

Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 16
AI and Probability; HMM

Today, we will discuss a very important topic, which is hidden Markov model, before that we would like to draw a perspective around this topic, the topic of hidden Markov model. And this perspective is about, what is the role of probability in artificial intelligence, why is it? That probability has started playing an extremely important role in artificial intelligence in recent times. And in particular natural language processing the field of statistical natural language processing, draws immensely from probability and probabilistic techniques.

When we were doing part of speech tagging, we had mentioned hidden Markov model and we had also described, how the part of speech tagging problem can be solved by hidden Markov model. We had seen that the part of speech tags are states and the words of a sentence are outputs of this machine and states, which are traversed of the machine, this sequence of states produce the tag sequence. Now, it is time to delve into the theory of hidden Markov model, in the context of natural language processing and artificial intelligence and we proceed to discuss this alright.

(Refer Slide Time: 01:51)

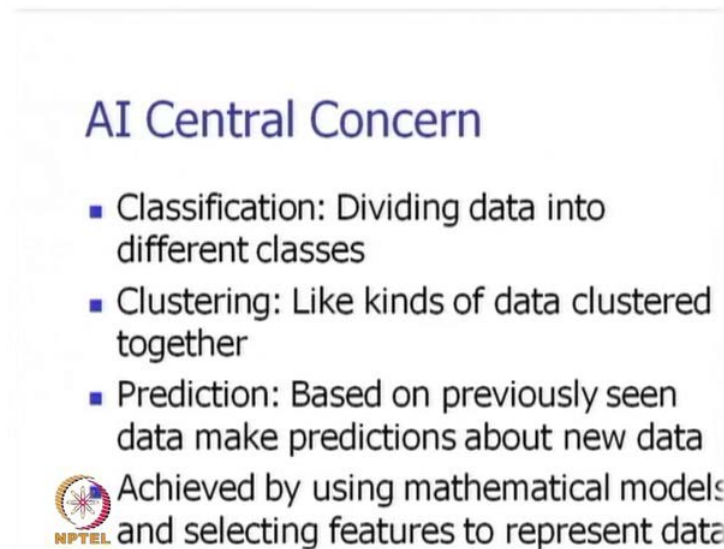
Fundamental Questions

- What is the relationship between AI and probability?
- Which situations need probability?
- Which models are used to model uncertainty?



So, the fundamental questions, which come in artificial intelligence are with respect to probability are, what is the relationship between A I and probability, which situation need probability, which models are used to model uncertainty. So, these are very basic questions at the foundation of intelligent processing and we would like to discuss them to an extent of detail.


(Refer Slide Time: 02:26)



The slide is titled "AI Central Concern" in a blue font. It contains a bulleted list of three items: "Classification: Dividing data into different classes", "Clustering: Like kinds of data clustered together", and "Prediction: Based on previously seen data make predictions about new data". Below the list is the NPTEL logo, which consists of a circular emblem with a star and the text "NPTEL" underneath. To the right of the logo, the text reads "Achieved by using mathematical models and selecting features to represent data".

AI Central Concern

- Classification: Dividing data into different classes
- Clustering: Like kinds of data clustered together
- Prediction: Based on previously seen data make predictions about new data

 Achieved by using mathematical models and selecting features to represent data

So, a central concern in artificial intelligence is the question of classification, this underlies, so many different problems of AI, that the whole problem merits a complete and exhaustive studying. Classification divides the data into 2 different classes, another problem is clustering, which is concerned with like kinds of data, being clustered together being put together.

And allied problem is prediction based on previously seen data, make predictions about the new data, this maybe a clustering problem or a classification problem in the sense that, the new data maybe given and existing class or it can be put to a particular cluster. So, these tasks are achieved by using mathematical models and selecting features to represent data.

(Refer Slide Time: 03:41)

Applications of probability

- Probabilistic planning
- Natural Language Processing
- Expert Inference
- Scene understanding



Applications of probability are probabilistic planning, natural language processing expert inference and scene understanding, what we will do is we will concentrate on n and p and see how probabilistic approaches are not only needed, but are critical.

(Refer Slide Time: 04:05)

