Blog Tuần 2 – Module 2

Xác suất và Ứng dụng trong Machine Learning

Tác giả: GRID034

Tuần thứ hai của Module 2 mở đầu hành trình tư duy xác suất – một nền tảng không thể thiếu trong Trí tuệ nhân tạo (AI) và Khoa học dữ liệu. Blog này không chỉ cung cấp kiến thức lý thuyết về xác suất cơ bản mà còn dẫn dắt người học đến những ứng dụng thực tiễn như phân loại email, lọc spam, dự đoán học lực, và xây dựng bộ phân loại Naive Bayes bằng Python.

Các nội dung nổi bật trong tuần:

- **Xác suất cơ bản:** Không gian mẫu, biến cố, quy tắc cộng – nhân, xác suất có điều kiện, định lý xác suất toàn phần và Bayes.
- **Bayesian Thinking:** Hiểu bản chất của tiên nghiệm – hậu nghiệm – khả năng, và cách cập nhật kiến thức bằng dữ liệu mới.
- **Phân loại Naive Bayes:** Phân loại rời rạc (Bernoulli), liên tục (Gaussian), áp dụng cho văn bản, dữ liệu học sinh.
- **Python thực chiến:** Cài đặt Naive Bayes từ đầu, tính xác suất thủ công, ứng dụng scikit-learn cho bài toán phân loại email spam.
- **MongoDB nâng cao (tiếp nối tuần 1):** Tìm hiểu Aggregation Pipeline, các toán tử và chỉ mục tối ưu truy vấn.

Bài viết được trình bày trực quan, kết hợp giữa giải thích công thức, ví dụ minh họa và mã nguồn Python, giúp người học kết nối xác suất với AI một cách tự nhiên và thực tế.

Xác suất Cơ bản

Bùi Đức Xuân

1. Giới Thiệu Về Xác Suất

Xác suất được định nghĩa là một thước đo về khả năng xảy ra của một sự kiện [1, 2].

1.1. Thí Nghiệm và Sự Kiện

- **Thí nghiệm (Experiment)**: Việc thực hiện một tập hợp các điều kiện cơ bản để quan sát một hiện tượng nhất định.
- **Kết quả (Outcome)**: Một kết quả của một thí nghiệm.
- **Không gian mẫu (Sample Space - S)**: Tập hợp tất cả các kết quả có thể xảy ra của một thí nghiệm.
  - Ví dụ: Tung một đồng xu có không gian mẫu $S = \{\text{heads, tails}\}$.
  - Ví dụ: Gieo một con xúc xắc có không gian mẫu $S = \{1, 2, 3, 4, 5, 6\}$.
- **Sự kiện (Event)**: Một tập con của không gian mẫu.

1.2. Mối Quan Hệ Giữa Các Sự Kiện

Giả sử A và B là hai sự kiện trong cùng một thí nghiệm.

- **Kéo theo (Implication)**: ”Sự kiện A kéo theo sự kiện B” có nghĩa là nếu sự kiện A xảy ra, thì sự kiện B cũng xảy ra. Ký hiệu $A \Rightarrow B$, tương đương với $A \subseteq B$.
- **Tương đương (Equivalent)**: ”Sự kiện A bằng sự kiện B” có nghĩa là nếu $A \Rightarrow B$ và $B \Rightarrow A$. Ký hiệu $A \Leftrightarrow B$.

1.3. Các Phép Toán Trên Sự Kiện

- **Giao của các sự kiện (Intersection of events - $A \cap B$)**: Tập hợp các kết quả chung của A và B.
  - Ví dụ: Trong thí nghiệm gieo một con xúc xắc.
  - Sự kiện A: ”số gieo được là số chẵn” $\Rightarrow A = \{2, 4, 6\}$ [4, 5].
  - Sự kiện B: ”số gieo được chia hết cho 3” $\Rightarrow B = \{3, 6\}$ [4, 5].
  - Giao của A và B là $A \cap B = \{6\}$.
- **Hợp của các sự kiện (Union of events - $A \cup B$)**: Tập hợp tất cả các kết quả có trong A hoặc B (hoặc cả hai).
  - Ví dụ: Với A và B như trên [5, 6].
  - Hợp của A và B là $A \cup B = \{2, 3, 4, 6\}$.

- **Phần bù (Complements - A' hoặc A^c)**: Tập hợp tất cả các kết quả trong không gian mẫu S mà không phải là phần tử của A [2]. Nó tương ứng với việc phủ định bất kỳ mô tả nào bằng lời của sự kiện A [2]. $A' \cup A = \Omega$.

2. Định Nghĩa Xác Suất

Xác suất là thước đo mức độ có thể xảy ra của một sự kiện.

2.1. Xác Suất Cổ Điển (Classical Probability)

Được tính bằng tỷ lệ giữa số kết quả thuận lợi cho một sự kiện trên tổng số kết quả có thể xảy ra.

$$P(A) = \frac{\text{số kết quả thuận lợi}}{\text{tổng số kết quả có thể xảy ra}} = \frac{n(A)}{n(\Omega)}$$

- Ví dụ: Xác suất gieo được số chẵn trên một con xúc xắc công bằng.
- Không gian mẫu có 6 mặt, $n(\Omega) = 6$.
- A: "số chẵn" $\Rightarrow A = \{2, 4, 6\} \Rightarrow n(A) = 3$.
- $P(A) = 3/6 = 0.5$.

2.2. Xác Suất Hình Học (Geometric Probability)

Dựa trên tỷ lệ giữa độ đo (chiều dài, diện tích) của miền thuận lợi và độ đo của không gian mẫu [7, 8].

$$P(A) = \frac{\text{độ đo miền A}}{\text{độ đo miền } \Omega} \quad [7, 8]$$

- Ví dụ 1D: Một số thực X ngẫu nhiên giữa 0 và 3. Xác suất X gần 0 hơn là gần 1 [7]. Miền thuận lợi là $X \in [0, 0.5]$, tổng miền là $[4]$. $P(A) = \frac{0.5}{3} = 1/6$.
- Ví dụ 2D: Một phi tiêu được ném vào một bảng phi tiêu hình tròn, hạ cánh ngẫu nhiên trên diện tích bảng. Xác suất nó rơi gần tâm hơn là gần cạnh [8]. Trong trường hợp này, độ đo là diện tích. Nếu bán kính của miền "gần tâm" là r_1 và bán kính của toàn bộ bảng là r_2, thì $P(A) = \frac{\pi r_1^2}{\pi r_2^2}$.

