

NGHIÊN CỨU VỀ CÁC THUẬT TOÁN HỌC ĐỂ XẾP HẠNG  
(LEARN TO RANK) TRONG NGỮ CẢNH ÁP DỤNG CÁC MÔ  
HÌNH HỌC MÁY CHO BÀI TOÁN XẾP HẠNG CÁC TÀI LIỆU TRẢ  
VỀ CỦA MỘT HỆ THỐNG TRUY HỒI THÔNG TIN

Môn học: Khai thác thông tin

Giảng Viên: Phạm Thế Anh Phú



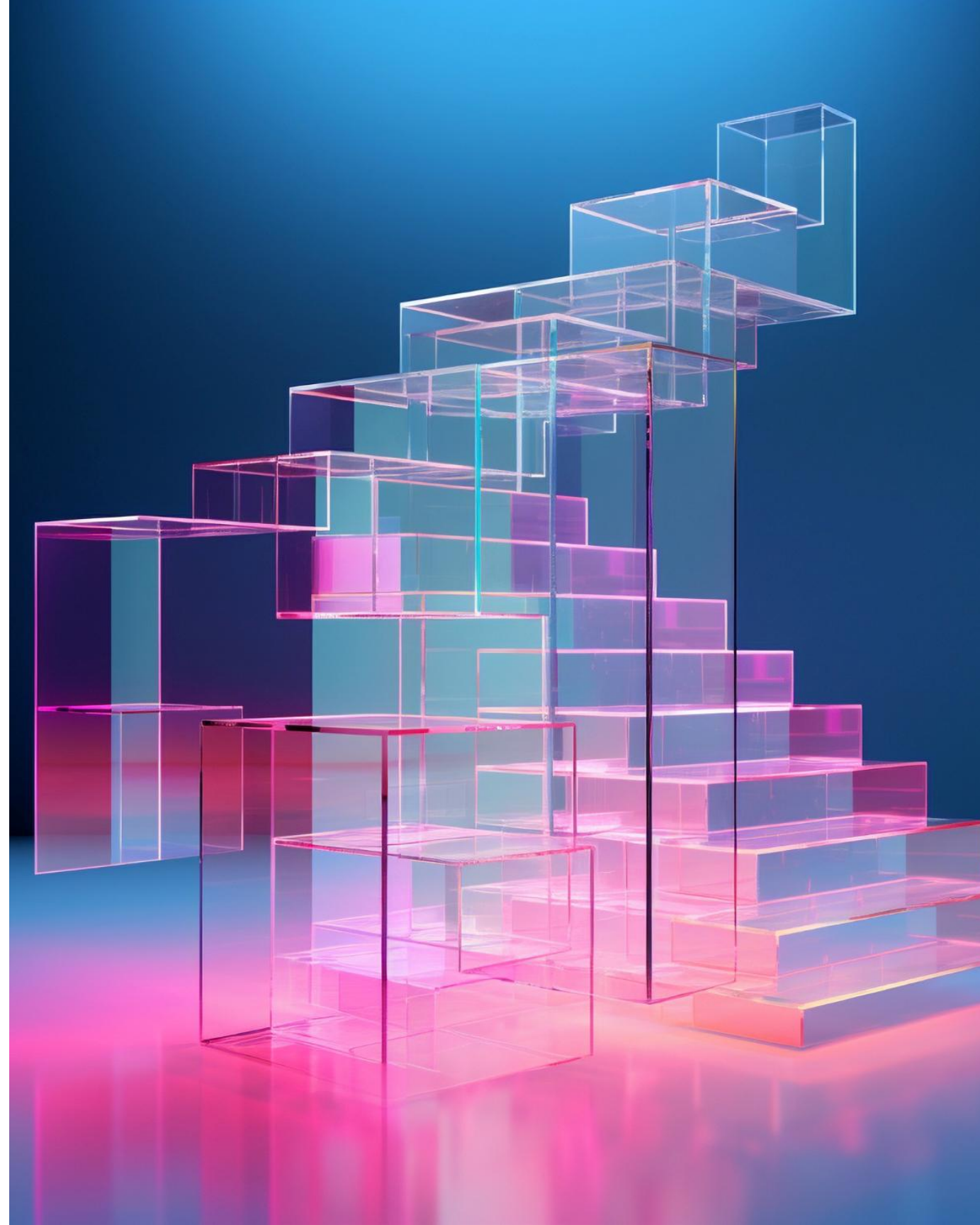
# Thành viên nhóm

- Lương Xuân Nhất
- Lê Quý Thiên
- Hồ Nhân Phước
- Đặng Thanh Hoà
- Lục Gia Yển



# Nội dung

1. Tổng quan về **Learning to Rank** (L2R)
2. Mô tả về **bài toán L2R** & Các **phương pháp tiếp cận**
3. Giới thiệu một số **thuật toán học máy nổi bật**
4. **Triển khai** huấn luyện học máy bằng **RankNet**



# 1. Tổng quan về (Learn to Rank)



Christopher Burges



Hang Li



Tie-Yan Liu



Thorsten Joachims

# Khái Niệm Cơ Bản về Learn to Rank (LTR)

Learn to Rank (LTR) là một tập hợp các kỹ thuật học máy được thiết kế để tự động học cách xếp hạng danh sách các mục. Thay vì xếp hạng từng mục riêng lẻ một cách độc lập, LTR tập trung vào việc tối ưu hóa toàn bộ thứ tự của danh sách.