Models for uncertainty

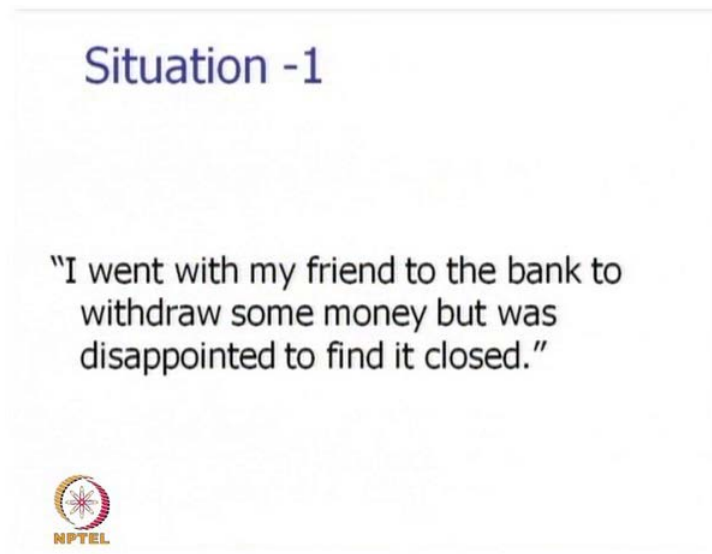
- At any point of time we have partial information about any non-trivial situation.
- Work with levels/layers
- Models
 - Fuzzy logic
 - Probability
 - Information Theory
 - Non-monotonic logic



So, this brings us to the question of, what are the models of uncertainty, at any point of time, we have partial information about any non trivial situation. So, this is the key point in any situation, which is of any level of seriousness, we have to work with partial information, this is a must and we also work at levels and layers, this is inevitable. The models, which are used for uncertainty are fuzzy logic probability information theory non-monotonic logic and so on and so forth all, these are very important models each


having its own characteristics. Now, the point to note here at this particular slide, is the point I am making about working with partial information, we will have any number of examples, but this is an inevitable situation. For any non trivial tasks whenever, we are asked to do a particular task, we find that the information that, we have to work with is very much partial.

(Refer Slide Time: 05:36)



Situation -1

"I went with my friend to the bank to withdraw some money but was disappointed to find it closed."




So, if we look at examples, we take examples from language processing and this is because language processing, we are all familiar with this task of language generation and language analysis. And discussing uncertainty in terms of natural language data, does not require any sophisticated instrument, but only a sequence of words, which need to be pondered upon. So, here I give a situation, the first situation is, I went with my friend to the bank to withdraw some money, but was disappointed to find it closed. So, this is the sentence, I went with my friend to the bank, to withdraw some money, but was disappointed to find it closed.

(Refer Slide Time: 06:29)

Situation -2

- Teacher – Student dialogue.
(Student was absent in the last class.)
Teacher (angrily): Did you miss the last class?
Student: No sir, not much.



Situation 2 is a teacher student dialogue, this is from times of India, this is presented as a kind of small joke, but look let us look at the language processing issue here, teacher student dialogue, student was absent in the last class, teacher asks him angrily did you miss the last class, student no sir not much. So, we will come back to this example, the whole punch line is based on, this particular segment of text not much.

(Refer Slide Time: 07:10)

Situation -1 (uncertainties)

- **Part of Speech ambiguity:**
- "bank" is Noun or Verb ('depend' e.g. I bank on you.)
- "closed" is Verb or Adjective
- POS ambiguity resolution:
 - Surface analysis: "bank" preceded by "the"
 - Deeper analysis: Semantics



So, we begin to analyze situation 1 and we investigate the kind of uncertainties, that prevail in situation 1 the sentence. The sentence you remember is I went with my friend

to the bank to withdraw some money, but was disappointed to find it closed, first ambiguity, we are dealing with is part of speech ambiguity ok. Part of speech is a well known concept, this are grammatical categories, like noun verb adjective adverb preposition conjunction etcetera, which are assigned to the words in a sentence.

Here, we find the in the sentence the word bank appears, first part of speech ambiguity, we mean it, on this word is that bank can be a noun or verb, the verb meaning of bank is depend for example, I bank on you, I depend on you and therefore, bank can be a noun or a verb. Similarly closed, I found the bank closed, closed can be a verb or an adjective, the closed bank denotes a the bank in a state and therefore, closed is an adjective, so it can be verb or adjective.

Now, if I ask you, what would you assign as the category of bank, I think most of you would say that, this is now and many times, I have seen that, this ambiguity is resolve from a very simple surface analysis where, the clue is that bank is preceded by the and we know that an article can precede only a noun, therefore bank is noun. So, there could be other clues like, how can verb bank be closed, only the noun bank can be closed or withdrawing money from bank withdrawing something from a verb does not make sense those things are deeper analysis. But one can see that a simple surface clue, if captured would disambiguate this, so the other kinds of analysis, which are deeper rely on semantics or the meaning content of the text or segments of text.

(Refer Slide Time: 09:51)

Situation -1 (uncertainties) ...

- **Sense ambiguity:**
- "bank" (pos: Noun) means "river bank" or "financial institution"
- "withdraw" means "take away" or "go away" (e.g. "If you cannot remain silent then withdraw from the meeting".)
- **Sense ambiguity resolution:**
 - Clue words: {money, withdraw} for bank




Next kind of ambiguity, you would like to look at is sense ambiguity also known as, the word sense disambiguation even after we have disambiguated back, in terms of this category verb or noun. There still remains another uncertainty about, what does bank do you mean, the financial intuition bank where, we deposit money or withdraw money from or does it mean, a river bank or embankment. So, there is this sense ambiguity, which again needs to be resolved and this is an uncertainty.

