Dưới đây là một bài thuyết trình cho slide về bài toán Multi-arm Bandit, bạn có thể tham khảo khi trình bày trước lớp:

**Xin chào các bạn,**

Hôm nay tôi sẽ giới thiệu về bài toán Multi-arm Bandit, một bài toán tối ưu hóa cơ bản trong học tăng cường, có ứng dụng rất rộng trong nhiều lĩnh vực.

### **1. Multi-arm Bandit - Bài toán tối ưu hóa cơ bản**

Bài toán Multi-arm Bandit (N cánh tay của tên cướp) mô phỏng một tình huống trong đó chúng ta phải chọn lựa tối ưu từ nhiều lựa chọn khác nhau. Mục tiêu của bài toán này là tối đa hóa phần thưởng, nhưng một thách thức lớn là chúng ta không biết trước kết quả của mỗi lựa chọn.

### **2. Mô phỏng tình huống**

Trong ví dụ, chúng ta có 10 máy chơi game, mỗi máy có khả năng trả thưởng khác nhau từ 0 đến 10 đô la mỗi lần chơi. Mục tiêu của chúng ta là tìm ra máy nào trả thưởng cao nhất. Tuy nhiên, để làm được điều này, chúng ta cần phải thử nghiệm và thu thập dữ liệu từ các máy để đánh giá phần thưởng kỳ vọng của mỗi máy.

### **3. Công thức tính phần thưởng**

Để tính toán phần thưởng kỳ vọng của một máy, chúng ta sử dụng công thức sau:

Trong đó:

* là các phần thưởng từ máy .
* là số lần chơi máy .

Công thức này giúp chúng ta tính được phần thưởng kỳ vọng của mỗi máy để có thể đưa ra quyết định tối ưu.

### Cập nhật giá trị trung bình khi có phần thưởng mới:

Để cập nhật, chúng ta làm như sau:

1. Cộng giá trị phần thưởng mới vào tổng.
2. Chia tổng đó cho số lượng phần tử mới (k + 1).

Công thức cập nhật giá trị trung bình mới () là:

* là số lượng phần tử đã quan sát.
* là giá trị trung bình hiện tại.
* là phần thưởng mới.

### **4. Khám phá và Khai thác**

Bài toán này đòi hỏi chúng ta phải cân bằng giữa việc khám phá (exploration) và khai thác (exploitation).

* **Khám Phá (Exploration):** Thử nghiệm các máy khác nhau để thu thập dữ liệu mới và cải thiện kiến thức.
* **Khai Thác (Exploitation):** Dựa trên những gì chúng ta đã biết, chọn máy có phần thưởng cao nhất hiện tại để tối đa hóa phần thưởng.

### **5. Chiến Lược Epsilon-Greedy**

Để giải quyết vấn đề cân bằng giữa khám phá và khai thác, chiến lược **Epsilon-Greedy** được áp dụng. Chiến lược này hoạt động như sau:

* Xác suất ϵ (epsilon): Chúng ta sẽ chọn một máy ngẫu nhiên để khám phá (tức là thử nghiệm máy khác).
* Xác suất 1 - ϵ: Chúng ta sẽ chọn máy có phần thưởng cao nhất hiện tại để khai thác.

Khi ϵ nhỏ, chúng ta khám phá ít và khai thác nhiều. Ngược lại, khi ϵ lớn, chúng ta khám phá nhiều và khai thác ít. Điều quan trọng là phải điều chỉnh ϵ sao cho hợp lý để tối ưu hóa quyết định.

"Tiếp theo, chúng ta sẽ cùng tìm hiểu về một chiến lược khác trong bài toán Multi-arm Bandit, đó là **Softmax**. Khác với chiến lược Epsilon-Greedy, nơi xác suất chọn máy ngẫu nhiên hoặc máy tốt nhất được điều chỉnh bằng một tham số ϵ, chiến lược **Softmax** tính toán xác suất lựa chọn mỗi máy dựa trên phần thưởng mà máy đó đã nhận được.

### **Softmax**

Trong chiến lược Softmax, mỗi máy chơi có một xác suất để được chọn, và xác suất này không phải là một giá trị cố định. Nó phụ thuộc vào phần thưởng trung bình của máy đó so với các máy còn lại. Cụ thể, xác suất chọn máy được tính bằng công thức:

Trong đó:

* là phần thưởng kỳ vọng của máy ,
* là một tham số điều chỉnh, gọi là **temperature**.

Khi nhỏ, chiến lược này sẽ giống như khai thác, tức là chọn máy có phần thưởng cao nhất. Còn khi lớn, xác suất chọn các máy sẽ trở nên đồng đều hơn, tương tự như việc khám phá.

### **Tóm lại:**

Softmax là một chiến lược hiệu quả giúp tối ưu hóa việc lựa chọn máy trong bài toán Multi-arm Bandit. Việc sử dụng tham số giúp điều chỉnh mức độ khám phá và khai thác, từ đó mang lại sự cân bằng tốt hơn giữa việc thu thập dữ liệu mới và tối đa hóa phần thưởng."

### **6. Ứng dụng thực tế**

Bài toán Multi-arm Bandit có rất nhiều ứng dụng trong đời sống:

* **Quảng Cáo:** Tối ưu vị trí và nội dung quảng cáo để thu hút khách hàng.
* **Đầu Tư:** Phân bổ danh mục đầu tư một cách hiệu quả.
* **Học Máy:** Cải thiện các mô hình dự đoán bằng cách tối ưu hóa quyết định.

### **7. Kết luận**

Như vậy, bài toán Multi-arm Bandit tuy đơn giản nhưng lại rất hiệu quả trong việc tối ưu hóa quyết định trong môi trường không chắc chắn. Cân bằng giữa khám phá và khai thác là yếu tố quan trọng để giúp chúng ta đưa ra các quyết định tối ưu. Và chiến lược **Epsilon-Greedy** là một công cụ hữu ích để đạt được mục tiêu này.

Cảm ơn các bạn đã lắng nghe!

Hy vọng bài thuyết trình này sẽ giúp bạn truyền đạt đầy đủ các ý tưởng trong slide!