Probability & Statistics – Xác Suất & Thống Kê

Nguyễn Quản Bá Hồng*

Ngày 22 tháng 10 năm 2024

Tóm tắt nội dung

This text is a part of the series *Some Topics in Advanced STEM & Beyond*:
URL: https://nqbh.github.io/advanced_STEM/.
Latest version:

- *Probability & Statistics – Xác Suất & Thống Kê*.
  PDF: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/probability_statistics/NQBH_probability_statistics.pdf.
  TEX: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/probability_statistics/NQBH_probability_statistics.tex.

Mục lục

1 Basic

Relationship among AL, ML, & DL. DL $\subset$ ML $\subset$ AI.

2 Probability – Xác Suất

Community – Cộng đồng. Andrey Nikolaevich Kolmogorov.
Resources – Tài nguyên.

1. Simon J. D. Prince. *Computer Vision: Models, Learning, & Inference*.

*A Scientist & Creative Artist Wannabe. E-mail: nguyenquanbahong@gmail.com. Bến Tre City, Việt Nam.

3 Statistics – Thống Kê

Community – Cộng đồng.
Resources – Tài nguyên.

1.

4 Stochastic – Ngẫu Nhiên

Community – Cộng đồng. Caroline Geiersbach, Michael Hintermüller.
Resources – Tài nguyên.

1.

5 Data Science (DS)

Community – Cộng đồng.
Resources – Tài nguyên.

1.

5.1 Wikipedia/Data Science

"*Data science* is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processing, scientific visualization, algorithms & systems to extract or extrapolate knowledge & insights from potentially noisy, structured, or unstructured data.

Data science also integrates domain knowledge from the underlying application domain (e.g., natural sciences, information technology, & medicine). Data science is multifaceted & can be described as a science, a research paradigm, a research method, a discipline, a workflow, & a profession.

Data science is "a concept to unify statistics, data analysis, informatics, & their related methods" to "understand & analyze actual phenomena" with data. It uses techniques & theories drawn from many fields within the context of mathematics, statistics, computer science, information science, & domain knowledge. However, data science is different from computer science & information science. Turing Award winner Jim Gray imagined data science as a "4th paradigm" of science (empirical, theoretical, computational, & now data-driven) & asserted that "everything about science is changing because of the impact of information technology" & the data deluge.

A *data scientist* is a professional who creates programming code & combines it with statistical knowledge to create insights from data.

5.2 Foundations

Data science is an interdisciplinary field focused on extracting knowledge from typically large data sets & applying the knowledge & insights from that data to solve problems in a wide range of application domains. The field encompasses preparing data for analysis, formulating data science problems, analyzing data, developing data-driven solutions, & presenting findings to inform high-level decisions in a broad range of application domains. As such, it incorporates skills from computer science, statistics, information science, mathematics, data visualization, information visualization, data sonification, data integration, graphic design, complex systems, communication & business. Statistician Nathan Yau, drawing on Ben Fry, also links data science to human-computer interaction: users should be able to intuitively control & explore data. In 2015, the American Statistical Association identified database management, statistics & machine learning, & distributed & parallel systems as the 3 emerging foundational professional communities.

5.2.1 Relationship to statistics

Many statisticians, including Nate Silver, have argued that data science is not a new field, but rather another name for statistics. Others argue that data science is distinct from statistics because it focuses on problems & techniques unique to digital data. Vasant Dhar writes that statistics emphasizes quantitative data & description. In contrast, data science deals with quantitative & qualitative data (e.g., from images, text, sensors, transactions, customer information, etc.) & emphasizes prediction & action. Andrew Gelman of Columbia University has described statistics as a non-essential part of data science.

Stanford professor David Donoho writes that data science is not distinguished from statistics by the size of datasets or use of computing & that many graduate programs misleadingly advertise their analytics & statistics training as the essence of a data-science program. He describes data science as an applied field growing out of traditional statistics.

5.3 Etymology

5.4 Data science & data analysis

Data science & data analysis are both important disciplines in the field of data management & analysis, but they differ in several key ways. While both fields involve working with data, data science is more of an interdisciplinary field that involves the application of statistical, computational, & machine learning methods to extract insights from data & make predictions, while data analysis is more focused on the examination & interpretation of data to identify patterns & trends.

Data analysis typically involves working with smaller, structured datasets to answer specific questions or solve specific problems. This can involve tasks such as data cleaning, data visualization, & exploratory data analysis to gain insights into the data & develop hypotheses about relationships between variables. Data analysts typically use statistical methods to test these hypotheses & draw conclusions from the data. E.g., a data analyst might analyze sales data to identify trends in customer behavior & make recommendations for marketing strategies.

Data science, on the other hand, is a more complex & iterative process that involves working with larger, more complex datasets that often require advanced computational & statistical methods to analyze. Data scientists often work with unstructured data such as text or images & use machine learning algorithms to build predictive models & make data-driven decisions. In addition to statistical analysis, data science often involves tasks such as data preprocessing, feature engineering, & model selection. E.g., a data scientist might develop a recommendation system for an e-commerce platform by analyzing user behavior patterns & using machine learning algorithms to predict user preferences.