Similarly we look at the withdraw, which means either take away or go away, for example, you could have a sentence where, the chairman angrily saying to somebody, if you cannot remain silent then withdraw from the meeting. I withdraw from the committee, which means I resign from the committee, so withdraw also can have multiple meanings like take away go away and so on. So, the sense ambiguity resolution in this particular case, can be obtained from the clue words money and withdraw. So, these are very strong indicators that, the word bank is being used in the sense of a financial place rather than a river bank, so this is the issue of sense ambiguity.

(Refer Slide Time: 11:24)

Situation -1 (uncertainties) ...

- Lexical loss ambiguity:
 - Pronoun dropped/elision/elipsis
 - "...but <I/friend/money/bank> was disappointed."

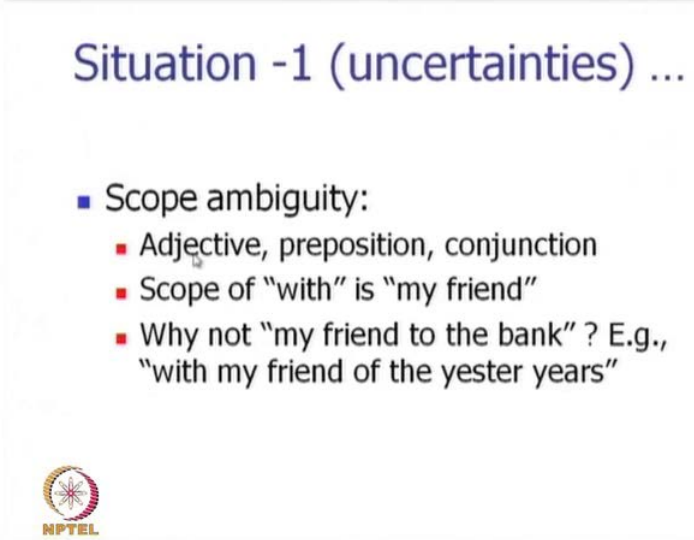


Then there is this lexical ambiguity our sentence was I went with my friend to the bank to withdraw some money, but was disappointed to find it closed. So, you see here, this particular segment or clause, but was disappointed to find it closed, is the fact that here, we have to fill in a noun, who was the disappointed I or friend or money or bank. So, this is the question to be resolved and we have to appropriately fill this blank, which is a case

of lexical loss, there is an element lexical item, which is missing from here and we have to fill it in there.


Now, one might come back and say putting money or bank there would be non sense, how can money or bank, take this slot, money or bank cannot be disappointed, but remember that, we are talking about natural language processing, we are talking about a machine being able to process textual information. And it is possible, that the machine is not endowed with common sense, the fact that money cannot be disappointed or bank cannot be disappointed comes from the fact that, these words denote entities, which are inanimate and are therefore, incapable of being disappointed. So, therefore, that leaves us with I and friend and both of them can qualify, for being disappointed and we have to put it here appropriately. So, how it is resolved is a different concern, but the main point, I am making here is that, there is an uncertainty here about, which entity should fill this slot.

(Refer Slide Time: 13:20)



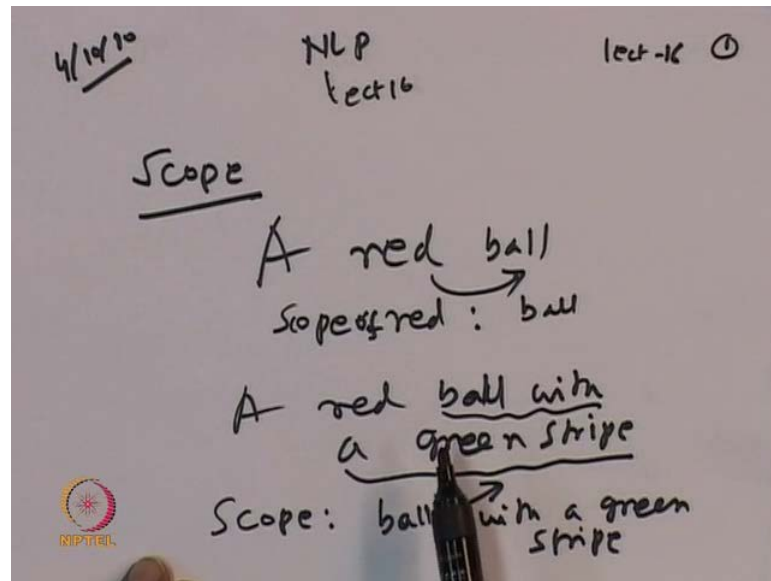
Situation -1 (uncertainties) ...