3. Các Quy Tắc của Xác Suất

3.1. Quy Tắc Cộng (Addition Rule)

- **Đối với các sự kiện xung khắc (Mutually exclusive events)**: Nếu A và B không thể xảy ra cùng lúc (nghĩa là $A \cap B = \emptyset$).

$$P(A + B) = P(A) + P(B)$$

- Ví dụ: Gieo một con xúc xắc công bằng. Xác suất để A = $\{1, 5\}$ [9]. Các sự kiện $\{1\}, \ldots, \{6\}$ là rời rạc [10]. $P(\{1\}) = P(\{5\}) = 1/6$ [10]. Vì $\{1\}$ và $\{5\}$ là rời rạc, $P(A) = P(\{1, 5\}) = P(\{1\}) + P(\{5\}) = 1/6 + 1/6 = 2/6 = 1/3$.

- **Tổng quát**: Đối với bất kỳ hai sự kiện A và B nào [9].

$$P(A+B) = P(A) + P(B) - P(AB)$$

3.2. Xác Suất của Phần Bù (Complement of an event)

Đối với bất kỳ sự kiện A nào [10]:

$$P(A^c) = 1 - P(A)$$

- Ví dụ: Tìm xác suất khi gieo xúc xắc, chúng ta nhận được một số khác 1 và 6 [10].
- Gọi A: "Nhận được số 1 hoặc 6" $\Rightarrow A = \{1, 6\}$.
- $P(A) = P(\{1\}) + P(\{6\}) = 1/6 + 1/6 = 2/6 = 1/3$.
- Xác suất nhận được số khác 1 và 6 là $P(A^c) = 1 - P(A) = 1 - 1/3 = 2/3$.

3.3. Xác Suất Có Điều Kiện (Conditional Probability)

Xác suất để A xảy ra **với điều kiện** B đã xảy ra.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- Ví dụ: Một con xúc xắc công bằng được gieo [11]. Không gian mẫu $S = \{1, 2, 3, 4, 5, 6\}$.
- A: "gieo được số năm" $\Rightarrow A = \{5\} \Rightarrow P(A) = 1/6$.
- B: "gieo được số lẻ" $\Rightarrow B = \{1, 3, 5\} \Rightarrow P(B) = 3/6 = 1/2$.
- $A \cap B = \{5\} \Rightarrow P(A \cap B) = 1/6$.
- a) Tìm xác suất gieo được số năm, **biết rằng** nó là số lẻ.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = 1/3 \quad [12]$$

- b) Tìm xác suất gieo được số lẻ, **biết rằng** nó là số năm.

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)} = \frac{1/6}{1/6} = 1$$

3.4. Quy Tắc Nhân (Multiplication Rule)

$$P(AB) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

Mở rộng cho nhiều sự kiện [13]:

$$P(A_1 A_2 \dots A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 A_2) \dots P(A_n|A_1 A_2 \dots A_{n-1})$$

- Ví dụ: Trong một nhà máy có 100 sản phẩm, trong đó có 5 sản phẩm lỗi. Chọn ngẫu nhiên 3 sản phẩm. Xác suất không có sản phẩm nào bị lỗi.
- Gọi A_i là sự kiện sản phẩm thứ i được chọn không bị lỗi, với $i = 1, 2, 3$.

- $P(A_1) = 95/100$.
- $P(A_2|A_1) = 94/99$ (vì đã chọn 1 sản phẩm không lỗi).
- $P(A_3|A_1A_2) = 93/98$ (vì đã chọn 2 sản phẩm không lỗi).
- $P(A_1A_2A_3) = \frac{95}{100} \cdot \frac{94}{99} \cdot \frac{93}{98} \approx 0.8560$.

4. Định Lý Xác Suất Toàn Phần (Total Probability Theorem)

Định lý này áp dụng khi kết quả của giai đoạn 2 phụ thuộc vào kết quả của giai đoạn 1 [16]. Các kết quả của giai đoạn 1 được chia thành n tập A_i, mỗi tập chứa một số kết quả có cùng ảnh hưởng đến xác suất xảy ra của H.

4.1. Hệ Thống Đầy Đủ Các Sự Kiện

Các sự kiện $A_1, A_2, \ldots, A_n$ của một thử nghiệm được gọi là hệ thống đầy đủ nếu:

- $A_i \cap A_j = \emptyset, \forall i \neq j$ (các sự kiện xung khắc).
- $\sum_{i=1}^{n} A_i = A_1 + A_2 + \cdots + A_n = \Omega$ (tổng của chúng bao phủ toàn bộ không gian mẫu).
- $P(A_1) + P(A_2) + \cdots + P(A_n) = 1$ (tổng xác suất bằng 1).

4.2. Công Thức

Nếu $A_1, A_2, \ldots, A_n$ là một hệ thống đầy đủ các sự kiện và H là bất kỳ sự kiện nào xảy ra chỉ khi một trong các sự kiện $A_1, A_2, \ldots, A_n$ xảy ra [17].

$$P(H) = \sum_{i=1}^{n} P(A_i) \cdot P(H|A_i)$$

4.3. Ví Dụ

- **Phát hiện tiện ích (Widget Detection)**: Công ty M cung cấp 80% tiện ích cho một cửa hàng ô tô và chỉ 1% sản phẩm của họ bị lỗi. Công ty N cung cấp 20% tiện ích còn lại và 3% sản phẩm của họ bị lỗi. Nếu một khách hàng mua ngẫu nhiên một tiện ích, xác suất nó bị lỗi là bao nhiêu [18]?
  - H: ”Tiện ích bị lỗi”.
  - A_M: ”Tiện ích đến từ công ty M”, A_N: ”Tiện ích đến từ công ty N”.
  - A_M và A_N tạo thành một hệ thống đầy đủ các sự kiện.
  - $P(A_M) = 0.8$, $P(A_N) = 0.2$.
  - $P(H|A_M) = 0.01$, $P(H|A_N) = 0.03$.
  - $P(H) = P(H|A_M) \cdot P(A_M) + P(H|A_N) \cdot P(A_N) = 0.01 \cdot 0.8 + 0.03 \cdot 0.2 = 0.014$.
- **Chọn bi**: Có ba túi, mỗi túi 100 viên bi. Túi 1: 75 đỏ, 25 xanh; Túi 2: 60 đỏ, 40 xanh; Túi 3: 45 đỏ, 55 xanh. Chọn ngẫu nhiên một túi và sau đó chọn ngẫu nhiên một viên bi từ túi đã chọn. Xác suất viên bi được chọn là màu đỏ [19, 20]?