## Kỹ Thuật Học Máy

LTR là một nhánh của học máy, chuyên biệt cho các bài toán mà đầu ra là một danh sách được sắp xếp theo mức độ phù hợp.



## Mục Tiêu

Mục tiêu chính là học một hàm xếp hạng (ranking function) để sắp xếp các đối tượng (tài liệu, sản phẩm, v.v.) theo cách tối đa hóa mức độ phù hợp với truy vấn hoặc sở thích của người dùng.



## Ứng Dụng Rộng Rãi

Phổ biến trong các hệ thống tìm kiếm thông tin, hệ thống gợi ý, quảng cáo trực tuyến, và nhiều lĩnh vực khác nơi thứ tự đóng vai trò quan trọng.

# Ứng Dụng Đa Dạng của Learn to Rank

LTR đã trở thành trái tim của nhiều hệ thống AI hiện đại, mang lại trải nghiệm cá nhân hóa và hiệu quả hơn.



## Công Cụ Tìm Kiếm

Sắp xếp hàng tỷ kết quả tìm kiếm để hiển thị những trang web, hình ảnh, hay video phù hợp nhất với truy vấn của người dùng.



Tìm trên Google

Xem trang đầu tiên tìm được

Google hỗ trợ các ngôn ngữ: [English](#) [Français](#) [繁體中文](#)