- Scope ambiguity:
 - Adjective, preposition, conjunction
 - Scope of "with" is "my friend"
 - Why not "my friend to the bank" ? E.g., "with my friend of the yester years"



Then, we come to scope ambiguity, which is sometimes say quite subtle, so adjectives prepositions conjunctions they have scope, in the sense that they apply to an extended length of text and whichever portion of text, they qualify of apply to is called the scope of these entities for example, the sentence a red ball, I will write it down.

(Refer Slide Time: 14:00)




Red ball, so the issue of scope, so if you have the sentence a red ball, then scope of red is the ball, but if you have a red ball with a green stripe then the word red qualifies the whole segment ball with a green stripes. So, the scope in this case is a ball with a green stripe. So, this shows that an adjective need not qualify only the next noun on the surface, but it could qualify a completely large phrase.

(Refer Slide Time: 15:01)

Situation -1 (uncertainties) ...

- Scope ambiguity:
 - Adjective, preposition, conjunction
 - Scope of "with" is "my friend"
 - Why not "my friend to the bank" ? E.g., "with my friend of the yester years"

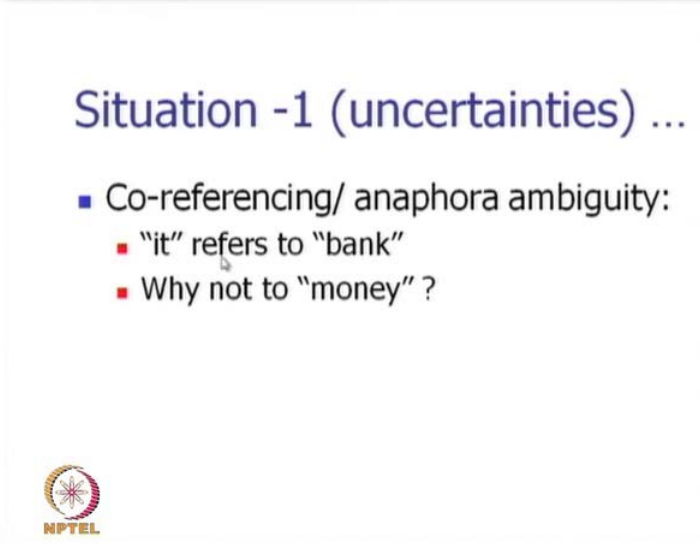


So, that is where, question of scope comes what is the scope of this adjective, so when we look at the issue of scope ambiguity, I find that adjective preposition conjunction all

of them can have scope. And therefore, in this particular sentence, which we are discussing, what is the scope of with I went with my friend to the bank to withdraw some money.


So, here the scope of with is my friend, but see that a machine can get confused, why would not it take this complete phrase my friend to the bank as the entity to be qualified for example, with my friend of the yester years. Here my friend of yester years can be a completely within the scope of with, so that ambiguity can arise.

(Refer Slide Time: 15:55)



Situation -1 (uncertainties) ...


- Co-referencing/ anaphora ambiguity:
 - "it" refers to "bank"
 - Why not to "money" ?



Next, we have this uncertainty with respect to co-referencing or anaphora as it is called, we have a pronoun it, I went with my bank to the with my friend to the bank to withdraw some money, but was disappointed to find it closed. So, which what is this it, disappointed to find it closed, it refers to bank our common sense tells us that, that is refers to bank, but it could as well refer to money, in absence of any other knowledge a machine could bind it to money. So, there are multiple possibilities of binding this pronoun to the noun a task, which is called anaphora and anaphora ambiguity, can arise very much in the process of text information gathering. So, this is called co-referencing anaphora ambiguity.

(Refer Slide Time: 16:54)

Situation -1 (uncertainties) summary		
POS	Bank (N/V)	closed (V/ adj)
Sense	Bank (financial institution)	withdraw (take away)
Pronoun drop	But	I/friend/money/bank was disappointed
SCOPE	With	my friend
Co-referencing	It ->	bank



We can now look at all the uncertainties in situation 1, which is the sentence and the discussion and we can layer these uncertainties or ambiguities. So, first level of ambiguity is at the part of speech, bank can be noun or verb closed can be verb or adjective. So, this layer is layer of pause ambiguity.

The next layer is layer of sense ambiguity, bank can be a financial institution or it could be the river bank, withdraw can be take away or to go away, in this particular case, it is take away bank is financial institution. Pronoun dropping is happening, so sentence segment is, but was disappointed, so here, we place I scope ambiguity is the next layer, which is with my friend and my friend is in the scope of with, the co-referencing layer, it gets bound to bank.

