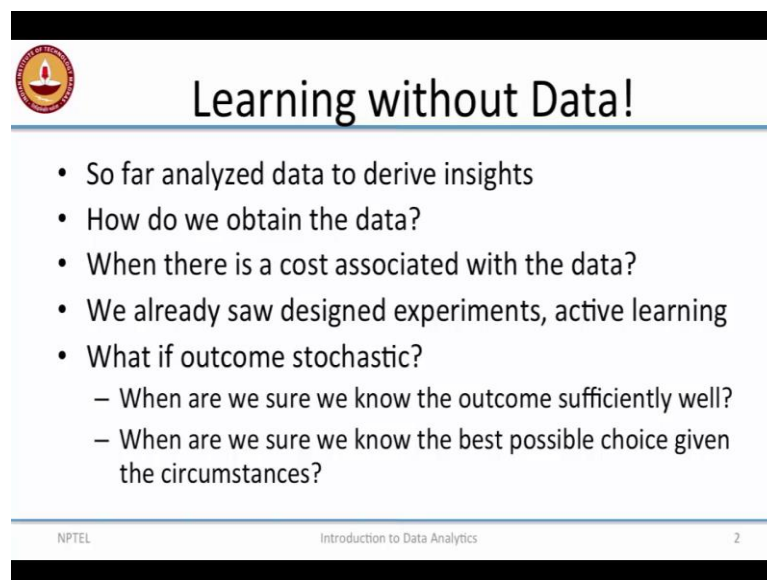


Introduction to Data Analytics
Prof. Nandan Sudarsanam and
Prof. B. Ravindran
Department of Management Studies and
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Module – 08
Lecture – 45
An Introduction to Online Learning - Reinforcement Learning

Hello and welcome to this module on Introduction to Online Learning, so where we will look at the basics of one form of Reinforcement Learning.

(Refer Slide Time: 00:20)



The slide features the IIT Madras logo in the top left corner. The title "Learning without Data!" is centered at the top. Below the title, a bulleted list of questions is presented. At the bottom, there is a footer with "NPTEL" on the left, "Introduction to Data Analytics" in the center, and the number "2" on the right.

- So far analyzed data to derive insights
- How do we obtain the data?
- When there is a cost associated with the data?
- We already saw designed experiments, active learning
- What if outcome stochastic?
 - When are we sure we know the outcome sufficiently well?
 - When are we sure we know the best possible choice given the circumstances?


NPTEL Introduction to Data Analytics 2

So, far in analyzed data to derive insights, but we never actually looked at the question of how do we obtained the data until the last couple of modules. We already saw designed experiments, we also looked at active learning, but what if there is a cost associated with the data and what if the outcome is stochastic. So, we will be finding to make measurement about the data, so we want to make a prediction about, what will happen if a patient is given a simple medicine.

But, the patient is sometimes cured sometimes he is not cured, so there is some amount of stochasticity about the outcome of this experiments or both this data measurement that you are trying to do. So, when are we sure we know the outcomes sufficiently well for us to treat it as a proper data point, when are we sure we know the best possible choice


given the circumstances. Suppose there are multiple recommendations that you can give to a patient and each of these could have an uncertain outcome. So, how are you sure, a) what is the outcome, b) which among the outcomes are, is the correct one. So, how do we go about answering this question?

(Refer Slide Time: 01:52)



Stochastic Multi-choice problems

- Which is the right drug to treat a disease?
- Which is the right advertisement to show to a user?
- Which is the right scheme to sell to a customer?
- What is the right move in a game?
- Model as a Multi-arm bandit problem



NPTEL Introduction to Data Analytics


So, these come up in the variety of settings, I was just talking to you about the drug and to treat a disease, you could think of more mundane things like, what is the right advertisement to show to a user, who comes to your web page or what is the right scheme or insurance scheme to sell to a customer whom you are trying to canvas or if you are playing a game what is the right move to make in a game, what is the right movie to recommend to the person, so there are many, many different situations, where we have similar situations arising.