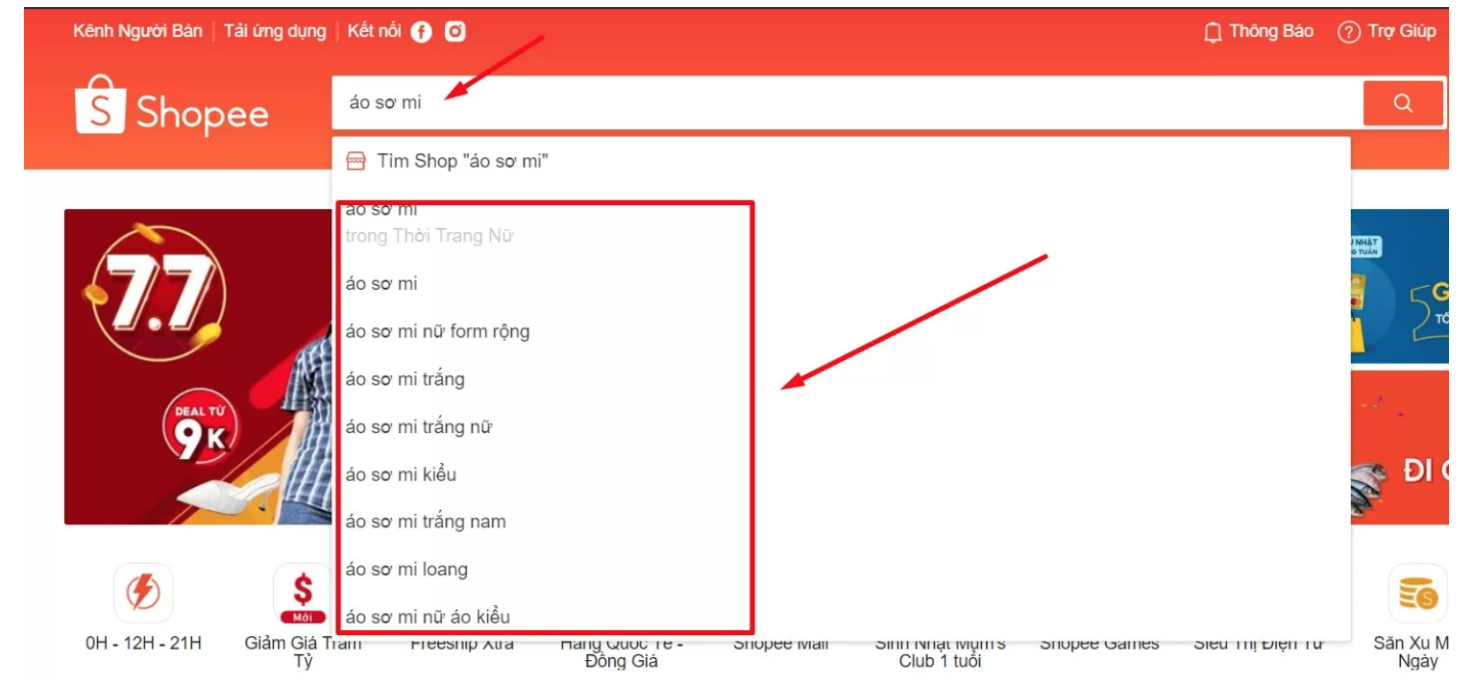
# Ứng Dụng Đa Dạng của Learn to Rank

LTR đã trở thành trái tim của nhiều hệ thống AI hiện đại, mang lại trải nghiệm cá nhân hóa và hiệu quả hơn.



## Thương Mại Điện Tử

Xếp hạng sản phẩm dựa trên khả năng được mua, giúp người dùng dễ dàng tìm thấy món đồ mong muốn và tối ưu hóa doanh số.



# Ứng Dụng Đa Dạng của Learn to Rank

LTR đã trở thành trái tim của nhiều hệ thống AI hiện đại, mang lại trải nghiệm cá nhân hóa và hiệu quả hơn.



## Mạng Xã Hội / Streaming

Đề xuất nội dung (bài đăng, phim, nhạc) dựa trên sở thích cá nhân, tăng cường sự tương tác và giữ chân người dùng.