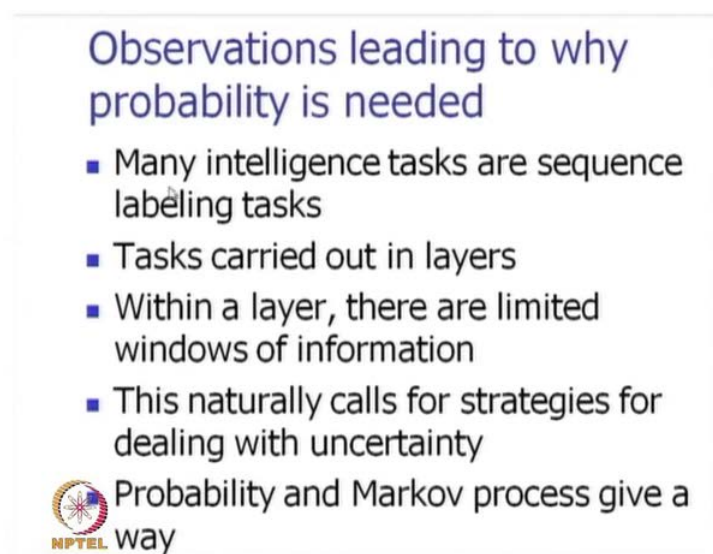
Now, a question that arises is that, are these layers, arranged really this way, is it true that, we first do part of speech disambiguation, then do sense disambiguation, then resolve all pronoun drops, then solve the scopes and then bind the co-references, it is true that, we operate this way, the answer to this is nobody knows and answer to this very fundamental cognitive question of language processing. However, when we are building a machine right now, our thinking algorithm design etcetera are limited by, but are also facilitated by this kind of separation of concerns.

So, in science and technology or the hallmark of science and technology is separation of concerns and we have separated the concerns of language processing into these levels,

which are once again part of speech tag sense processing, pronoun dropping scope and co-referencing. However, we can make one remark, about the sequences at least, if you levels have sequencing amongst themselves, for example, it is convenient to do part of speech tagging.


First and then proceed to do sense tagging, because once you have decided that, this bank cannot be a verb, I need not be concerned about its verb senses. So, that is convenient then I concentrate only, on the noun senses and then disambiguate and isolate a particular sense. Similarly, co-referencing comes pretty late, because co-referencing of an is a discourse processing task where, we take an element of text and bind it with another element of text, which is within the same clause or outside the same clause or even it is outside the text and even it could be far away from the particular sentence, which is being processed ok.

(Refer Slide Time: 20:20)



Observations leading to why probability is needed

- Many intelligence tasks are sequence labeling tasks
- Tasks carried out in layers
- Within a layer, there are limited windows of information
- This naturally calls for strategies for dealing with uncertainty

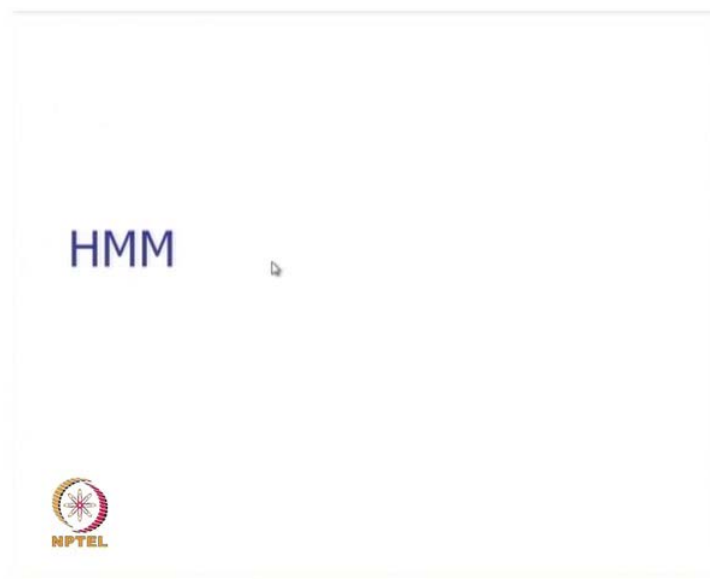
 Probability and Markov process give a way

So, these are the layers, now if you come to think of these a little deeply, we find that these are the reasons, why probabilistic framework is needed to process textual information to do natural language processing, so we record this observations. Many intelligence tasks are sequence labeling tasks for example, to understand a sentence maybe mechanically, we produce the levels on the words of the sentence in terms of their, part of speech in terms of their sense, in terms of the co-references, they should be resolved for and so on and so forth.

So, many intelligent tasks are actually sequence labeling tasks, it applies to many other branches of artificial intelligence, then it is found to be convenient to carry out the tasks in layers. Whether human beings do with completely or partially is not completely known, but machines as they stand now have to carry out this tasks in layers, within a layer there are limited windows of information. The processing can happen only taking as a small window and then doing the task of producing the levels, these windows provide the clues for various uncertainty resolution.

And since the window of information is necessarily limited it calls for strategies for dealing with uncertainty, because a information is never complete and it is here, that probability and Markov processes gives us a way of carrying out this tasks. So, I suppose the motivation for using probability and using Markov model is clear now, it namely that, we have to deal with uncertain information and which arises, because we process information in a layer within a small window al right.