So, these kinds of problems are sometimes called the stochastic multiple choice problems and are frequently modeled by, what are known as multi arm bandits. So, people might have heard of slot machines, which are gambling machines where you put a coin in and pull the lever. And then, if all the three pictures on the machine turn out to be the same, you get some kind of pay off, so you get a lot of coins back.

So, in the multi arm case you do not have a single arm to pull, a single lever to pull you have multiple levers and each lever has a different chances of success and you have to pull the each of those lever often enough to understand, which of those levers is the best

one to pull, it's a very similar to the kind of scenario that we have been talking about so far.

(Refer Slide Time: 03:26)



Problem Description

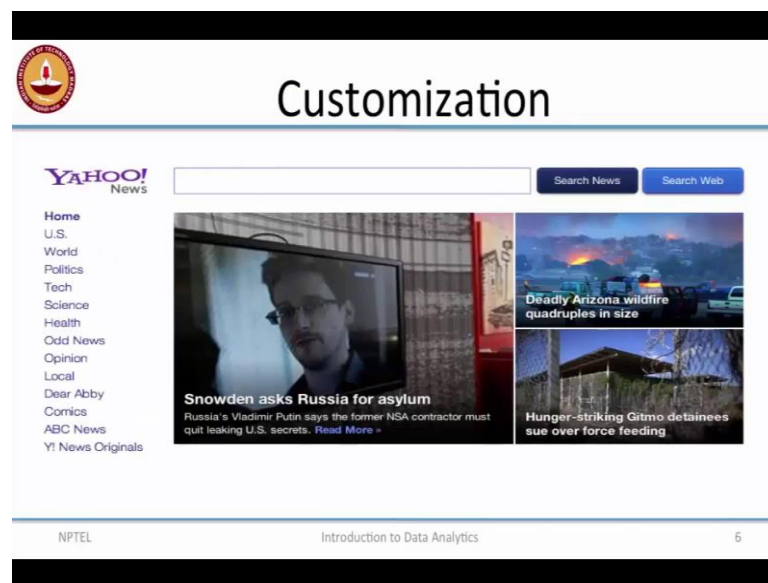
- n-arm bandit problem is to learn to preferentially select a particular action (arm) from a set of n actions $(1, 2, 3, \dots, n)$
- Each selection results in Rewards derived from the respective probability distribution
- Arm i has a reward distribution with mean μ_i and
$$\mu^* = \max \{\mu_i\}$$

NPTEL Introduction to Data Analytics 5

So, little formally, so the n arm bandit problem is to learn to preferentially select a particular action or an arm from a set of n actions, which we will conveniently labeled to 1 to n. So, each selection results in some kind of an outcome, which we call as a rewards derived from a probability distribution associated with the respective arm. So, arm could be a medicine that you give to a patient and the reward could be that the patient survives or not and the probability of survival of the patient is determined by the system.

So, what disease the patient is suffering, what are the other conditions that are prevalent at that time and what is the frequency of medicine, so on and so forth, variety of factors determine what is the probability with which the patient to recover. And, so the assumption is that each arm i is going to have a reward that is distributed with a mean μ_i and it is the one best arm and the pay off or the reward for the arm will denote by μ^* which is the max over i of all the μ_i s. This is the basic problem description.

(Refer Slide Time: 04:44)




So, you can look at a few examples of problems where this is used, here is a Yahoo news front page from some time ago. And, so they have the editor at the Yahoo news select a few tense of stories among, which the front page news items are selected and then, you could think of a reinforcement learning algorithm or a bandit like problem, where you are trying to maximize the reward. And the case of this kind of new story customization is going to be, how many times that the user clicks on a new story.

So, I have something like 20 stories to choose from, I choose one story to put make my lead story and the payoff or the reward that I am going to get is how many times you clicked on the lead story, this is like a bandit problem. So, now the question is unless I show something as the lead story, I am not going to know if the person is going to click on it or not. So, how many times should I show the news item as a lead story till I know, what is the probability of a person clicking on the lead story.

And how sure can I be that this is the best new story and that did not have to show anything else to the user, because I know, what is the right estimate for the clicks for all the other stories.