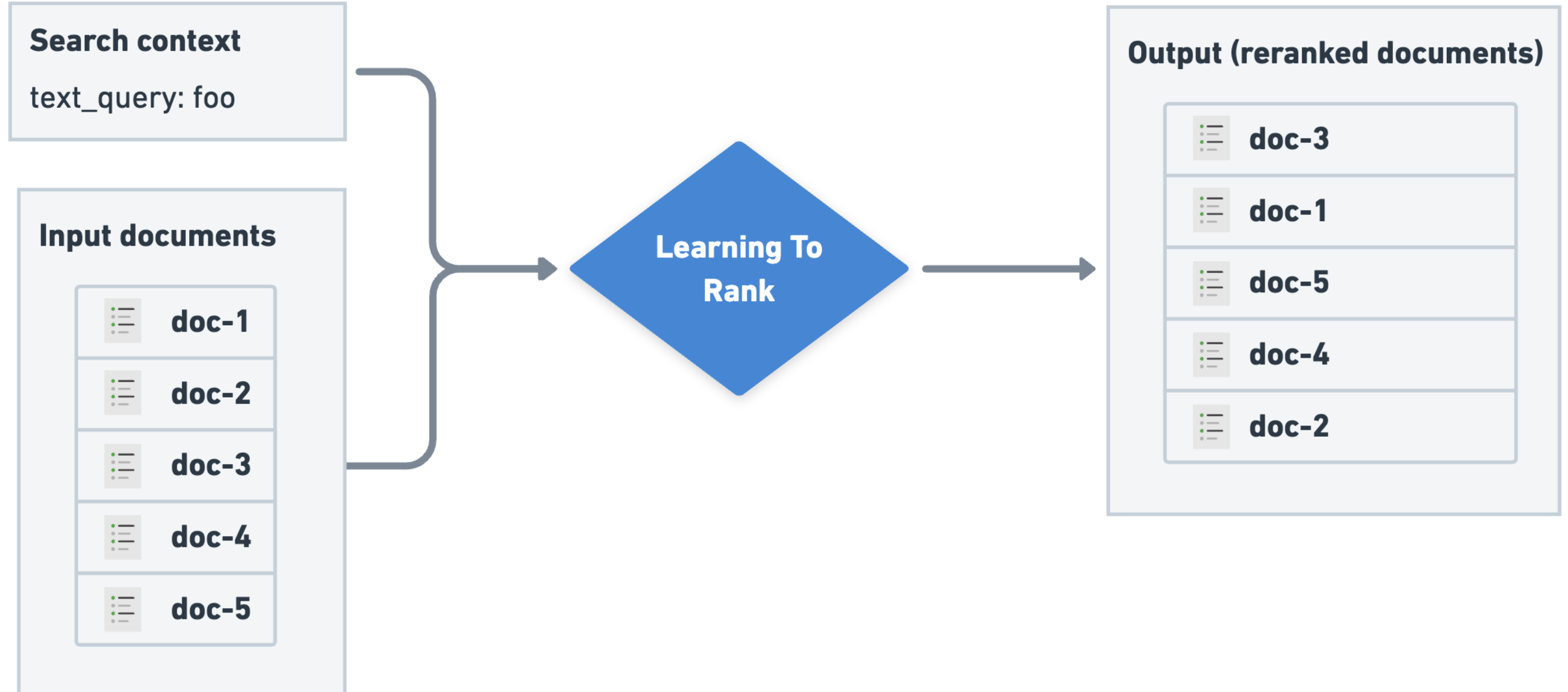
## Quảng Cáo Trực Tuyến

Tối ưu thứ tự hiển thị quảng cáo để tối đa hóa tỷ lệ nhấp chuột (CTR) và doanh thu, đảm bảo quảng cáo đến đúng đối tượng.



## 2. Mô tả về bài toán L2R

### Các phương pháp tiếp cận



# Cách L2R giải quyết bài toán

Thu thập dữ liệu

Trích xuất đặc trưng

Huấn luyện mô hình

Dự đoán & xếp hạng

Đánh giá

## Ví dụ

1

Truy vấn: "best laptop 2025"

2

Dữ liệu huấn luyện:

- (q, doc1, label=3), (q, doc2, label=2), (q, doc3, label=0)