- R: ”viên bi được chọn là màu đỏ”.
- B_i: ”tôi chọn Túi i”.
- B_1, B_2, B_3 tạo thành một hệ thống đầy đủ các sự kiện.
- $P(B_1) = 1/3, P(B_2) = 1/3, P(B_3) = 1/3$.
- $P(R|B_1) = 0.75, P(R|B_2) = 0.60, P(R|B_3) = 0.45$.
- $P(R) = P(R|B_1) \cdot P(B_1) + P(R|B_2) \cdot P(B_2) + P(R|B_3) \cdot P(B_3) = 0.75 \cdot 1/3 + 0.60 \cdot 1/3 + 0.45 \cdot 1/3 = 0.60$.

5. Định Lý Bayes (Bayes' Rule)

Định lý Bayes được sử dụng để tính xác suất của một nguyên nhân (A_i) xảy ra, dựa trên việc quan sát một kết quả (H).

5.1. Các Thành Phần của Công Thức Bayes

- **Hậu nghiệm (Posterior)**: $P(c|X)$ - Xác suất của ”c” là đúng, với điều kiện ”X” là đúng.
- **Khả năng (Likelihood)**: $P(X|c)$ - Xác suất của ”X” là đúng, với điều kiện ”c” là đúng.
- **Tiên nghiệm (Prior)**: $P(c)$ - Xác suất của ”c” là đúng. Đây là kiến thức ban đầu.
- **Chuẩn hóa/Lề (Marginalization)**: $P(X)$ - Xác suất của ”X” là đúng.

5.2. Công Thức

Nếu $A_1, A_2, \ldots, A_n$ là một hệ thống đầy đủ các sự kiện và H là bất kỳ sự kiện nào với $P(H) \neq 0$:

$$P(A_i|H) = \frac{P(A_i) \cdot P(H|A_i)}{P(H)} = \frac{P(A_i) \cdot P(H|A_i)}{\sum_{j=1}^{n} P(A_j) \cdot P(H|A_j)}, \quad i = 1, 2, \ldots, n$$

5.3. Ví Dụ

- **Chọn bi (Tiếp theo)**: Giả sử chúng ta quan sát thấy viên bi được chọn là màu đỏ. Xác suất Túi 1 đã được chọn là bao nhiêu [22, 23]?
  - Từ ví dụ trước, $P(R) = 0.60$.
  - $P(B_1) = 1/3$, $P(R|B_1) = 0.75$.
  - Áp dụng Định lý Bayes:

$$P(B_1|R) = \frac{P(R|B_1) \cdot P(B_1)}{P(R)} = \frac{0.75 \cdot 1/3}{0.60} = \frac{0.25}{0.60} = \frac{5}{12} \quad [23, 24]$$

- **Phát hiện Email Spam**: Giả sử từ 'offer' xuất hiện trong 80% email spam và 10% email mong muốn. Nếu 30% email nhận được là spam và bạn nhận được một email mới chứa từ 'offer', xác suất email đó là spam là bao nhiêu?
  - Gọi A_1: ”Email là Spam”, A_2: ”Email không phải Spam” (hoặc Email mong muốn). A_1, A_2 là một hệ thống đầy đủ các sự kiện.

- H: "Email chứa từ 'offer'".
- $P(H|A_1) = 0.8$ (từ 'offer' trong spam).
- $P(A_1) = 0.3$ (30% email là spam).
- $P(A_2) = 1 - P(A_1) = 0.7$ (70% email không phải spam).
- $P(H|A_2) = 0.1$ (từ 'offer' trong email mong muốn).
- Tính $P(H)$ bằng Định lý Xác suất Toàn phần:

$$P(H) = P(A_1)\cdot P(H|A_1)+P(A_2)\cdot P(H|A_2) = 0.3\cdot 0.8+0.7\cdot 0.1 = 0.24+0.07 = 0.31$$

- Áp dụng Định lý Bayes để tìm $P(A_1|H)$:

$$P(A_1|H) = \frac{P(H|A_1)\cdot P(A_1)}{P(H)} = \frac{0.8\cdot 0.3}{0.31} = \frac{0.24}{0.31} \approx 0.774$$

Probability in AI

Dao Lam Hoang

1. Các khái niệm cơ bản

1.1 Khái niệm

- **Phép thử:** Việc thực hiện một tập hợp các điều kiện cơ bản để quan sát một hiện tượng nhất định.
  Ví dụ: Tung một con xúc xắc.
- **Không gian mẫu:** Tập hợp tất cả các kết quả có thể xảy ra.
  Ví dụ: $\{1, 2, 3, 4, 5, 6\}$ là không gian mẫu khi tung xúc xắc.
- **Biến cố:** Là một tập con của không gian mẫu.
  Ví dụ: Biến cố "ra số chẵn" là tập $\{2, 4, 6\}$.

1.2 Biến cố

- **Biến cố:** Là một tập con của không gian mẫu, đại diện cho một hoặc nhiều kết quả mong muốn. *Ví dụ:* Biến cố A là "tung xúc xắc được số chẵn" thì:

$$A = \{2, 4, 6\}$$

- **Biến cố rỗng ($\varnothing$):** Là biến cố không chứa phần tử nào, tức là không thể xảy ra. *Ví dụ:* Biến cố B là "tung xúc xắc được số 8" thì:

$$B = \varnothing$$

- **Giao của hai biến cố ($A \cap B$):** Là biến cố xảy ra khi cả hai biến cố A và B cùng xảy ra. *Ví dụ:* $A = \{2, 4, 6\}$: số chẵn $B = \{4, 5, 6\}$: số lớn hơn 3

$$A \cap B = \{4, 6\}$$

- **Hợp của hai biến cố ($A \cup B$):** Là biến cố xảy ra khi ít nhất một trong hai biến cố A hoặc B xảy ra. *Ví dụ:* $A = \{1, 2, 3\}$, $B = \{3, 4, 5\}$

$$A \cup B = \{1, 2, 3, 4, 5\}$$

- **Hai biến cố xung khắc:** Là hai biến cố không thể xảy ra đồng thời, nghĩa là $A \cap B = \varnothing$. *Ví dụ:* $A = \{1, 2\}$: số nhỏ hơn 3 $B = \{4, 5, 6\}$: số lớn hơn 3