(Refer Slide Time: 22:33)




So, with this background and perspective, which is absolutely fundamental, we proceed to understand hidden Markov model or HMM.


(Refer Slide Time: 22:43)

A Motivating Example


Colored Ball choosing



Urn 1
of Red = 30
of Green = 50
of Blue = 20




Urn 2
of Red = 10
of Green = 40
of Blue = 50



Urn 3
of Red = 60
of Green = 10
of Blue = 30

Probability of transition to another Urn after picking a ball:

	U ₁	U ₂	U ₃
U ₁	0.1	0.4	0.5
U ₂	0.6	0.2	0.2
U ₃	0.3	0.4	0.3



We take a motivating example, which is very well known in most of the tutorials on hidden Markov model textbooks discussing Markov model 1 finds example of this kind or some variation of this example. The example, I believe first appeared in celebrated tutorial paper on hidden Markov model. So, here we have this 3 urns, urn 1 urn 2 and urn 3 and we are given the number of red balls, green balls and blue balls in these urns, in urn 1, we have 30 red balls, 50 green balls and 20 blue balls. In urn 2, we have 10 red balls, 40 green balls and 50 blue balls, In urn 3, we have 60 red balls, 10 green balls and 30 blue balls.

(Refer Slide Time: 24:00)

Example (contd.)

Given :

	U ₁	U ₂	U ₃
U ₁	0.1	0.4	0.5
U ₂	0.6	0.2	0.2
U ₃	0.3	0.4	0.3


and

	R	G	B
U ₁	0.3	0.5	0.2
U ₂	0.1	0.4	0.5
U ₃	0.6	0.1	0.3

Observation : RRGGBRGR

State Sequence : ??

Not so Easily Computable.



And there is a person, who is picking out balls from, these urns and he gives us a sequence of the drawing of balls and their colors. So, we find that, the person has first drawn a red ball, then another red ball, then a green ball, then a green ball, then a blue ball, then a red ball, then a green ball and then finally, a red ball. So, this is the way the balls have been drawn.

So, now the question that, we are asking what is the state sequence or in other words, what is the sequence of urns, from which he has drawn the balls alright. So, for this sequence of colors of balls, we have to produce, urn sequences $U_1 U_2 U_3$ and so on. Some U_i should be assigned to, each of these alphabets, denoting the color of the ball, but we are not completely in dark though, we do not everything, we still know that, the following probability exists. If a person is drawing a ball from urn 1, next time he can draw a ball from any of the 3 urns and they this probability. So, from urn U_1 , the probability of drawing, another drawing the next ball from urn 1, again is 0.1 and he can draw from urn 2 with probability 0.4 and the probability of drawing from urn 3 is 0.5 ok.

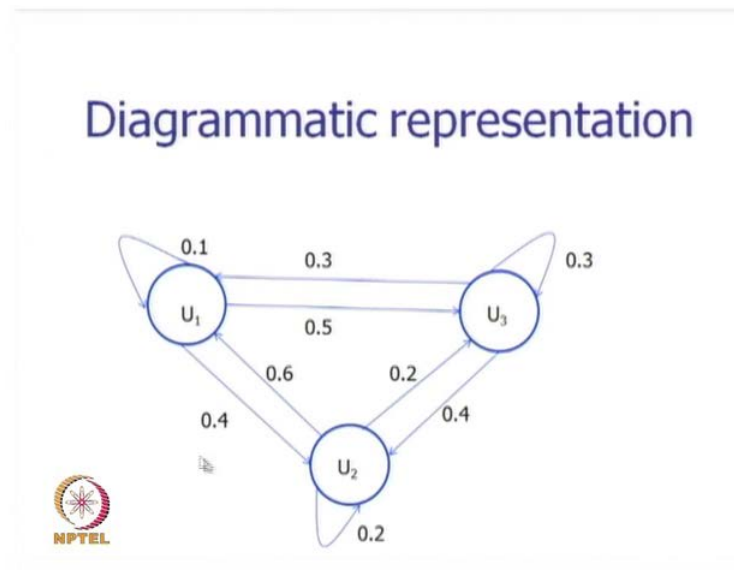
So, the so this is the meaning of probability, in each cell the column is the next state or the next urn and the row is the current state or urn. So, if the current urn, the meaning of this cell for example, is that if the current urn is U_2 , then the probability of the next draw being from urn 2 again is 0.2 and so on and so forth. And we since, we know the number of balls and their colors and the distribution of balls with their colors in the urns, we know the probability of drawing these colors also.

We had 30 red balls and 50 green balls and 20 blue balls in urn 1, which gives rise to the probability of drawing a red ball from urn 1 as 0.3 probability of drawing a green ball from urn 1 as 0.5 probability of drawing a blue ball from urn 1 as 0.2. So, this the way, it is recorded in urn 3, the probability of drawing red ball is 0.6 probability of drawing a green ball is 0.1, probability of drawing a blue ball is 0.3.

