ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

LÝ THUYẾT XÁC SUẤT VÀ THỐNG KÊ DÀNH CHO MÁY HỌC

Theory of Probability and Statistics for Machine Learning

Liên kết Github

Lĩnh vực nghiên cứu

CƠ SỞ TOÁN TRONG KHOA HỌC MÁY TÍNH

Báo cáo viên

NGUYỄN XUÂN HIỀN

Người hướng dẫn khoa học

Tiến sỹ **NGUYỄN AN KHƯƠNG**

Tiến sỹ **TRẦN TUẤN ANH**

GIỚI THIỆU CHUNG

- **Tầm quan trọng:** Xác suất và Thống kê là trụ cột toán học thiết yếu cho máy học.
- **Vai trò trung tâm:** Mô hình hóa và xử lý sự không chắc chắn trong dữ liệu và các tác vụ dự đoán.
- **Mục tiêu tiểu luận:** Tổng hợp lý thuyết nền tảng, cung cấp kiến thức vững chắc cho người mới bắt đầu nghiên cứu và xây dựng mô hình máy học.
- **Ứng dụng trong máy học:** Học có giám sát, học không giám sát, học tăng cường.

NỘI DUNG CHÍNH

- Tổng quan về Xác suất và Thống kê trong nghiên cứu máy học.
- Các khái niệm xác suất cơ bản.
- Biến ngẫu nhiên (rời rạc, liên tục), hàm mật độ xác suất.
- Đa biến ngẫu nhiên: xác suất đồng thời, có điều kiện, Định lý Bayes.
- Các tham số thống kê quan trọng: kỳ vọng, phương sai, độ lệch chuẩn, ma trận hiệp phương sai.
- Các loại không chắc chắn trong máy học.
- Ví dụ minh họa và bài tập tình huống.

TỔNG QUAN VỀ XÁC SUẤT VÀ THỐNG KÊ TRONG NGHIÊN CỨU MÁY HỌC

- ***Sự không chắc chắn (Uncertainty):*** Luôn tồn tại trong máy học dù lượng dữ liệu và phương pháp gia tăng.
  - Học có giám sát: Dự đoán mục tiêu dựa trên thuộc tính đã biết, cố gắng dự đoán giá trị khả năng nhất hoặc định lượng sự không chắc chắn.
  - Học không giám sát: Quan tâm đến sự không chắc chắn để xác định khả năng bất thường.
  - Học tăng cường: Suy luận về ảnh hưởng của thay đổi môi trường và hệ quả kỳ vọng.
- ***Xác suất (Probability):*** Bộ môn toán học lý luận trong điều kiện không chắc chắn.
  - Quan điểm Tần suất (Frequentist): Áp dụng cho sự kiện lặp lại.
  - Quan điểm Bayes (Bayesian): Hợp lý hóa lý luận trong điều kiện không chắc chắn, mức độ tin tưởng vào sự kiện không lặp lại, tính chủ quan.
- ***Thống kê (Statistics):*** Suy luận ngược từ dữ liệu, tìm kiếm hình mẫu đặc trưng cho quần thể rộng hơn.

BÀI TOÁN CƠ BẢN
TUNG ĐỒNG XU CÂN BẰNG

Thí nghiệm: Tung đồng xu cân đối, lượng hóa khả năng xuất hiện mặt ngửa (h) và mặt sấp (t).