(Refer Slide Time: 06:17)



Explore-Exploit Dilemma

- One key question - the dilemma between exploration and exploitation
- Explore to find profitable actions
- Exploit to act according to the best observations already made
- Bandit problems encapsulate 'Explore vs Exploit'

NPTEL Introduction to Data Analytics 7

So, this basic problem is called as the dilemma between exploration and exploitation. We really want to explore to find the truly profitable action, we do not want to just stick with the first story that we show to the user, we really want to explore figure out which of the story is the best story. But, at the same time you would also like to exploit whatever knowledge we have, so that we can get a lot of reward.

So, you can get more click through this, so when do we stop exploring or start exploiting. So, bandit problems in some sense encapsulate these explore versus exploit dilemma and give us different approaches for solving this.

(Refer Slide Time: 07:02)

The slide, titled "Ad Selection", displays a screenshot of a Google search results page for the query "hotels in redondo beach". The search results show approximately 15,000,000 results in 0.29 seconds. On the left, there are navigation links for "Everything", "Images", "Maps", "Videos", "News", and "More". Below these, it says "Chennai, Tamil Nadu" and "Change location". The main content area features several advertisements for hotels in Redondo Beach, including "Hotels Redondo Beach | Embassy Suites Hilton.com", "Redondo Beach Hotels CA - Lowest price guarantee", "Hotel Deals at Expedia - Enjoy Your Trip to Redondo Beach", "Redondo Beach Hotels", "Redondo Beach Lux Hotel", and "Redondo Beach Hotels". Each ad includes a brief description and a link to the hotel's website. The slide footer contains "NPTEL", "Introduction to Data Analytics", and the number "8".

In other few more examples, so somebody types in a query or types hotels in some place and then, the search engines are going to throw a lot of ads. In fact, people leaves similar algorithm to the bandit algorithm to decide, which are the acts that need to be shown on the, so again, which can each of it we can see each of these acts is like an arm on the bandit problem and if you click on the arm you get the payoff.

(Refer Slide Time: 07:34)

The slide, titled "Recommendation", displays a screenshot of a movie recommendation interface. The main heading is "People who liked this also liked...". Below this, there is a grid of movie posters. The first row includes "Malcolm X", "The Last Emperor", "Gandhi", and "Rafana". The second row includes "The Last Emperor", "Gandhi", "Rafana", and "Malcolm X". To the right of the grid, there is a section for "Malcolm X" with a "PG-13" rating, a star rating of "☆☆☆", and the text "The biop influential". Below the grid, there are navigation buttons: "Prev 6", "Next 6", and "Next >". At the bottom right, there is a button "Add to Watchlist" and a section for "Director Stars: D". The slide footer contains "NPTEL", "Introduction to Data Analytics", and the number "9".

And you can also think about the recommendation engines, so if you people would have seen on Amazon or in other places where people like this also like this people bought

this also bought this, so these kinds of recommendations. And, so you could think of these recommendations also as bandit problems. So, if you show recommendation to user and if user clicks on it, then you get a reward.

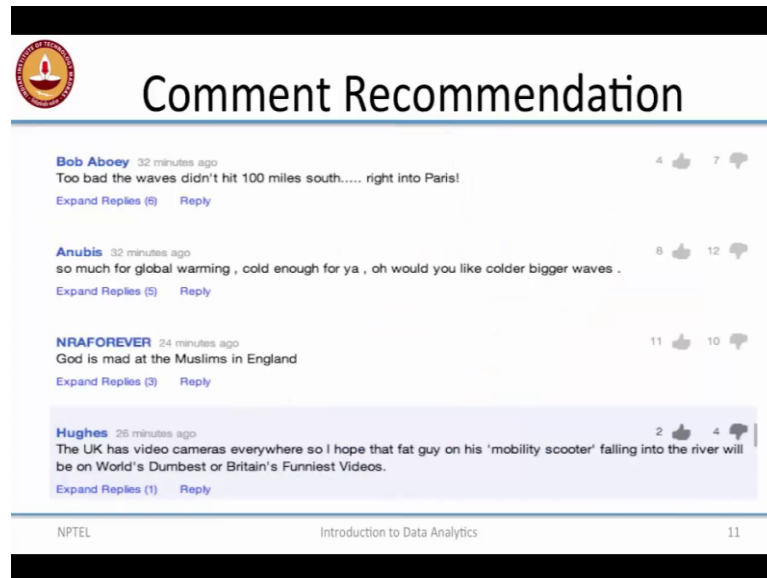
(Refer Slide Time: 08:02)



But, but really fall down a minute, is it really a bandit problem you know people typically look and think about customers, who bought this also bought this customers who bought item in your recent history also bought.