$$A \cap B = \varnothing$$

- **Biến cố đối ($\overline{A}$):** Là biến cố gồm tất cả các phần tử trong không gian mẫu mà không thuộc A. *Ví dụ:* Không gian mẫu $\Omega = \{1, 2, 3, 4, 5, 6\}$ $A = \{1, 2, 3\}$

$$\overline{A} = \{4, 5, 6\}$$

- **Biến cố độc lập:** Hai biến cố A và B được gọi là độc lập nếu:

$$P(A \cap B) = P(A) \cdot P(B)$$

Ví dụ: Tung hai con xúc xắc. - A: "con thứ nhất ra số chẵn" $\Rightarrow P(A) = \frac{3}{6} = \frac{1}{2}$ - B: "con thứ hai ra số 5" $\Rightarrow P(B) = \frac{1}{6}$ Vì hai xúc xắc độc lập:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12}$$

2. Xác suất và các tính chất cơ bản

2.1 Khái niệm xác suất

Xác suất của một biến cố là khả năng xảy ra của biến cố đó. Ký hiệu: $P(A)$, với A là một biến cố.
Các tính chất cơ bản:

- $0 \leq P(A) \leq 1$
- $P(\Omega) = 1$ (biến cố chắc chắn)
- $P(\varnothing) = 0$ (biến cố không thể xảy ra)

2.2 Xác suất cổ điển

Nếu một thí nghiệm có n kết quả đồng khả năng xảy ra, và biến cố A có m kết quả thuận lợi, thì:

$$P(A) = \frac{m}{n}$$

Ví dụ: Tung một con xúc xắc, không gian mẫu: $\Omega = \{1, 2, 3, 4, 5, 6\}$. Biến cố A: "ra số chẵn" $= \{2, 4, 6\} \Rightarrow m = 3, n = 6$

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

2.3 Xác suất có điều kiện

Xác suất biến cố A xảy ra với điều kiện B đã xảy ra:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{với } P(B) > 0$$

Ví dụ: $A = \{2, 4, 6\}$: "ra số chẵn" $B = \{4, 5, 6\}$: "ra số lớn hơn 3"

$$A \cap B = \{4, 6\}, \quad P(A \cap B) = \frac{2}{6}, \quad P(B) = \frac{3}{6}$$

$$P(A|B) = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$$

2.4 Các phép tính trong xác suất

2.4.1 Quy tắc cộng (Additive Rule)

Nếu A và B là hai biến cố bất kỳ, thì:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Trường hợp đặc biệt: Nếu $A \cap B = \varnothing$ (hai biến cố xung khắc), thì:

$$P(A \cup B) = P(A) + P(B)$$

Ví dụ: Một lớp học có 20 học sinh nam và 15 học sinh nữ. Chọn ngẫu nhiên một học sinh.

Gọi A: ”chọn được học sinh nam” Gọi B: ”chọn được học sinh nữ”

Vì mỗi học sinh thuộc chỉ một giới tính, nên $A \cap B = \varnothing$, và:

$$P(A) = \frac{20}{35}, \quad P(B) = \frac{15}{35}$$

$$P(A \cup B) = P(A) + P(B) = \frac{20 + 15}{35} = 1$$

2.4.2 Quy tắc nhân (Multiplicative Rule)

- **Quy tắc tích tổng quát:**

  $$P(A_1, A_2, \ldots, A_n) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1, A_2) \cdots P(A_n|A_1, A_2, \ldots, A_{n-1})$$

- **Công thức xác suất giao khi biết xác suất có điều kiện:**

  $$P(A \cap B) = P(A|B) \cdot P(B)$$

- **Nếu A và B độc lập, thì:**

  $$P(A \cap B) = P(A) \cdot P(B)$$

Ví dụ: Một hộp có 3 bi đỏ và 2 bi xanh. Lấy lần lượt 2 viên bi không hoàn lại.

Gọi: - A: ”bi đầu là đỏ” - B: ”bi thứ hai là đỏ”

Ta có:

$$P(A) = \frac{3}{5}, \quad P(B|A) = \frac{2}{4} \Rightarrow P(A \cap B) = P(A) \cdot P(B|A) = \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{10}$$

3. Định lý xác suất toàn phần và Định lý Bayes

3.1 Định lý xác suất toàn phần

Giả sử không gian mẫu Ω được chia thành các biến cố $B_1, B_2, \ldots, B_n$ sao cho:
- $B_i \cap B_j = \varnothing \quad$ với $i \neq j$ (các biến cố không giao nhau),
- $\bigcup_{i=1}^{n} B_i = \Omega$ (phủ toàn bộ không gian mẫu),
- $P(B_i) > 0 \quad$ với mọi i.

Khi đó, với mọi biến cố $A \subset \Omega$, ta có công thức:

$$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)$$

Ví dụ: Có ba hộp bi:

- Hộp 1: 2 bi đỏ, 3 bi xanh
- Hộp 2: 4 bi đỏ, 1 bi xanh
- Hộp 3: 3 bi đỏ, 3 bi xanh

Chọn ngẫu nhiên một hộp (xác suất chọn mỗi hộp là $\frac{1}{3}$), sau đó lấy ngẫu nhiên 1 viên bi từ hộp đó.

Tính xác suất lấy được bi đỏ.

Gọi:

- B_1, B_2, B_3: chọn hộp 1, 2, 3 tương ứng
- A: biến cố ”lấy được bi đỏ”

Ta có:

$$P(A) = \sum_{i=1}^{3} P(B_i) \cdot P(A \mid B_i) = \frac{1}{3} \cdot \frac{2}{5} + \frac{1}{3} \cdot \frac{4}{5} + \frac{1}{3} \cdot \frac{3}{6} = \frac{1}{3}\left(\frac{2}{5} + \frac{4}{5} + \frac{1}{2}\right) = \frac{1}{3} \cdot \left(\frac{6}{5} + \frac{1}{2}\right) = \frac{1}{3} \cdot \frac{17}{10} = \frac{17}{30}$$

3.2 Định lý Bayes

Nếu $B_1, B_2, \ldots, B_n$ không giao nhau, và phủ toàn bộ Ω, với $P(B_i) > 0$, và A là biến cố có $P(A) > 0$, thì xác suất có điều kiện của B_k khi biết A được tính bởi:

$$P(B_k \mid A) = \frac{P(B_k) \cdot P(A \mid B_k)}{\sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)}$$

Ví dụ: Tiếp tục từ ví dụ trên, giả sử ta đã rút được một viên bi đỏ. Tính xác suất viên bi đó đến từ hộp 2.