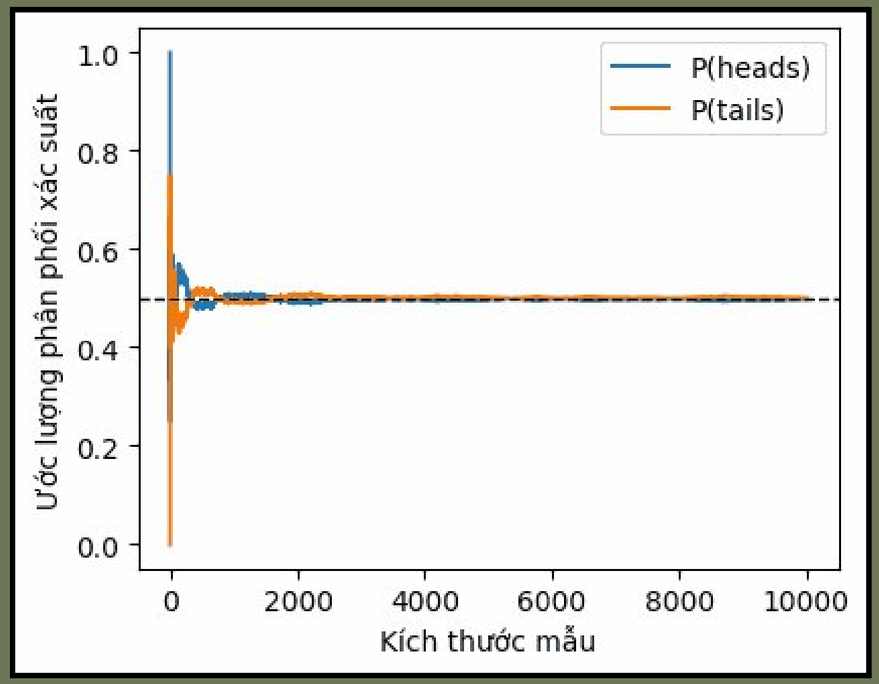
Đại lượng lý thuyết: xác suất P(heads) = 0.50.

Thống kê: Đại lượng thực nghiệm từ dữ liệu n_h/n.

Tính nhất quán (Consistency): Ước tính hội tụ về xác suất tương ứng khi dữ liệu tăng.

Luật số lớn (Law of Large Numbers): Khi số lần lặp lại tăng, giá trị ước lượng hội tụ về xác suất cơ bản.

Định lý Giới hạn trung tâm (Central Limit Theorem - CLT): Sai số giảm theo tỷ lệ $1 / \sqrt{n}$.



CÁC KHÁI NIỆM CƠ BẢN VỀ XÁC SUẤT

- **Không gian mẫu (Sample Space):** Tập hợp tất cả kết quả duy nhất có thể xảy ra. Ví dụ: Tung 1 đồng xu S = {ngửa,sấp}. Tung 1 xúc xắc S = {1,2,3,4,5,6}.
- **Biến cố (Event):** Một tập hợp con của không gian mẫu.

Ví dụ: Xúc xắc ra mặt lẻ A = {1,3,5}.

- **Hàm xác suất P:** Gán giá trị thực trong [0,1] cho biến cố.
- **Ba tiên đề của xác suất (Kolmogorov):**
  - $0 \leq P(A) \leq 1$
  - $P(S) = 1$
  - $P(U_{i=1}^{\infty} A_i) = \Sigma_{i=1}^{\infty} P(A_i)$ với A_i loại trừ lẫn nhau
- **Phần bù (Complement):** $P(A') = 1 - P(A)$

BIẾN NGẪU NHIÊN

RANDOM VARIABLES

Định nghĩa: Phép ánh xạ từ không gian mẫu sang một tập hợp các giá trị. Ký hiệu P(X=v).

Phân loại:

- **Biến ngẫu nhiên rời rạc (Discrete):** Nhận giá trị hữu hạn hoặc đếm được (ví dụ: kết quả tung xúc xắc).
- **Biến ngẫu nhiên liên tục (Continuous):** Nhận vô số giá trị trong một khoảng (ví dụ: chiều cao, cân nặng).

Hàm mật độ xác suất (Probability density function - PDF):

Dùng cho biến ngẫu nhiên liên tục.

Xác suất chiều cao nằm trong khoảng [1.79m, 1.81m] được tính bằng tích phân của mật độ trên khoảng đó.

ĐA BIẾN NGẪU NHIÊN VÀ ĐỊNH LÝ BAYES

Tương tác giữa nhiều biến: Hầu hết thuật toán ML liên quan đến nhiều biến ngẫu nhiên. Biết giá trị một biến có thể cập nhật niềm tin về biến khác (ví dụ: triệu chứng bệnh và khả năng nhiễm bệnh).

