EE675A

Lecture 1: Introduction

09/01/2023

Lecturer: Prof. Subrahmanya Swamy Peruru *Scribe: Harshvardhan Arya | Rishabh Katiyar*

1 Introduction

As we all know, the generation in which we are living is a digital book and it revolves around machines and technologies. One of the most prominent parts of these new innovations is artificial intelligence. The development of Artificial Intelligence (AI) has occurred in these generations in so-called waves. Classifying AI in such simple categories is a simplification, for which can be found numerous counter-examples. One of the significant parts of artificial intelligence is machine learning which is basically a sub-field of artificial intelligence which is broadly defined as the capability of a machine to imitate intelligent human behavior.
Machine Learning is divided into 3 major categories: Supervised Learning, **Unsupervised Learning** and **Reinforcement Learning**.

2 Supervised And Unsupervised Learning

2.1 Supervised Learning

It is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. This occurs as part of the cross-validation process to ensure that the model avoids **overfitting** or **underfitting**.

- **Example**: Let's assume we have a set of images tagged as "dog". A machine learning algorithm is trained with these dog images so it can easily distinguish whether an image is a dog or not.

2.2 Unsupervised Learning

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster **unlabeled datasets**. These algorithms discover hidden patterns or data groupings without the need for human intervention. The method's ability to discover similarities and differences in information makes it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image and pattern recognition.

- **Example**: Organize computing clusters : The geographic areas of servers is determined on the basis of clustering of web requests received from a specific area of the world. The local server will include only the data frequently created by people of that region.

3 Reinforcement Learning

Many times its difficult to obtain sufficient information about the environment that we are in and its even more difficult to gather a sufficient amount of training data to train our Supervised Machine Learning model. Reinforcement Learning is different from unsupervised learning as it is trying to maximize a reward signal instead of trying to find a hidden structure. In such scenarios or where the agent needs to consciously take actions adjusting to the environment in each step similar to a human brain, Reinforcement Learning is used. Reinforcement Learning has five main parameters: State, Action, Reward, Policy & Terminal State.

- State(S_t) - Current situation of the agent or the information available to the agent about its environment.
- Action(A_t) - This is what the agent performs.
- Reward(R_t) - It is the feedback we get from the environment after performing a certain action on the current state.
- Policy(π) - It a map from perceived states of the environment to actions to be taken when in those states. Is is of two types:
  - Deterministic Policy : a = π(s)
  - Stochastic Policy: it gives the probability to take a particular action given a particular state
$$\pi(a|s) = P[A_t = a|S_t = s] \tag{1}$$
- Terminal State(S_T) - It is the final state reached after performing a set of actions on a given environment.

An agent interacts with the environment by doing an action(A_t) and based on the rewards(R_t) it gets from the environment which is the only feedback signal for the agent to train, agent chooses the action(A_{t+1}) to take in the next time step according to the policy(π) and thus gets the reward(R_{t+1}) of the state that the agent would end up in by taking that particular action. A learning agent must be able to sense the state of its environment to some extent and must be able to take actions that affect the state. The agent also must have a goal or goals relating to the state of the environment. All the Reinforcement Learning algorithms are designed to maximize the cumulative reward as much as possible before reaching the terminal state(S_T). It is also known as goal-directed learning from interaction.

Objective - The main objective of Reinforcement Learning is to maximize the cumulative reward in all the time steps i.e. **max** $\sum_{i\geq t}^{T-1} R_{i+1}$

4 Bandits Problem in RL

Multi-Armed Bandit (MAB) is a Machine Learning framework in which an agent has to select actions (arms) in order to maximize its cumulative reward in the long term. This problem can be generalized for 1 or more than one arms and there are many ways to approach this problem in order to maximize the reward. There is no time dependence in this problem and also there is no correlation between the previous actions and the current rewards.
Let's say there are k arms which correspond to k actions $A_1, A_2, ..., A_k$ and we have to select one of the arms.

Some of the approaches to tackle the multi-arm bandit problem are:

- Explore then Commit Algorithm: We initially choose each arm for a fixed number of times and then choose the arm which was the best(had the best sample mean) for the remaining number of times. Let's say there are k arms, T total trials. So we play each arm for 'm' times and then play the best arm for the remaining T-mk trials.

- Greedy Strategy: We play each arm once and then greedily choose the arm which has the best sample mean in the subsequent trials. Let's say there are k arms and T total trials, so we choose each arm once and then greedily choose the arm for the remaining T-k trials.

$$X_t(a) = \frac{\text{sum of rewards when a taken prior to t}}{\text{number of times a taken prior to t}} = \frac{\sum_{i=1}^{t-1} R_i . \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}} \tag{2}$$

$$A_t = \arg\max_a X_t(a) \tag{3}$$

If our objective is Best Arm Identification, then we need to explore mostly. However, if our objective is to maximize the return or regret minimization, then we can use the epsilon greedy

method, which we will discuss ahead. Expected Regret is given as:

$$\mathbb{E}[\text{Regret}] = min \sum_{i=1}^{T} \mu^* - \mu(A_t) \tag{4}$$

μ is the mean reward of each arm. μ^* is the mean reward of the best arm. Our aim is to minimize this regret.

4.1 Exploration vs Exploitation Tradeoff

In this situation, the agent is in a dilemma about whether to continue to pull the arm with the best sample reward mean or explore the other sub-optimal arms at a specific time.
Exploitation is defined as a greedy approach in which agents try to get more rewards. So, in this technique, agents make the best decision based on current information.
In **exploration** techniques, agents primarily focus on improving their knowledge about each action instead of getting more rewards so that they can get long-term benefits. So, in this technique, agents work on gathering more information to make the best overall decision.
Exploitation is the right thing to do to maximize the expected reward on the one step, but exploration may produce the greater total reward in the long run

4.2 Epsilon Greedy Algorithm

Epsilon-Greedy is a simple method to balance exploration and exploitation by choosing between exploration and exploitation randomly. The epsilon-greedy, where epsilon(ϵ) refers to the probability of choosing to explore. With a small probability ϵ, we pick one arm randomly out of the k arms, and with probability $1 - \epsilon$, we pick the arm with the best sample estimate. Thus, the agent explores as well as exploits.

We took some material from [1]

References

[1] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2020.