So, this 2 things are known, now the first table is technically known as, the transition probability table, if we look upon, the colors of the balls drawn as observation, then the sequence of urns from, which they came is called the state sequence. So, this is the transition probability table where, we show the probability of going from one state to another or one urn to another and this is called the output probability table. So, transition probability table and observation probability table ok.

So, using this 2, what is the most probable urn sequence, that the person has visited can be computed, this is the question. This however, is not easily computable and the this is the observations, the states are not shown to us and the states are so to say hidden from us, we have to guess the states, albeit very systematically as in the probability from this observations.

(Refer Slide Time: 28:15)



So, this is the diagram, which shows the situation under discussion, so urn 1 from urn 1, we can go to urn 2 with a probability of 0.4, we remain in urn 1 with a probability 0.1, we can go to urn 3 with a probability 0.5. So, this is the way the probability is organized of the urns, their transitions and the way, they remain in the same state or go to a different state, out of this state's again, we will have red blue and green coming out with different probabilities.

(Refer Slide Time: 28:58)

Example (contd.)


- Here :
 - $S = \{U_1, U_2, U_3\}$
 - $V = \{R, G, B\}$
- For observation:
 - $O = \{o_1 \dots o_n\}$
- And State sequence
 - $Q = \{q_1 \dots q_n\}$
- π is $P(q_1 = U_i)$

A =

	U ₁	U ₂	U ₃
U ₁	0.1	0.4	0.5
U ₂	0.6	0.2	0.2
U ₃	0.3	0.4	0.3

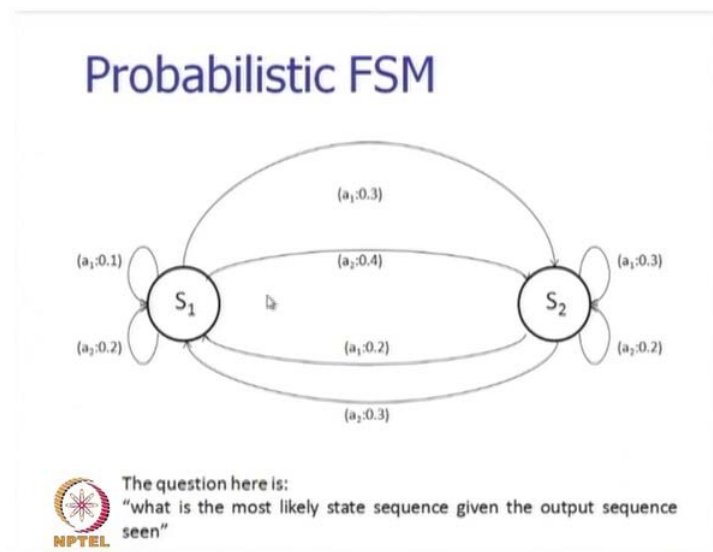
B =

	R	G	B
U ₁	0.3	0.5	0.2
U ₂	0.1	0.4	0.5
U ₃	0.6	0.1	0.3



Introducing a bit of notation here, we have the set of states as U_1, U_2 and U_3 and V is the set of outputs, which is R, G and B , the observation sequence is 1 to up to O_n and the state sequence is q_1 up to q_n . And there is a notion of initial probability π_i which is probability of q_1 , what is the probability, that the system starts in the initial state U_1 , typically this is taken as 1 and this is the transition probability table, the observation probability table.

(Refer Slide Time: 29:42)

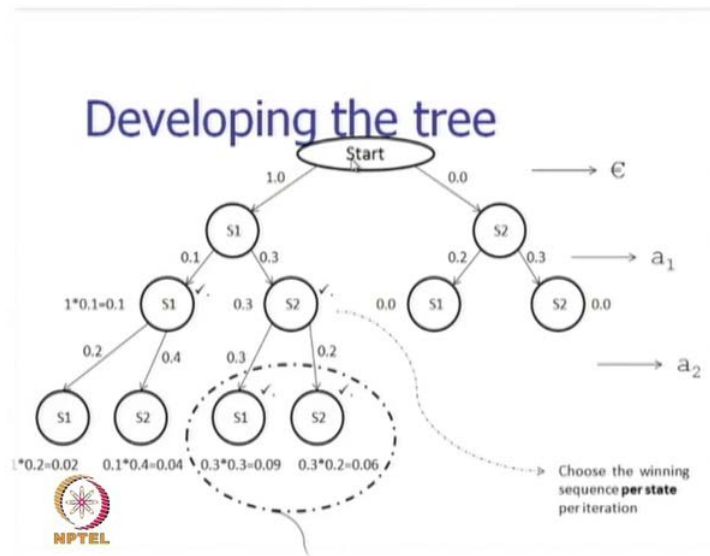


Now, we discuss the question of how, we can obtain the state sequence from an observation sequence, for this we have simplified the problem a bit, we are taking another automaton, an automaton with probabilities assigned on them. And here, we show an automaton with 2 states, S_1 and S_2 the outputs from S_1 and S_2 are the symbols a_1 and a_2 . So, we can look upon this as a 2 urn problem S_1 and S_2 are 2 urn and suppose, there are only 2 kinds of balls, red and green and a_1 a_2 correspond to red and green balls.