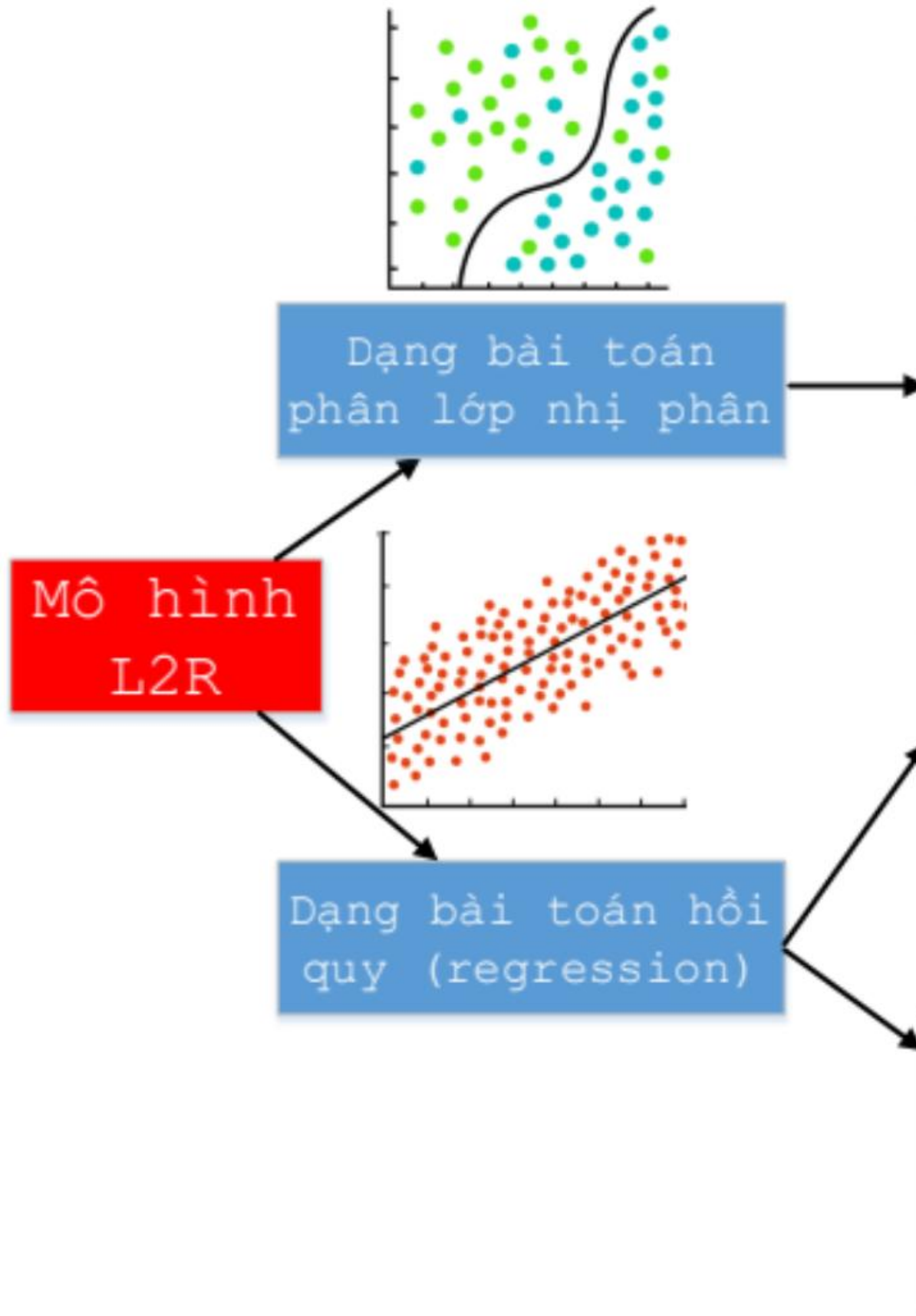
3

Mô hình học được rằng: doc1 > doc2 > doc3

4

Khi gặp truy vấn tương tự, hệ thống xếp tài liệu quan trọng nhất lên trên.

# Các Hướng Tiếp Cận Chính



## Pairwise

So sánh từng cặp đối tượng để xác định thứ tự tương đối.

## Pointwise

Xem mỗi đối tượng một cách độc lập.



## Listwise

Xem xét toàn bộ danh sách đối tượng cùng một lúc.





## Pairwise – So sánh từng cặp tài liệu

- Ý tưởng: Học cách so sánh 2 tài liệu cùng một truy vấn  $\rightarrow$  tài liệu nào tốt hơn.
- Mục tiêu học: Giảm số lượng cặp bị xếp sai thứ tự.
- Loss function: Binary Cross-Entropy dựa trên xác suất  $P(\text{docA} > \text{docB})$ .
- Ví dụ:
- Query: "apple"
  - $(\text{doc1}, \text{doc2}) \rightarrow \text{doc1}$  tốt hơn  $\text{doc2}$
  - $(\text{doc2}, \text{doc3}) \rightarrow \text{doc2}$  tốt hơn  $\text{doc3}$
  - $(\text{doc1}, \text{doc3}) \rightarrow \text{doc1}$  tốt hơn  $\text{doc3}$
- Ưu điểm: Xét mối quan hệ giữa tài liệu, thường chính xác hơn Pointwise.
- Nhược điểm: Số lượng cặp tăng nhanh  $\rightarrow$  tốn tài nguyên, chưa tối ưu toàn cục.
- Thuật toán tiêu biểu: RankNet, LambdaRank.



