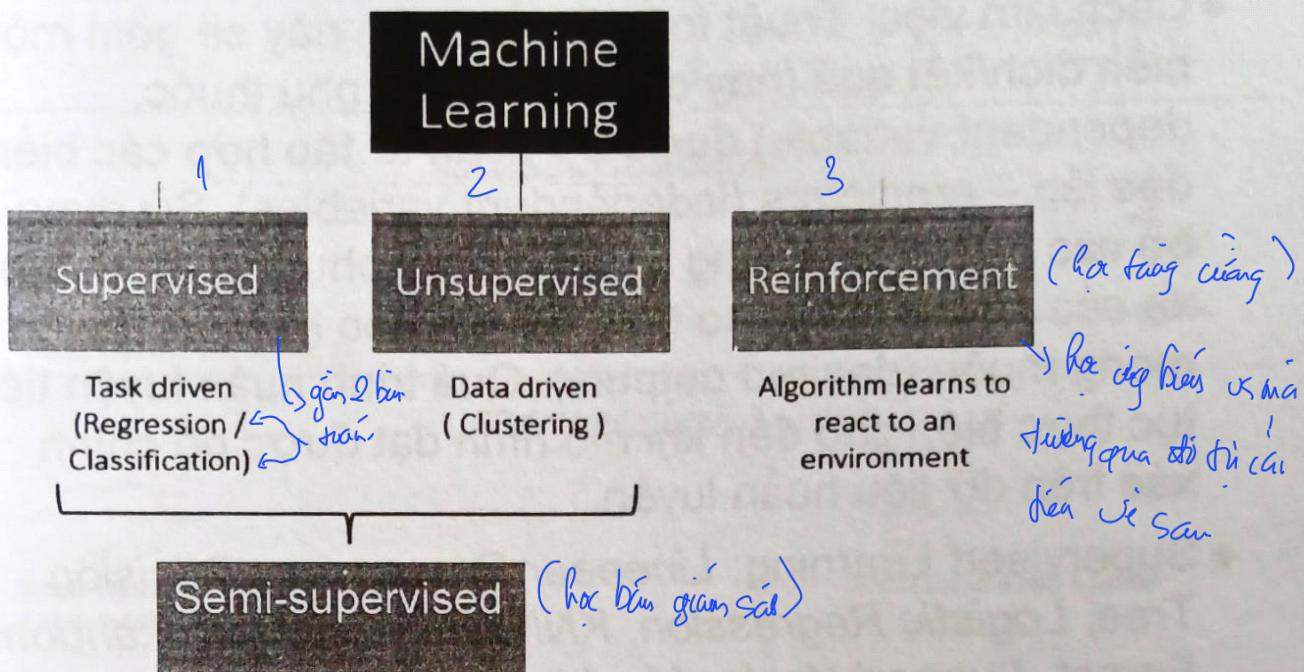


Nội dung

1. Giới thiệu
2. Phân loại
3. Thuật ngữ
4. Làm việc với dự án Machine Learning



Types of Machine Learning



R programming language for Data Science

17

Phân loại

Supervised (target is available)

Classification (label)

Regression^(Số)

Unsupervised (target is not available)

Cluster Analysis

Association Analysis

Phân loại

□ Supervised Learning

- Cách làm việc: Thuật toán trong nhóm này sẽ gồm một biến đích/kết quả (hay còn gọi là biến phụ thuộc, dependent variable) được dự đoán từ tập hợp các biến độc lập – predictors (independent variables). Sử dụng bộ các biến này, chúng ta tạo ra một phương thức ánh xạ các dữ liệu đầu vào (inputs) cho các kết quả đầu ra mong muốn (desired outputs). Quá trình huấn luyện tiếp tục thực hiện cho đến khi mô hình đạt được độ chính xác trên dữ liệu huấn luyện.
- Supervised Learning: Linear Regression, Decision Tree, Logistic Regression, KNN, Naïve Bayes, Random Forest, Support Vector Machine, ...



Phân loại

□ Supervised Approaches (Các bước cần làm trong Supervised)

- Target (thứ mà mô hình dự đoán) được cung cấp
- Dữ liệu được gán nhãn ('Labeled' data)
- Gồm có: Classification & regression analysis



Phân loại

□ Supervised Learning (

Input

Model

Predicted



✓ New email

Classification



SPAM
vs.
Not SPAM

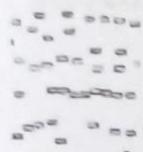
Data: E-mail

Label: Spam / not Spam

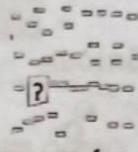
Categories: Spam / Not Spam

Regression

(Bài toán dự đoán
giá nhà)