Ta cần tính $P(B_2 \mid A)$, với:

$$P(B_2 \mid A) = \frac{P(B_2) \cdot P(A \mid B_2)}{P(A)} = \frac{\frac{1}{3} \cdot \frac{4}{5}}{\frac{17}{30}} = \frac{\frac{4}{15}}{\frac{17}{30}} = \frac{4}{15} \cdot \frac{30}{17} = \frac{120}{255} = \frac{8}{17}$$

4. Simple Classification

4.1 Định Lý Bayes với một feature

Giả sử ta muốn tính xác suất của một lớp $C = c$ dựa trên một đặc trưng đầu vào $X = x$. Khi đó, theo định lý Bayes:

$$P(C = c \mid X = x) = \frac{P(X = x \mid C = c) \cdot P(C = c)}{P(X = x)}$$

Ví dụ với hai lớp:

$$P(C = c_1 \mid X = x) =?, \quad P(C = c_2 \mid X = x) =?$$

4.2 Ví dụ: Kết quả học tập

Giả sử ta có dữ liệu xác suất về việc học sinh có học hay không và kết quả thi:

$$P(res = pass \mid stud = yes) = \frac{P(stud = yes \mid res = pass) \cdot P(res = pass)}{P(stud = yes)}$$
$$P(res = fail \mid stud = yes) = \frac{P(stud = yes \mid res = fail) \cdot P(res = fail)}{P(stud = yes)}$$

4.3 Mã Python minh họa

```python
# Giả sử dữ liệu xác suất đã biết:
P_stud_given_pass = 0.8
P_pass = 0.6
P_stud_given_fail = 0.3
P_fail = 0.4
P_stud = P_stud_given_pass * P_pass + P_stud_given_fail * P_fail

# Áp dụng định lý Bayes:
P_pass_given_stud = (P_stud_given_pass * P_pass) / P_stud
P_fail_given_stud = (P_stud_given_fail * P_fail) / P_stud

print("P(pass | stud = yes):", round(P_pass_given_stud, 3))
print("P(fail | stud = yes):", round(P_fail_given_stud, 3))
```

Kết quả: Nếu sinh viên học (stud = yes), thì:

$$P(pass \mid stud = yes) \approx 0.8, \quad P(fail \mid stud = yes) \approx 0.2$$

Hiểu Đúng Về Xác Suất và Ứng Dụng Trong Machine Learning: Từ Định Lý Đến Thực Tiễn

Vũ Thái Sơn

Hành trình từ xác suất cơ bản đến Naive Bayes và những điều thú vị trong AI

1. Giới thiệu: Xác suất và ứng dụng trong Machine Learning

Xác suất không chỉ là một khái niệm toán học khô khan mà còn là nền tảng cho nhiều quyết định trong đời sống và các thuật toán hiện đại như Naive Bayes trong Machine Learning. Bài viết này sẽ giúp bạn:

- Hiểu các định lý xác suất cơ bản qua ví dụ thực tế.
- Kết nối lý thuyết xác suất với ứng dụng AI, đặc biệt là Naive Bayes.
- Làm quen với công thức, mã nguồn Python và hình minh họa.

2. Xác suất cơ bản và sự kiện độc lập

2.1 Khái niệm xác suất

Xác suất của một sự kiện A là tỉ lệ số trường hợp thuận lợi cho A trên tổng số trường hợp có thể xảy ra [1]:

$$P(A) = \frac{\text{Số trường hợp thuận lợi}}{\text{Tổng số trường hợp}}$$

Ví dụ: Tung một đồng xu, xác suất ra mặt ngửa là $1/2$. Rút một lá bài từ bộ bài Tây, xác suất rút được lá Át Cơ là $1/52$.

2.2 Sự kiện độc lập và phụ thuộc

Hai sự kiện A và B là độc lập nếu việc xảy ra A không ảnh hưởng đến xác suất xảy ra B [1]:

$$P(A \cap B) = P(A) \cdot P(B)$$

Ví dụ: Tung 2 đồng xu, xác suất cả hai cùng ra ngửa là $1/2 \times 1/2 = 1/4$.

Giải thích:

Không gian mẫu gồm 4 trường hợp: HH, HT, TH, TT.
Sự kiện A: đồng xu thứ nhất ra ngửa (HH, HT) $\rightarrow P(A) = 1/2$
Sự kiện B: đồng xu thứ hai ra ngửa (HH, TH) $\rightarrow P(B) = 1/2$
$P(A \cap B)$ là xác suất cả hai đồng xu cùng ra ngửa, chỉ có trường hợp HH, nên $P(A \cap B) = 1/4$.

3. Định lý xác suất toàn phần và Bayes

3.1 Định lý xác suất toàn phần

Nếu một sự kiện có thể xảy ra qua nhiều kịch bản khác nhau, tổng xác suất là [1]:

$$P(A) = \sum_i P(B_i) \cdot P(A|B_i)$$

Ví dụ: Xác suất chọn được học sinh đeo kính khi chọn ngẫu nhiên một học sinh nam hoặc nữ.

3.2 Định lý Bayes

Định lý Bayes cho phép ”đảo ngược” xác suất có điều kiện [1, 2]:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Ý nghĩa: Giúp cập nhật xác suất của một giả thuyết A khi có thêm bằng chứng B.

Ví dụ:

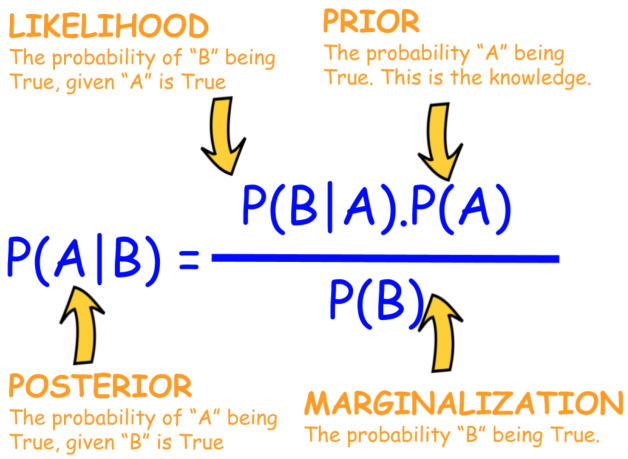
- Xác suất email là spam: $P(\text{Spam}) = 0.3$
- Xác suất email chứa từ ”free” nếu là spam: $P(\text{”free”}|\text{Spam}) = 0.4$
- Xác suất email chứa từ ”free” nếu không phải spam: $P(\text{”free”}|\text{Not Spam}) = 0.05$

$$P(\text{”free”}) = 0.4 \times 0.3 + 0.05 \times 0.7 = 0.155$$

$$P(\text{Spam}|\text{”free”}) = \frac{0.4 \times 0.3}{0.155} \approx 0.774$$

Nếu bạn nhận được email có từ ”free”, xác suất nó là spam là khoảng 77.4% [3].