# Pointwise – Xem từng tài liệu một



- **Ý tưởng:** Mỗi tài liệu được đánh giá độc lập với một truy vấn.
  - **Mục tiêu học:** Dự đoán điểm số liên quan → sắp xếp theo điểm.
  - **Loss function:**
    - Hồi quy → MSE (Mean Squared Error)
    - Phân loại → Binary Cross-Entropy (BCE)
  - **Ví dụ:**
    - Query: "apple"
      - Điểm dự đoán: doc1: 0.8 – doc2: 0.6 – doc3: 0.2
      - Xếp hạng: doc1 > doc2 > doc3
  - **Ưu điểm:** Đơn giản, dễ triển khai, dùng được nhiều thuật toán ML.
  - **Nhược điểm:** Không xét quan hệ giữa tài liệu, không tối ưu trực tiếp thứ tự.
- 

# Listwise

–

## Tối ưu cả danh sách

Ý tưởng: Xem toàn bộ danh sách kết quả cho một truy vấn.

Mục tiêu học: Tối ưu trực tiếp thứ tự toàn cục theo các chỉ số như NDCG, MAP.

Loss function: ListNet, ListMLE, hoặc tối ưu trực tiếp  $\Delta$ NDCG (LambdaMART).

Ví dụ:

- Query: "apple"
  - Nhãn liên quan: doc1: 3 – doc2: 2 – doc3: 1
  - Mô hình học thứ tự tốt nhất cho cả danh sách.

Ưu điểm: Tối ưu trực tiếp ranking metrics, độ chính xác cao nhất.

Nhược điểm: Khó triển khai, tốn tính toán, cần dữ liệu gán nhãn đầy đủ.

# So Sánh Các Phương Pháp Learn to Rank

Mỗi phương pháp LTR đều có ưu và nhược điểm riêng, phù hợp với các loại dữ liệu và mục tiêu khác nhau.

Phương pháp	Đơn vị học	Mục tiêu	Độ phức tạp	Độ chính xác	Thuật toán
Pointwise	(query, doc)	Dự đoán điểm	Thấp	Thấp – Trung	Logistic Regression, RankSVM (phân loại)
Pairwise	(query, doc1, doc2)	So sánh ưu tiên	Trung bình	Trung – Cao	RankNet, LambdaRank
Listwise	(query, list)	Tối ưu toàn cục	Cao	Cao nhất	LambdaMART, ListNet, ListMLE





### 3. Giới thiệu một số thuật toán học máy nổi bật

LambdaRank

# Các Thuật Toán L2R Nổi Bật

Dưới đây là ba thuật toán L2R phổ biến và được ứng dụng rộng rãi trong thực tiễn.

Pairwise



## RankNet

Nền tảng: Sử dụng mạng neural để học một hàm xếp hạng.

Huấn luyện: Dựa trên việc so sánh các cặp tài liệu. Hàm mất mát (loss function) được thiết kế để giảm thiểu số cặp tài liệu bị xếp hạng sai.

Pairwise



## LambdaRank

Cải tiến của RankNet: Kết hợp ý tưởng từ các chỉ số đánh giá xếp hạng (như NDCG) vào quá trình tính toán gradient.

"Lambda" trong tên gọi: Đại diện cho việc điều chỉnh gradient để ưu tiên các thay đổi mang lại lợi ích lớn cho chỉ số xếp hạng tổng thể, thay vì chỉ tập trung vào lỗi cặp.

Listwise



## LambdaMART

Kết hợp: Là sự kết hợp mạnh mẽ giữa LambdaRank và Gradient Boosted Decision Trees (GBDT).

Hiệu quả: Được coi là một trong những thuật toán L2R hiệu quả nhất, được sử dụng rộng rãi trong các hệ thống tìm kiếm thương mại.