Data: House features
Label: Price



850k

House Price: a real number



R programming language for Data Science

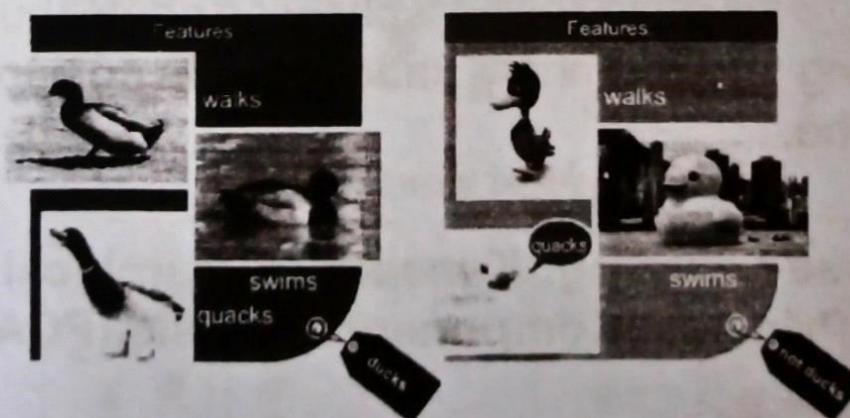
21

Phân loại

□ Classification (để đoán phải là con vịt hay hổ)

- Mục tiêu: dự đoán loại (category) từ các biến dữ liệu được cung cấp

If it Walks/Swims/Quacks Like a Duck Then It Must Be a Duck



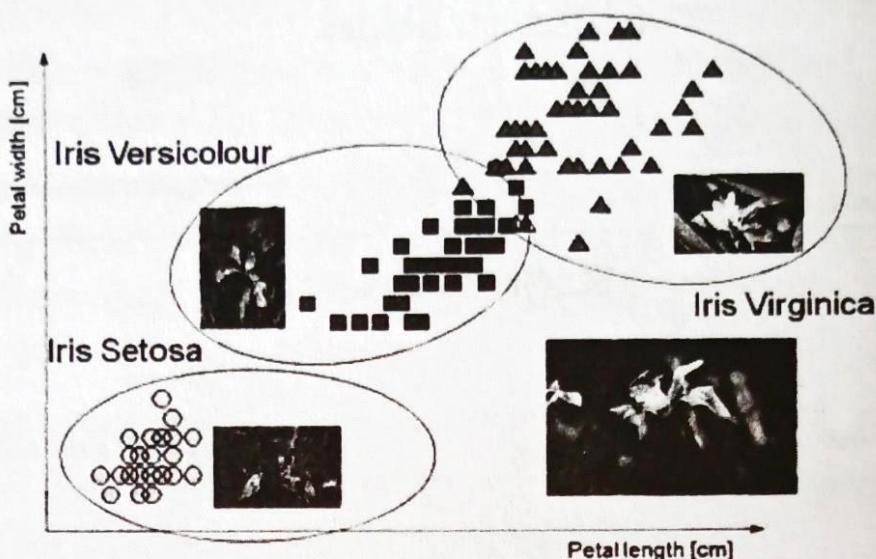
R programming language for Data Science

22

Phân loại

❑ Regression

- Mục tiêu: dự đoán giá trị số



Phân loại

❑ Unsupervised Learning

- Cách làm việc: trong thuật toán này chúng ta **không có bất kỳ biến target/outcome nào để dự đoán/ ước tính**
- Nó được sử dụng để **phân cụm vào các nhóm khác nhau**, được sử dụng rộng rãi cho việc phân khúc khách hàng vào các nhóm khác nhau để có hành động cụ thể.

❑ **Unsupervised Learning: K-means, Hierarchical clustering, Principal Component Analysis (PCA), Apriori,...**



Phân loại

□ Unsupervised Approaches

- Target không xác định
- Dữ liệu không được gán nhãn ('unlabeled' data)
- Bao gồm Cluster analysis & association analysis.

Phân loại

□ Cluster Analysis

- Mục tiêu: sắp xếp các item tương tự vào các nhóm

What is a natural grouping among these objects?