3.3 Minh họa thực tế



Hình 1: Minh họa trực quan về định lý Bayes: Từ quan sát (B) suy ra nguyên nhân (A) [1].

4. Bernoulli Naive Bayes: Phân loại rời rạc

4.1 Ví dụ với dữ liệu rời rạc

Studied	Result
Yes	Pass
No	Pass
Yes	Fail
No	Fail
Yes	Pass
No	Fail

Bảng 1: Dữ liệu minh họa Bernoulli Naive Bayes [3]

Tính $P(\text{Result} = \text{Pass}|\text{Studied} = \text{Yes})$:

$$P(\text{Pass}) = 3/6 = 0.5$$

$$P(\text{Studied} = \text{Yes}|\text{Pass}) = 2/3$$

$$P(\text{Studied} = \text{Yes}) = 3/6 = 0.5$$

$$P(\text{Pass}|\text{Studied} = \text{Yes}) = \frac{2/3 \times 0.5}{0.5} = 2/3$$

5. Naive Bayes với nhiều đặc trưng

5.1 Ví dụ với 3 đặc trưng

Confident	Studied	Sick	Result
Yes	Yes	No	Pass
No	Yes	Yes	Fail
Yes	No	No	Pass
Yes	Yes	No	Pass
No	No	Yes	Fail
No	Yes	No	Fail

Bảng 2: Bảng dữ liệu minh họa cho Naive Bayes [3]

Tính xác suất một học sinh sẽ ”Pass” nếu biết: Confident = Yes, Studied = Yes, Sick = No.

$$P(\text{Pass}) = 3/6 = 0.5$$

$$P(\text{Confident} = \text{Yes}|\text{Pass}) = 3/3 = 1$$

$$P(\text{Studied} = \text{Yes}|\text{Pass}) = 2/3$$

$$P(\text{Sick} = \text{No}|\text{Pass}) = 3/3 = 1$$

$$P(\text{All}|\text{Pass}) = 1 \times \frac{2}{3} \times 1 \times 0.5 = \frac{1}{3}$$

```python
# The dataset:
# Confident | Studied | Sick | Result
# Yes | Yes | No | Pass
# No | Yes | Yes | Fail
# Yes | No | No | Pass
# Yes | Yes | No | Pass
# No | No | Yes | Fail
# No | Yes | No | Fail

p_pass = 3/6
p_conf_yes_pass = 3/3
p_studied_yes_pass = 2/3
p_sick_no_pass = 3/3

p_all_yes_pass = p_conf_yes_pass * p_studied_yes_pass * p_sick_no_pass *
    p_pass
print(p_all_yes_pass) # Output: 0.333...
```

Listing 1: Probability calculation using Python

6. Gaussian Naive Bayes: Xác suất cho dữ liệu liên tục

6.1 Công thức phân phối chuẩn

Khi dữ liệu là số thực (ví dụ: chiều dài, chiều rộng cánh hoa), xác suất được tính qua phân phối chuẩn (Gaussian) [1]:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

6.2 Ví dụ thực tế

Giả sử chiều dài cánh hoa của hai loài hoa có phân phối chuẩn với các tham số khác nhau, bạn có thể dùng Gaussian Naive Bayes để tính xác suất một bông hoa thuộc về loài nào dựa trên chiều dài đo được [1].

7. Ứng dụng Naive Bayes trong Machine Learning

7.1 Phân loại email spam với Python

```python
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer

emails = ["Free money now", "Hi friend", "Win a prize", "Meeting tomorrow"]
labels = [1, 0, 1, 0] # 1: spam, 0: not spam

vectorizer = CountVectorizer()
X = vectorizer.fit_transform(emails)

clf = MultinomialNB()
clf.fit(X, labels)

# Predict for a new email
```

```
new_mail = ["Free prize for you"]
X_new = vectorizer.transform(new_mail)
print(clf.predict(X_new)) # Output: [1]
```

Listing 2: Email classification with scikit-learn

Tham khảo: [4]

7.2 So sánh Naive Bayes và Logistic Regression

	Naive Bayes	**Logistic Regression**
Ưu điểm	Đơn giản, nhanh, hiệu quả với dữ liệu lớn	Mô hình hóa tốt mối quan hệ phi tuyến
Nhược điểm	Giả định độc lập, không phù hợp khi feature liên quan mạnh	Cần nhiều dữ liệu để ổn định
Ứng dụng	Phân loại văn bản, lọc spam	Dự đoán xác suất, phân loại nhị phân

Bảng 3: So sánh Naive Bayes và Logistic Regression [5, 6]

8. Kiến thức mở rộng và kết luận

8.1 Ứng dụng xác suất trong Deep Learning

Dù Deep Learning thường dùng các hàm mất mát phức tạp, xác suất vẫn là nền tảng cho các hàm như Cross-Entropy, Softmax. Việc hiểu xác suất giúp bạn hiểu sâu hơn về cách máy học ”ra quyết định” [2].

8.2 Đọc thêm

- *Pattern Recognition and Machine Learning* - C. Bishop [1]
- *Probabilistic Machine Learning* - K. Murphy [2]
- Naïve Bayes Classifiers, Quang-Vinh Dinh, 2025 [3]
- Scikit-learn: Machine Learning in Python [4]

Kết luận:

Việc hiểu đúng và áp dụng các định lý xác suất không chỉ giúp bạn giải quyết các bài toán đời thường mà còn mở ra cánh cửa đến thế giới Machine Learning hiện đại.

Hãy thử áp dụng Naive Bayes vào dữ liệu của bạn và khám phá điều kỳ diệu của xác suất!