While data analysis focuses on extracting insights from existing data, data science goes beyond that by incorporating the development & implementation of predictive models to make informed decisions. Data scientists are often responsible for collecting & cleaning data, selecting appropriate analytical techniques, & deploying models in real-world scenarios. They work at the intersection fo mathematics, computer science, & domain expertise to solve complex problems & uncover hidden patterns in large datasets.

Despite these differences, data science & data analysis are closely related fields & often require similar skills sets. Both fields require a solid foundation in statistics, programming, & data visualization, as well as the ability to communicate findings effectively to both technical & non-technical audiences. Both fields benefit from critical thinking & domain knowledge, as understanding the context & nuances of the data is essential for accurate analysis & modeling.

In summary, data analysis & data science are distinct yet interconnected disciplines within the broader field of data management & analysis. Data analysis focuses on extracting insights & drawing conclusions from structured data, while data science involves a more comprehensive approach that combines statistical analysis, computational methods, & machine learning to extract insights, build predictive models, & drive data-driven decision-making. Both fields use data to understand patterns, make informed decisions, & solve complex problems across various domains.

5.5 Cloud computing for data science

Cloud computing can offer access to large amounts of computational power & storage. In big data, where volumes of information are continually generated & processed, these platforms can be used to handle complex & resource-intensive analytical tasks.

Some distributed computing frameworks are designed to handle big data workloads. These frameworks can enable data scientists to process & analyze large datasets in parallel, which can reducing processing times.

5.6 Ethical consideration in data science

Data science involve collecting, processing, & analyzing data which often including personal & sensitive information. Ethical concerns include potential privacy violations, bias perpetuation, & negative societal impacts.

Machine learning models can amplify existing biases present in training data, leading to discriminatory or unfair outcomes." – Wikipedia/data science

6 Deep Learning (DL)

Community – Cộng đồng.
Resources – Tài nguyên.

1. [LBH15]. Yann LeCun, Yoshua Bengio, Geoffrey Hinton. *Deep Learning.*

Những năm gần đây, sự phát triển của các hệ thống tính toán cùng lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn đã giúp machine learning tiến thêm 1 bước dài. 1 lĩnh vực mới được ra đời được gọi là *học sâu* (deep learning, DL). Deep learning đã giúp máy tính thực thi những việc vào 10 năm trước tưởng chừng là không thể: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói & chữ viết, giao tiếp với con người, chuyển đổi ngôn ngữ, hay thậm chí cả sáng tác văn thơ & âm nhạc." – [Tiệ25, p. 15]

Φ: Start with simple things – Luôn bắt đầu từ những điều đơn giản. Khi bắt tay vào giải quyết 1 bài toán ML hay bất cứ bài toán nào, nên bắt đầu từ các thuật toán đơn giản. Không phải chỉ có các thuật toán phức tạp mới có thể giải quyết được vấn đề. Các thuật toán phức tạp thường có yêu cầu cao về khả năng tính toán & đôi khi nhạy cảm với cách chọn tham số. Ngược lại, các thuật toán đơn giản giúp ta nhanh chóng có 1 bộ khung cho mỗi bài toán. Kết quả của các thuật toán đơn giản

cũng mang lại cái nhìn sơ bộ về sự phức tạp của mỗi bài toán. Việc cải thiện kết quả sẽ được thực hiện dần ở các bước sau. " – [Tiệ25, p. 17]

Approach. Để giải quyết mỗi bài toán ML, cần chọn 1 mô hình phù hợp. Mô hình này được mô tả bởi bộ các tham số ta cần đi tìm. Thông thường, lượng tham số có thể lên tới hàng triệu & được tìm bằng cách giải 1 bài toán tối ưu. Khi viết về các thuật toán ML, VKTiệp sẽ bắt đầu từ các ý tưởng trực quan. Các ý tưởng này được mô hình hóa dưới dạng 1 bài toán tối ưu. Các suy luận toán học & ví dụ mẫu trên Python sẽ giúp hiểu rõ hơn về nguồn gốc, ý nghĩa, & cách sử dụng mỗi thuật toán. Xen kẽ giữa các thuật toán ML, trình bày các kỹ thuật tối ưu cơ bản, với hy vọng giúp hiểu rõ hơn bản chất của vấn đề.

Audiences. Cuốn sách được thực hiện hướng tới nhiều nhóm độc giả khác nhau. Nếu không thực sự muốn đi sâu vào phần toán, vẫn có thể tham khảo mã nguồn & cách sử dụng các thư viện. Nhưng để sử dụng các thư viện 1 cách hiệu quả, cũng cần hiểu nguồn gốc của mô hình & ý nghĩa của các tham số. Còn nếu thực sự muốn tìm hiểu nguồn gốc, ý nghĩa của các thuật toán, có thể học được nhiều điều từ cách xây dựng & tối ưu các mô hình.