Clustering is subjective



Simpson's Family



School Employees



Females



Males

❑ Association Analysis

- Mục tiêu: tìm các quy luật để nắm bắt mối liên kết giữa các item

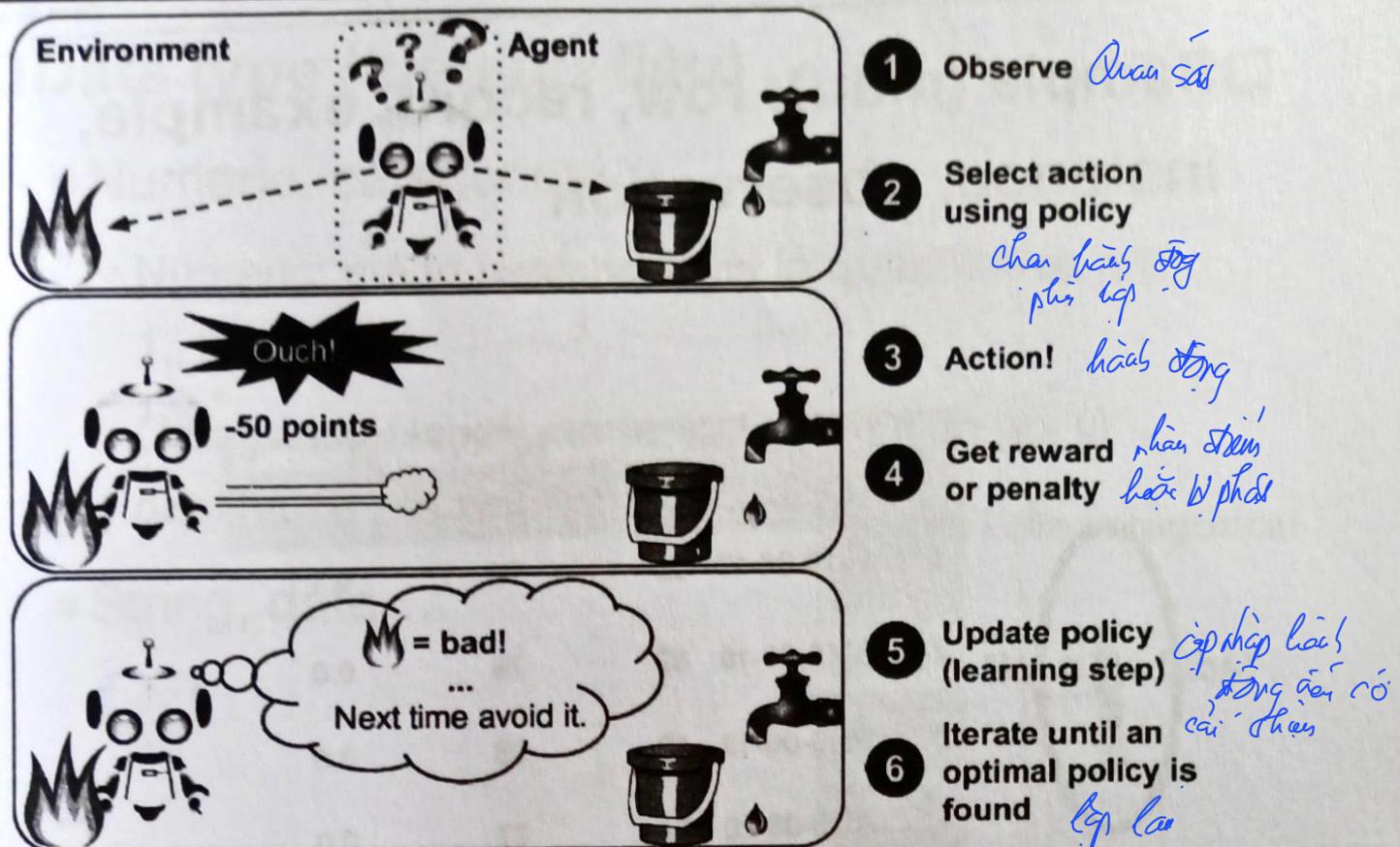
Bản thân máy tính có mua thêm sữa không



❑ Reinforcement Learning

- Cách làm việc: sử dụng thuật toán, máy được huấn luyện để đưa ra các quyết định cụ thể.
- Nó hoạt động theo cách: máy được tiếp xúc với môi trường nơi nó liên tục tự thử và sai. Máy sẽ học từ kinh nghiệm trong quá khứ và cố gắng nắm bắt kiến thức tốt nhất có thể để đưa ra những quyết định nghiệp vụ chính xác.
- Reinforcement Learning: Markov Decision Process...





R programming language for Data Science

29

Xem bài cũng là bài toán học tăng cường

Nội dung

1. Giới thiệu
2. Phân loại
3. Thuật ngữ
4. Làm việc với dự án Machine Learning

- ☐ **Sample (mẫu): row, record, example, instance, observation**

Variables					
	Date	MinTemp	MaxTemp	Rainfall	
1	2010-06-17	55	75	0.1	
2	2010-06-18	52	78	0.0	
3	2010-06-19	50	78	0.0	
4	2010-06-20	54	77	0.0	

R programming language for Data Science

31

- ☐ **Variable (biến): attribute, field, feature, column, dimension**

Variables					
	Date	MinTemp	MaxTemp	Rainfall	
1	2010-06-17	55	75	0.1	
2	2010-06-18	52	78	0.0	
3	2010-06-19	50	78	0.0	
4	2010-06-20	54	77	0.0	

R programming language for Data Science

32

Thuật ngữ

□ Data type (kiểu dữ liệu)

(1 feature chỉ có duy nhất 1 kiểu dữ liệu)

• Numeric, categorical (có 2 kiểu dữ liệu chính)

▪ Numeric: giá trị là số, còn gọi là quantitative. Vd: 1, 10⁵, -0.452

▪ Categorical: label, name, category, còn gọi là qualitative hay nominal

• String, date...