Tài liệu

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[3] Q.-V. Dinh, *Naïve Bayes Classifiers*, AI Vietnam – AIO2025, 2025, lecture notes, internal document (Vietnamese).

[4] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in python," https://scikit-learn.org/stable/modules/naive_bayes.html, 2011.

[5] I. Rish, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, pp. 41–46.

[6] I. Wickramasinghe and H. Kalutarage, "Naive bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation," *Soft Computing*, vol. 25, no. 3, pp. 2277–2293, 2021.

NoSQL – MongoDB: Từ Aggregation đến Tối ưu hóa Truy vấn

Đàm Nguyên Khánh

Giới thiệu

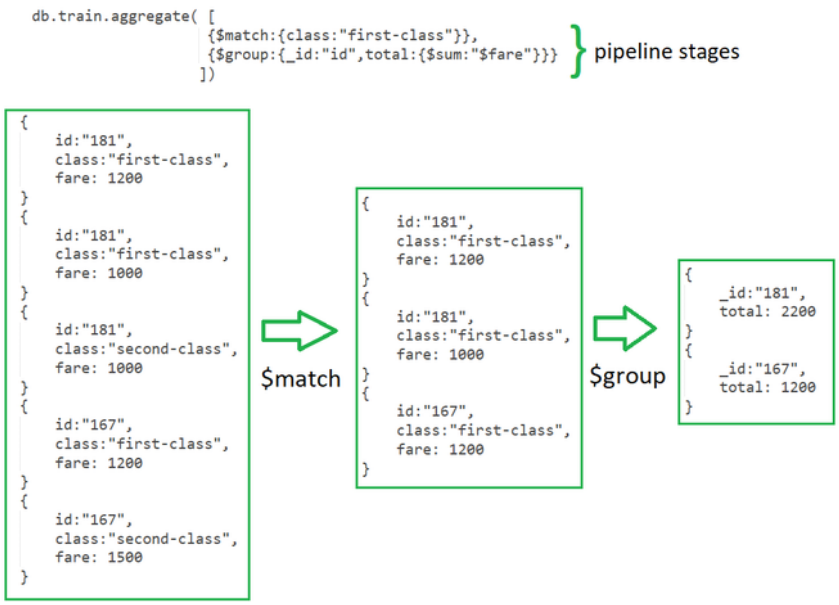
MongoDB là một cơ sở dữ liệu NoSQL phổ biến, đặc biệt hiệu quả với dữ liệu phi cấu trúc và ứng dụng thời gian thực. Trong bài viết này, chúng ta tìm hiểu về Aggregation Framework, chỉ mục (Indexes) và sử dụng PyMongo để thao tác với MongoDB qua Python.

1. Aggregation Framework

1.1 Vì sao cần Aggregation?

Mặc dù Mongo Query Language (MQL) cung cấp cú pháp đơn giản và trực quan để truy xuất dữ liệu, nó chỉ phù hợp với các truy vấn cơ bản. Đối với các nhu cầu xử lý dữ liệu phức tạp hơn như tính toán, tổng hợp, hoặc biến đổi cấu trúc document, Aggregation Framework là công cụ mạnh mẽ và linh hoạt hơn. Cụ thể, Aggregation Pipeline cho phép tổ chức truy vấn thành các giai đoạn (stages) liên tiếp, mỗi giai đoạn thực hiện một thao tác cụ thể trên tập dữ liệu trung gian. Một số thao tác phổ biến bao gồm:

- **Lọc dữ liệu** với $match: Chỉ giữ lại các document thỏa mãn điều kiện nhất định, tương tự như mệnh đề WHERE trong SQL.
- **Chiếu trường và biến đổi cấu trúc** với $project, $addFields: Chọn trường cần thiết, tạo trường mới hoặc chuyển đổi biểu diễn dữ liệu.
- **Nhóm dữ liệu** với $group: Gom các document theo một hoặc nhiều trường, và áp dụng các hàm tổng hợp như đếm, tính tổng, trung bình, v.v.
- **Thực hiện tính toán số học và logic** với các toán tử như $sum, $avg, $round, $multiply, $cond, giúp linh hoạt xử lý dữ liệu ngay trong truy vấn.



Hình 2: Minh họa Aggregation Pipeline: kết hợp $match và $group để lọc và tổng hợp dữ liệu theo ID

1.2 Các Stage cơ bản trong Aggregation Pipeline

Aggregation Pipeline bao gồm chuỗi các giai đoạn xử lý dữ liệu (gọi là *stage*). Mỗi stage nhận đầu vào là tập document trung gian từ stage trước và xuất ra một tập document mới. Dưới đây là ba stage cơ bản và thường gặp nhất:

$match – Lọc document theo điều kiện:

Stage này tương đương với mệnh đề WHERE trong SQL. Nó cho phép lọc các document thỏa mãn điều kiện cụ thể trước khi tiếp tục xử lý.

```
db.trips.aggregate([
  { $match: { "stop time": { $gt: ISODate("2016-01-05") } } }
])
```

$project – Chọn và tính toán lại trường:

Stage này dùng để chỉ định những trường nào sẽ được giữ lại hoặc tạo mới trong mỗi document. Có thể dùng kèm các biểu thức toán học, logic hoặc biến đổi kiểu dữ liệu.

```
db.trips.aggregate([
  { $project: { "tripduration_hrs": { $divide: ["$tripduration", 60] } } }
])
```

$group – Gom nhóm và tính toán tổng hợp:

Stage này nhóm các document theo một khoá chung (ví dụ: theo loại người dùng) và cho phép áp dụng các phép tổng hợp như $sum, $avg, $max,...

```
db.trips.aggregate([
  { $group: { _id: "$usertype", total: { $sum: 1 } } }
])
```

1.3 Các toán tử phổ biến

- **Số học:** $add, $divide, $round
- **Chuỗi:** $concat, $toUpper
- **Ngày:** $month, $dateDiff
- **So sánh:** $gt, $eq, $lte
- **Mảng:** $isArray, $first, $size
- **Điều kiện:** $cond, $ifNull