So, you could think of variety of different rules that allow you to take this kind of recommendation and we stop and think about it these rules actually could give you many, many choices and which, among those choices are you going to show to this particular user who came to your web page right now, so that can be treated as a bandit problem and people have right look at it in that fashion.

(Refer Slide Time: 08:41)



Comment Recommendation

Bob Aboey 32 minutes ago
Too bad the waves didn't hit 100 miles south..... right into Paris!
Expand Replies (6) Reply 4 7

Anubis 32 minutes ago
so much for global warming , cold enough for ya , oh would you like colder bigger waves .
Expand Replies (5) Reply 6 12

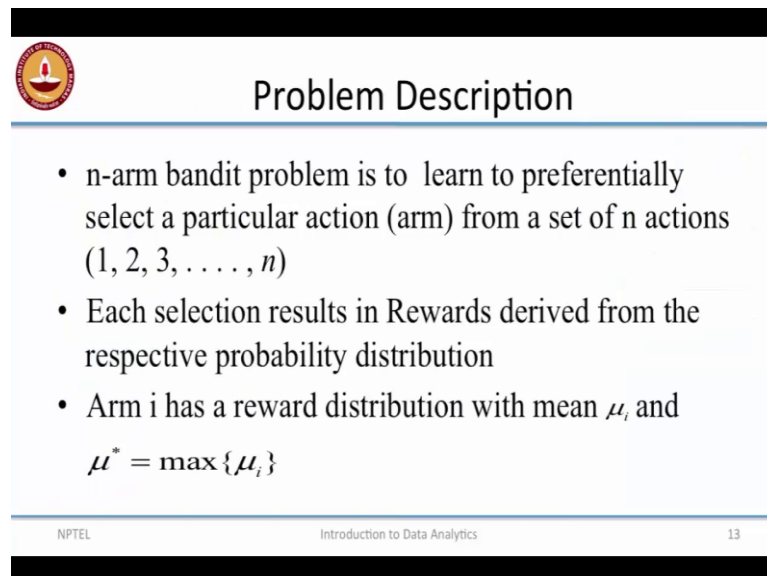
NRAFOREVER 24 minutes ago
God is mad at the Muslims in England
Expand Replies (3) Reply 11 10

Hughes 26 minutes ago
The UK has video cameras everywhere so I hope that fat guy on his 'mobility scooter' falling into the river will be on World's Dumbest or Britain's Funniest Videos.
Expand Replies (1) Reply 2 4

NPTEL Introduction to Data Analytics 11

It is the last example that we look at it is on comment recommendation. So, if you think about many blogs people are pretty active in commenting on the main new story and when a user comes to the blog page you would really like to show the user the comments on the story that are most relevant to the user. And, so once you show a comment of the user, it could give a thumbs up and a thumbs down and depending on the reaction of the user you know whether it was the good action to take or not and these thumbs up and thumbs down serve as your rewards.

(Refer Slide Time: 09:18)



Problem Description

- n-arm bandit problem is to learn to preferentially select a particular action (arm) from a set of n actions $(1, 2, 3, \dots, n)$
- Each selection results in Rewards derived from the respective probability distribution
- Arm i has a reward distribution with mean μ_i and $\mu^* = \max \{\mu_i\}$

NPTEL Introduction to Data Analytics 13

Just to recapitulate, so the main problem description is that you have n possible actions and the learner has to preferentially select a particular action. So, as to get as much reward as possible each selection results in some rewards derived from the probability distribution; that is unknown to us at upfront. And making a simplifying assumption that each arm i has a reward distribution with mean μ_i and that is one arm that gives us a mean reward of μ^* , which is the maximum μ of all mean. So, I will stop here now and then, I will continue in the next module.