Color	Biến categorical
Red	
Silver	Giá trị là label
Blue	
White	
Black	

R programming language for Data Science

33

Thuật ngữ

Sample

Variable

- Feature
- Field
- Column
- ...

Variables					
ID	Date	MinTemp	MaxTemp	Rainfall	WindGustDir
1	2010-06-17	55	75	0.1	
2	2010-06-18	52	78	0.0	
3	2010-06-19	50	78	0.0	
4	2010-06-20	54	77	0.0	

- Instance
- Record
- Row
- Observation
- ...

Numeric
Quantitative

Categorical
Qualitative
Nominal

Nội dung

1. Giới thiệu
2. Phân loại
3. Thuật ngữ
4. Làm việc với dự án Machine Learning



Làm việc với dự án Machine Learning

❑ Các bước chính cần thực hiện

- Tìm hiểu bức tranh toàn cảnh của bài toán => bài toán cần giải là gì?
- Lấy dữ liệu
- Khám phá dữ liệu xem có gì bất thường hay không (thiếu dữ liệu, trùng lặp, outlier) và trực quan hóa dữ liệu để có được thông tin chi tiết
- Chuẩn bị dữ liệu/chuẩn hoá dữ liệu cho thuật toán Machine Learning algorithms
- Chọn lựa mô hình và huấn luyện
- Tinh chỉnh mô hình
- Trình bày giải pháp
- Thực thi, giám sát và duy trì hệ thống



Làm việc với dự án Machine Learning

☐ Làm việc với dữ liệu thực tế

- Khi tìm hiểu về Machine Learning, tốt nhất là nên thử nghiệm trên các bộ dữ liệu thực tế (real-world data)
 - Có khá nhiều nguồn có thể nhận dữ liệu (miễn phí/trả phí) như:
 - UC Irvine Machine Learning Repository
 - Kaggle datasets
 - Amazon's AWS datasets
 - Wikipedia's list of Machine Learning datasets
- ...

R programming language for Data Science

37

Làm việc với dự án Machine Learning

☐ Môi trường làm việc

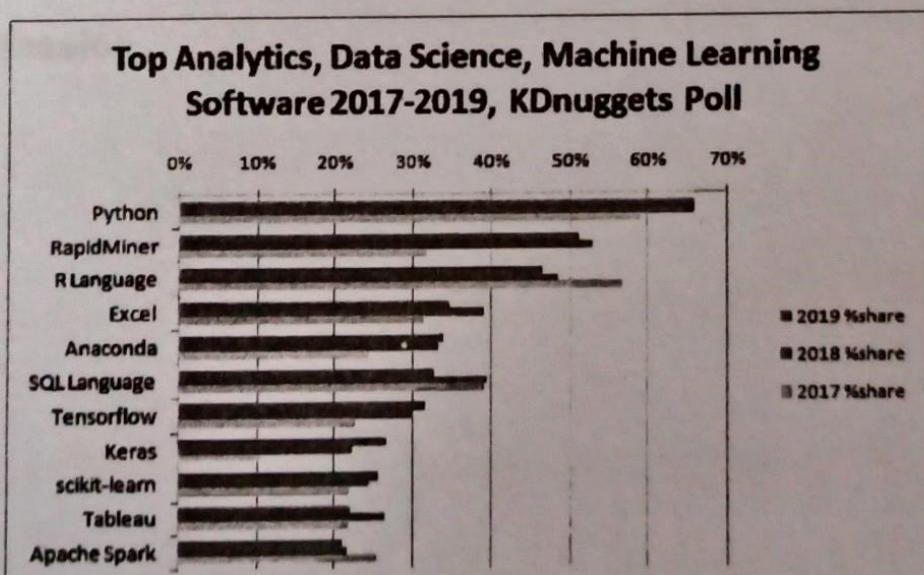
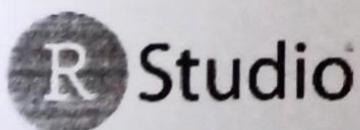


Fig 1: KDnuggets Analytics/Data Science 2019 Software Poll: top tools in 2019, and their share in the 2017, 2018 polls