1.4 Các Stage nâng cao

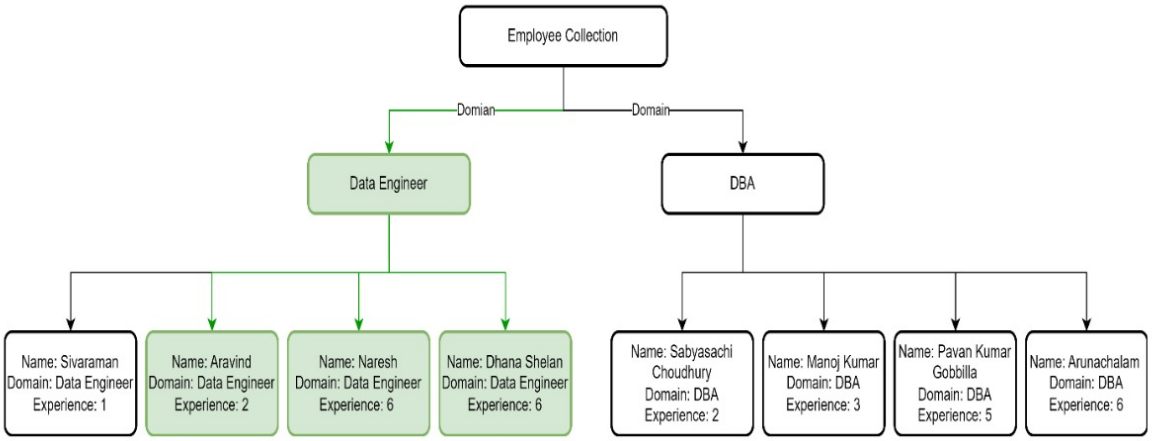
- $addFields – Thêm trường mới không xoá dữ liệu cũ
- $sort, $limit, $skip, $count – Phân trang dữ liệu
- $bucket, $bucketAuto – Phân nhóm theo khoảng giá trị
- $facet – Chạy nhiều pipeline đồng thời
- $sortByCount – Đếm số lượng theo nhóm giá trị

2. Tối ưu truy vấn với Indexes

2.1 Index là gì?

Indexes tăng tốc truy vấn bằng cách tránh duyệt toàn bộ collection. MongoDB dùng cấu trúc **B-Tree** để lưu chỉ mục.

```
db.collection.createIndex({ name: 1 })
```



Hình 3: Cây phân cấp từ Employee Collection: phân nhóm nhân viên theo trường Domain và hiển thị thông tin chi tiết.

2.2 Các loại Index quan trọng

Chỉ mục (Index) là thành phần quan trọng trong tối ưu hóa hiệu suất truy vấn dữ liệu. MongoDB hỗ trợ nhiều loại chỉ mục phù hợp với từng nhu cầu cụ thể:

- **Single field index:** Chỉ mục được tạo trên một trường duy nhất. Đây là loại chỉ mục cơ bản, thường dùng cho các truy vấn lọc theo một điều kiện cụ thể.
- **Compound index:** Chỉ mục kết hợp trên nhiều trường, hỗ trợ các truy vấn lọc hoặc sắp xếp theo nhiều tiêu chí. MongoDB sử dụng thứ tự khai báo trường trong index để xác định khả năng tận dụng.
- **Partial index:** Chỉ mục chỉ áp dụng cho một tập con document thỏa mãn điều kiện nhất định. Rất hữu ích khi truy vấn thường xuyên tập dữ liệu có điều kiện rõ ràng (ví dụ: chỉ những bản ghi có giá trị lớn hơn một ngưỡng).

Ví dụ tạo partial index cho các chuyến đi có thời lượng lớn hơn 100:

```
db.trips.createIndex(
  { tripduration: 1 },
  { partialFilterExpression: { tripduration: { $gt: 100 } } }
)
```

3. Làm việc với MongoDB qua Python (PyMongo)

3.1 Kết nối MongoDB

PyMongo là thư viện chính thức để tương tác với MongoDB thông qua Python. Đoạn mã sau thiết lập kết nối với cơ sở dữ liệu MongoDB cục bộ:

```
from pymongo import MongoClient
client = MongoClient("mongodb://localhost:27017/")
db = client.mydatabase
```

Sau khi kết nối, ta có thể thực hiện mọi thao tác CRUD (Create, Read, Update, Delete) và các truy vấn nâng cao.

3.2 Aggregation với PyMongo

Aggregation Pipeline có thể được định nghĩa dưới dạng danh sách Python (dạng dictionary) và truyền trực tiếp vào phương thức aggregate() của collection:

```
pipeline = [
  {"$match": {"tripduration": {"$gt": 100}}},
  {"$group": {"_id": "$usertype", "count": {"$sum": 1}}}
]
result = db.trips.aggregate(pipeline)
```

Kết quả trả về là một iterator chứa các document đã xử lý theo đúng các giai đoạn trong pipeline.

4. Tổng kết

MongoDB cung cấp một hệ sinh thái phong phú phục vụ cho việc lưu trữ, xử lý và truy vấn dữ liệu phi quan hệ. Bảng dưới đây tóm tắt các thành phần chính đã được trình bày và vai trò của chúng trong quá trình xây dựng hệ thống dữ liệu hiệu quả:

Thành phần	Mô tả và Mục tiêu
Aggregation Framework	Cung cấp các công cụ mạnh mẽ để xử lý dữ liệu phức tạp theo chuỗi các bước (pipeline stages). Cho phép lọc ($match), chiếu ($project), nhóm ($group), tính toán ($sum, $avg, $round), phân nhóm động ($bucket, $facet) và nhiều thao tác khác. Đây là thành phần tương đương với các phép toán trong GROUP BY, HAVING, và SELECT của SQL.
Indexes	Giúp tăng tốc độ truy vấn dữ liệu bằng cách tạo cấu trúc dữ liệu phụ (B-Tree) để tra cứu nhanh hơn, thay vì phải duyệt toàn bộ collection. MongoDB hỗ trợ nhiều loại chỉ mục: chỉ mục đơn, chỉ mục kết hợp (compound), chỉ mục điều kiện (partial index) nhằm tối ưu các trường hợp truy vấn cụ thể và tiết kiệm không gian lưu trữ.
PyMongo (MongoDB Driver cho Python)	Thư viện chính thức giúp kết nối và thao tác MongoDB thông qua Python. Cung cấp đầy đủ khả năng thực hiện CRUD (Create, Read, Update, Delete), chạy aggregation pipeline, và thực hiện các thao tác nâng cao như indexing, filter động, xử lý dữ liệu trả về dưới dạng dictionary. Đây là cầu nối thiết yếu giữa ứng dụng và cơ sở dữ liệu trong các hệ thống sử dụng Python.