# So Sánh Chi Tiết Các Mô Hình

Tiêu chí	RankNet	LambdaRank	LambdaMART
Loại học	Pairwise	Pairwise	Listwise (thực tế)
Mô hình học	Mạng nơ-ron	Mạng nơ-ron	Cây quyết định (GBDT)
Tối ưu NDCG		 (gián tiếp)	 (gián tiếp)
Hiệu suất	Tốt	Rất tốt	Xuất sắc
Xử lý phi tuyến	Cao	Cao	Rất cao
Rủi ro quá khớp	Trung bình	Trung bình	Cao
Tốc độ huấn luyện	Nhanh	TB	Chậm (nhiều cây)

# Cơ chế hoạt động và Hàm mất mát của RankNet

## Cơ chế hoạt động

RankNet sử dụng một mạng nơ-ron truyền thẳng để tính điểm tương ứng cho mỗi tài liệu ( $s_A$  và  $s_B$ ). Sau đó, hàm sigmoid được áp dụng để chuyển đổi sự khác biệt điểm này thành xác suất  $P(A \succ B)$ , thể hiện khả năng tài liệu A xếp trên B. Quá trình huấn luyện tối ưu hóa mạng bằng cách giảm thiểu một hàm mất mát cụ thể.

## Hàm mất mát Binary Cross-Entropy

Hàm mất mát của RankNet dựa trên Binary Cross-Entropy, tính toán sự khác biệt giữa xác suất dự đoán của mô hình và xác suất thực tế (được cung cấp từ dữ liệu huấn luyện).

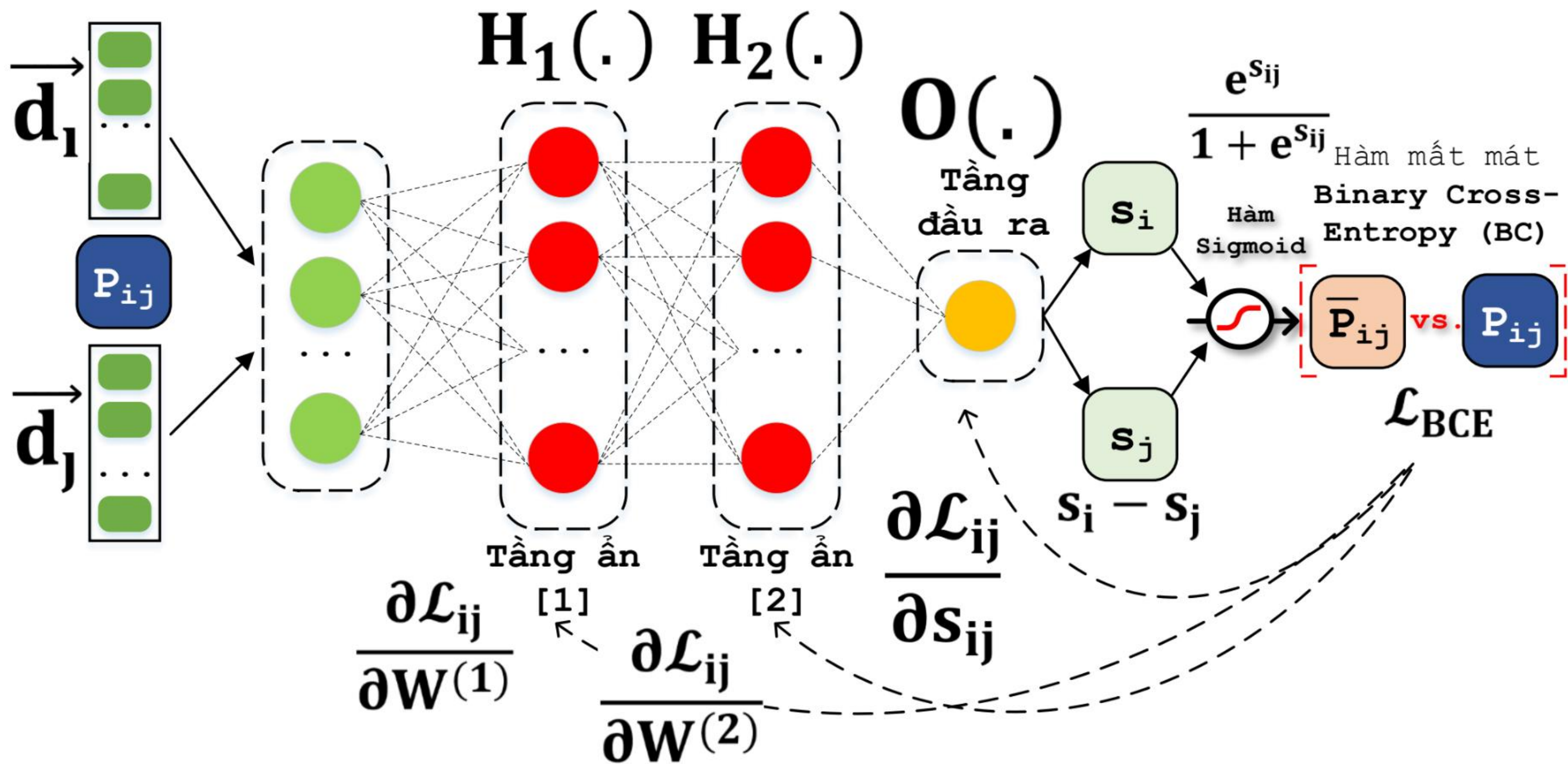
Mô hình được huấn luyện bằng cách sử dụng Backpropagation và Gradient Descent để điều chỉnh trọng số mạng, giúp giảm thiểu giá trị của hàm mất mát này và cải thiện khả năng xếp hạng.