Python. Python là 1 ngôn ngữ lập trình miễn phí, có thể được cài đặt dễ dàng trên các nền tảng hệ điều hành khác nhau. Có rất nhiều thư viện hỗ trợ ML cũng như DL trên Python. Có 2 thư viện Python chính thường được sử dụng là `numpy`, `scikit-learn`.

- `numpy` www.numpy.org là 1 thư viện phổ biến giúp xử lý các phép toán liên quan đến các mảng nhiều chiều, hỗ trợ các hàm gần gũi với đại số tuyến tính. Cách xử lý các mảng nhiều chiều.
- `scikit-learn/sklearn` scikit-learn.org: 1 thư viện chứa đầy đủ các thuật toán ML cơ bản & rất dễ sử dụng. Tài liệu của scikit-learn cũng là 1 nguồn tham khảo chất lượng cho MLer. Scikit-learn được dùng để kiểm chứng các suy luận toán học & các mô hình được xây dựng thông qua `numpy`.

Inevitability of mathematics in ML. Có rất nhiều thư viện giúp tạo ra các sản phẩm ML/DL mà không yêu cầu nhiều kiến thức toán. Hướng tới việc giúp hiểu bản chất toán học đằng sau mỗi mô hình trước khi áp dụng các thư viện sẵn có. Việc sử dụng thư viện + yêu cầu kiến thức nhất định về việc lựa chọn mô hình & điều chỉnh các tham số.

7 Machine Learning (ML)

Resources – Tài nguyên.

1. Andrew Ng. *Machine Learning Course* on Coursera.
2. Machine Learning Mastery: Making Developers Awesome at Machine Learning: https://machinelearningmastery.com.
   - Machine Learning Mastery/8 Inspirational Applications of Deep Learning.
3. Machine Learning cơ bản: https://machinelearningcoban.com/.
4. [Tiệ25]. Vũ Hữu Tiệp. *Machine Learning Cơ Bản*.

   Mã nguồn cuốn ebook "Machine Learning Cơ Bản": https://github.com/tiepvupsu/ebookMLCB.

Definition 1. "Machine learning (ML) *is a field of study in AI concerned with the development & study of statistical algorithms that can learn from data & generalize to unseen data, & thus perform tasks without explicit instructions. Quick progress in the fields of deep learning, beginning in 2010s, allowed neural networks to surpass many previous approaches in performance." – Wikipedia/machine learning*

Định nghĩa 1. *"Học máy (machine learning, ML) là 1 tập con của trí tuệ nhân tạo. Machine learning là 1 lĩnh vực nhỏ trong Khoa học Máy tính, có khả năng tự học hỏi dựa trên dữ liệu được đưa vào mà không cần phải được lập trình cụ thể: "Machine Learning is the subfield of computer science, that "gives computers the ability to learn without being explicitly programmed" – Wikipedia." – [Tiệ25, p. 15]*

"ML finds application in many fields, including natural language processing, computer vision, speech recognition, email filtering, agriculture, & medicine. The application of ML to business problems is known as predictive analysis.

Statistics & mathematical optimization/mathematical programming methods comprise the foundations of machine learning. Data mining is related field of study, focusing on exploratory data analysis (EDA) via unsupervised learning.

From a theoretical viewpoint, probably approximately correct (PAC) learning provides a framework for describing machine learning." – Wikipedia/machine learning

Relationships of ML to AI. As a scientific endeavor, machine learning grew out of the quest for AI. In the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks"; these were mostly perceptrons & other models e.g. ADALINE that were later found to be reinventions of the generalized linear models of statistics. Probabilistic reasoning was also employed, especially in automated medical diagnosis. However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI & machine learning. Probabilistic systems were plagued by theoretical & practical problems of data acquisition & representation.

8 Artificial Intelligence (AI)

Resources – Tài nguyên.

1. [BV14]. Lê Hoài Bắc, Tô Hoài Việt. *Cơ Sở Trí Tuệ Nhân Tạo*.
2. [Aou14]. Joseph E. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*.
3. [Aou19]. Joseph E. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence – Chạy Dua Với Robot: Học Tập Thời Trí Tuệ Nhân Tạo*.

9 Miscellaneous

Tài liệu

[Aou14] Joseph E. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. MIT Publisher, 2014, p. 187.

[Aou19] Joseph E. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence – Chạy Dua Với Robot: Học Tập Thời Trí Tuệ Nhân Tạo*. Trịnh Huy Nam dịch. Nhà Xuất Bản Thế Giới, 2019, p. 241.

[BV14] Lê Hoài Bắc and Tô Hoài Việt. *Cơ Sở Trí Tuệ Nhân Tạo*. Nhà Xuất Bản Khoa Học & Kỹ Thuật, 2014, p. 229.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521 (2015), pp. 436–444. DOI: 10.1038/nature14539. URL: https://doi.org/10.1038/nature14539.

[Tiệ25] Vũ Khắc Tiệp. *Machine Learning Cơ Bản*. 2025, p. 422.