Hàm xác suất đồng thời (Joint Probability Function): $P(X = x, Y = y)$

$$P(X = x, Y = y) \leq P(X = x) \text{ và } P(X = x, Y = y) \leq P(Y = y)$$

Xác suất có điều kiện (Conditional Probability):

$$P(Y = y|X = x) = P(X = x, Y = y) / P(X = x)$$

Định lý Bayes (Bayes' Rule):

$$P(A|B) = P(B)P(B|A) / P(A)$$

Dạng đầy đủ: $P(A|B) = P(B|A)P(A) / [P(B|A)P(A) + P(B|A')P(A')]$

Thomas Bayes
(1701-1761)

ĐA BIẾN NGẪU NHIÊN VÀ ĐỊNH LÝ BAYES

Ví dụ minh họa: Xét nghiệm COVID-19. Dữ liệu:
Dương tính giả: $P(D_1 = 1|C = 0) = 0.01$
Dương tính thật: $P(D_1 = 1|C = 1) = 1$
Xác suất tiên nghiệm: $P(C = 1) = 0.0015$
Xác suất nhiễm bệnh khi xét nghiệm lần 1 dương tính:
$P(C = 1|D_1 = 1)$
Xét nghiệm lần 1 dương tính:
$P(C = 1|D_1 = 1) \approx 0.1306$
Xét nghiệm lần 2 dương tính:
$P(C = 1|D_1 = 1, D_2 = 1) \approx 0.8307$
Nhiều bằng chứng (độc lập có điều kiện) giúp tăng độ chắc chắn.




THAM SỐ THỐNG KÊ

Giá trị kỳ vọng (Expected value / Mean - μ):

Biến rời rạc: $\boldsymbol{\mu} = E[X] = \sum_x xP(X = x)$

Biến liên tục :$\boldsymbol{\mu} = E[X] = \int xp(x)dx$

Phương sai (Variance - σ^2):

$Var[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

Độ lệch chuẩn (Standard Deviation - σ):

$SD[X] = \sqrt{Var(X)}$

Ma trận hiệp phương sai (Covariance Matrix - Σ):

$\Sigma = E[(Z - \mu)(Z - \mu)^T]$.

SỰ KHÔNG CHẮC CHẮN

Nguồn gốc không chắc chắn: giá trị nhãn, giá trị ước tính tham số, sự khác biệt phân phối dữ liệu huấn luyện và triển khai.

Không chắc chắn ngẫu nhiên (Aleatoric Uncertainty):

Bản chất do tính ngẫu nhiên vốn có trong dữ liệu, không thể giải thích bởi các biến quan sát. Không thể giảm bằng cách thu thập thêm dữ liệu (cùng loại).

Ví dụ: Nhiễu trong phép đo của cảm biến do hạn chế phần cứng.

Không chắc chắn nhận thức (Epistemic Uncertainty):