$$C(P_{ij}, P_{ij}^{\text{predict}}) = -P_{ij} \log(P_{ij}^{\text{predict}}) - (1 - P_{ij}) \log(1 - P_{ij}^{\text{predict}})$$



## 4. Triển khai huấn luyện học máy bằng RankNet

# KIẾN TRÚC MÔ HÌNH RANKNET



# CÁCH TRIỂN KHAI

## Các bước triển khai

Dữ liệu đầu vào

16

Huấn luyện mô hình

13

Chạy demo

1

Đánh giá kết quả

6

# CÁCH TRIỂN KHAI

## Dữ liệu đầu vào

✓ Bộ tập dữ liệu ca dao, tục ngữ (CD-TN) Việt Nam

3

✓ Bộ tập dữ liệu vector hóa

6

✓ Cặp tài liệu huấn luyện (Query-document pairs)

4



# CÁCH TRIỂN KHAI

## Huấn luyện mô hình

Mô tả

Đầu vào là cặp các vector đặc trưng tài liệu ( $d_i, d_j$ ) sẽ được truyền qua các tầng ẩn và tầng đầu ra để sinh ra ( $s_i, s_j$ ) số điểm tương ứng cho cặp tài liệu đó để xếp hạng

①

Quá trình xử lý

③

Phương pháp sử dụng

①

Kết quả ở tầng output sẽ được tối ưu với hàm mất mát để tìm điểm hội tụ

Hàm mất mát (**loss function**) là "kim chỉ nam" cho quá trình học của mô hình. Nhiệm vụ của nó là đo lường mức độ "sai" của dự đoán mô hình so với "sự thật".

Tối ưu hàm mất mát

Trong RankNet, hàm **BCE** (Binary Cross-Entropy) được sử dụng để đánh giá sự khác biệt giữa xác suất dự đoán của mô hình:  $P_{ij}$  và xác suất thực tế  $P_{ij}$

# CÁCH TRIỂN KHAI

## Đánh giá kết quả

Kiểm tra hàm mất mát

So sánh độ chính xác bằng Pairwise Accuracy

là một trong những chỉ số được sử dụng để đánh giá hiệu suất của các mô hình học để xếp hạng dựa trên phương pháp cặp (pairwise learning to rank), trong đó có RankNet. Nó đo lường tỷ lệ các cặp tài liệu được mô hình sắp xếp đúng thứ tự so với thứ tự thực tế.

Tính toán Pairwise Accuracy: = Số lượng cặp được sắp xếp đúng thứ tự / Tổng số cặp có thể so sánh