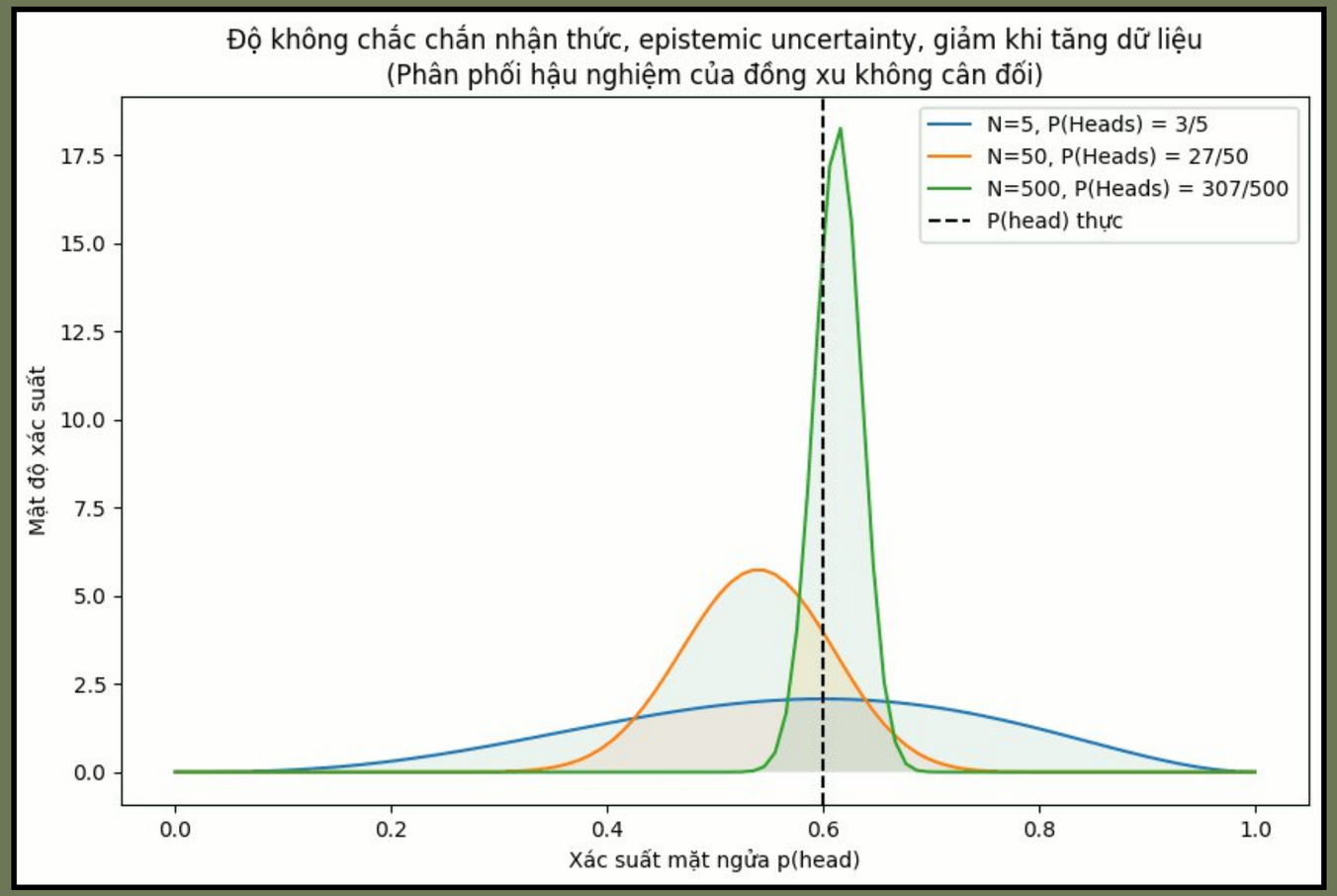
Bản chất do sự thiếu hiểu biết về mô hình hoặc tham số của mô hình. Có thể giảm bằng cách thu thập thêm dữ liệu.

Ví dụ: Ước tính xác suất mặt ngửa của đồng xu không rõ tính cân đối.

SỰ KHÔNG CHẮC CHẮN

Bài tập tình huống:

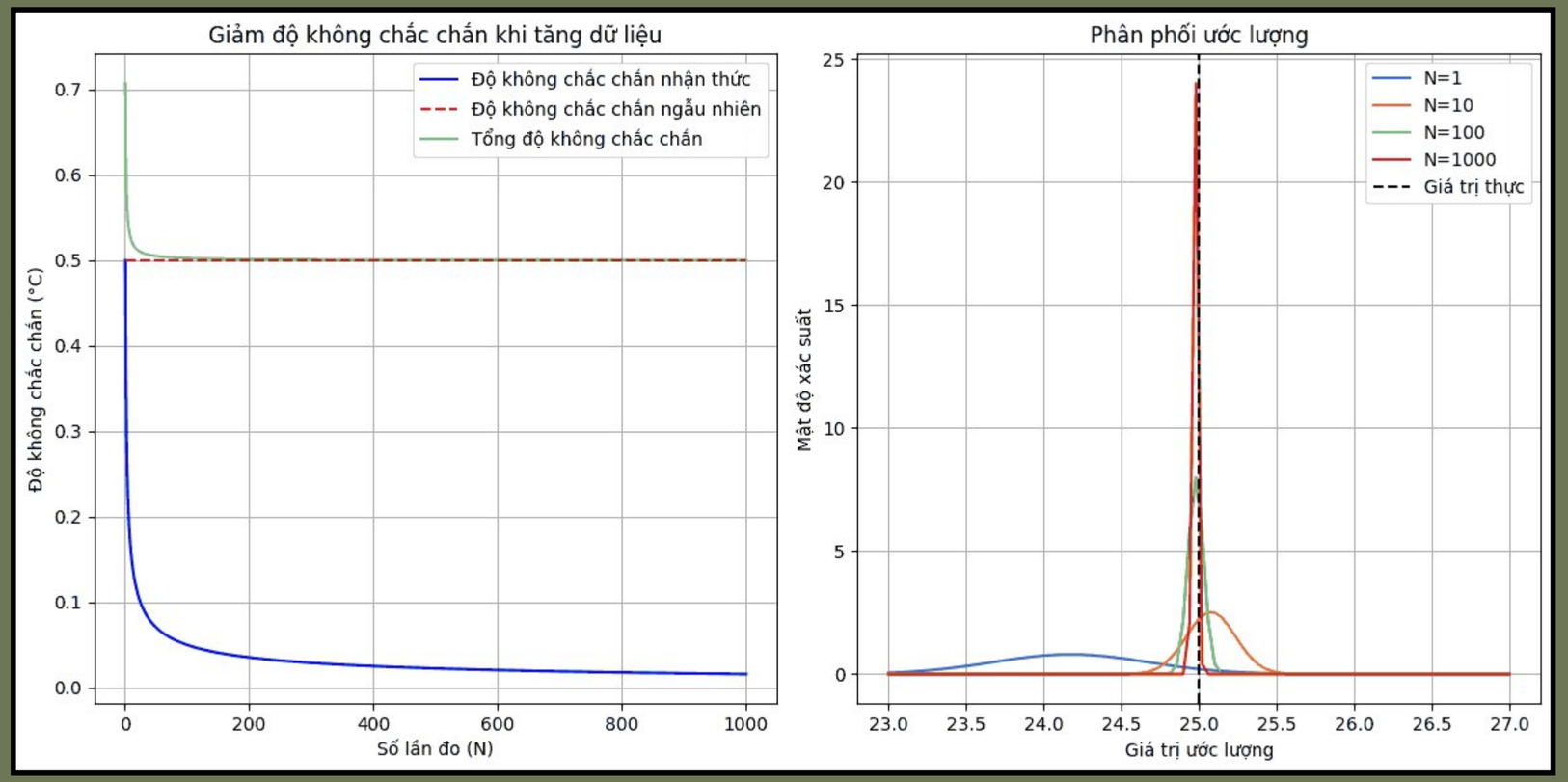
Bài tập 1 (Đồng xu): Quan sát thêm dữ liệu (tăng N) giúp ước lượng xác suất mặt ngửa hội tụ, giảm độ không chắc chắn nhận thức.



SỰ KHÔNG CHẮC CHẮN

Bài tập tình huống:

Bài tập 2 (Cảm biến nhiệt độ): Tăng số lần đo ban đầu giúp giảm sai số ước tính (giảm epistemic), nhưng đến một ngưỡng thì sai số không giảm thêm do nhiễu cố hữu của cảm biến (aleatoric).



TỔNG KẾT

- Xác suất và Thống kê là nền tảng toán học không thể thiếu, cung cấp công cụ để mô hình hóa và xử lý sự không chắc chắn trong Máy học.
- Nắm vững các khái niệm từ cơ bản (không gian mẫu, biến cố, hàm xác suất) đến nâng cao (biến ngẫu nhiên, định lý Bayes, các tham số thống kê, các loại không chắc chắn) là rất quan trọng.
- Hiểu biết này giúp xây dựng mô hình hiệu quả, đánh giá độ tin cậy của dự đoán và đưa ra quyết định dựa trên dữ liệu một cách khoa học.

TÀI LIỆU THAM KHẢO

- Morris H. DeGroot, Mark J. Schervish. 2012. “Probability and Statistics”. 4th edition.
- Robert V. Hogg, Joseph W. McKean, Allen T. Craig. 2013. “Introduction to Mathematical Statistics”. 7th edition.
- Sheldon Ross. 2020. “A first course in probability”. 10th edition.
- William Mendenhall, Robert J. Beaver, Barbara M. Beaver. 2009. “Introduction to Probability and Statistics”. 13th edition.

THẢO LUẬN

LỜI CẢM ƠN