So, S_1 can remain in state S_1 , with an output of a_1 with the probability 0.1, S_1 can remain in state S_1 with an output of a_2 with a probability 0.2. S_1 can go to S_2 outputting a_1 with a probability of 0.3, S_1 can go to S_2 with output a_2 with a probability of 0.4, S_2 can remain in S_2 with output a_1 with a probability of 0.32 can remain in S_2 with output of a_2 with a probability of 0.2, S_2 can go to S_1 with a with an output of a_1 with a probability 0.2 and S_2 can go to S_1 with output a_2 with a probability of 0.3.

So, given this automaton where, the probabilities of the outputs and the probabilities of transition, which are sort of merged together, for such a finite state machine how can, we compute the state sequence probability given the observation sequence. So, this say simplified problem and you would have possibly noted also that, this is a slightly different problem, than the problem, we have discussed with urn where, the transition probability and the observation probability were separated, in 2 different tables. Here, we are talking about a transition into a state and the output being produced together with their probability. So, in this situation also how do, we compute the best possible state sequence, in the sense of having the highest probability.

(Refer Slide Time: 32:37)



So, it is done in the following way, suppose we are at the start state and assume that, the start state is always S 1, so this transition here 1.0 shows that, at the starting, the state is S 1 with probability 1. The other state does not exist, now S 2 and this is with epsilon transition that means, no input has been consumed, so now, we have this S 1, so effectively this part of the 3 is of interest.

Now, we get S 1 and this is the state, now we have got the symbol a 1, we would be given a complete sequence a 1 a 2 a 2 a 1 and so on and we have to the best possible state sequence. So, here from S 1 on a 1, we can go to S 1 or we can go to S 2, that is remain in S 1 or we can go to S 2, now the probability of going from S 1 to S 1 with output a 1, that probability, we see from the diagram is 0.1, so that is recorded here ok.

So, the probability is multiplied here, this is the probability of the sequence S 1 S 1, which is 1.0 into 0.1*0.1 and here, it is 0.3, so the probability of this sequence S 1 S 2 is 1 into 0.3, which is 0.3. Now, we take the next symbol, which is a 2 and when, we have seen a 2, S 1 on S 2 can go to S 2 with probability 0.4 or it can continue to remain in the state S 1 itself and that probability is 0.2.

Now, we if take the product of probabilities, why are we taking the product of probabilities, we will see that very soon, then the probability of this sequence S 1 S 1 S 1, which is recorded here is 0.1 into 0.2, which is 0.02. The sequence S 1 S 1 S 2, the

sequence probability will be 0.1 into 0.4, which is 0.04 from S 2, I can go to S 1 with probability 0.3 and this is giving rise to the probability here as 0.3 into 0.3, which is 0.09.

So, the sequence of S 1 S 2, S 1 is 0.09, similarly the probability of the sequence S 1 S 2 S 2 is 0.3 into 0.2, which is 0.06 alright. So, we have computed the probabilities at the leaf levels, you must understand that, the leaf probabilities are the probabilities of the accumulated sequences. Now, here comes the issue of Markov assumption, the Markov assumption kicks in at this stage and we immensely save on computation. The computation becomes much, much more efficient, then it would have been otherwise means, without the assumption of a Markov process, what is happening here is that, we find that having consumed epsilon a 1 a 2 or having produced. This sequence a 1 a 2, the machine can be in state S 1 S 2 or S 1 S 2 as shown here in this branch of the tree or in this branch of the tree. So, the question is out of this 4 leaves, which leaves should, we keep in under consideration, which are the leaves, which would remain alive for further processing.

So, the decision is made as follows, look at all those leaves, which are same states, so here S 1 is a leaf here, similarly S 1 is leaf here of course, the back of sequence is different here, the sequence S 1 S 2 S 1 here it is S 1 S 1 S 1. So, this 2 leaves are compared their, because they end in the same state and out of this, whichever has higher probability is retained. That means, on the next symbol, this is the state, which will be advanced just like predecessor, we will have 2 odds coming out, we will have 2 odds coming out with S 1 and S 2. And similarly out of this 2 leaves, which are both ending in S 2, we will keep this S 2 alive and not the other one, this particular state will survive, this S 2 will survive, there are ticks here similarly S 1 will survive here.

So, this is because the probability here is 0.09, which is much more than 0.02 and this is probability 0.06 much more than 0.04 and therefore, these are the 2 states, which will survive. Now, these 2 states are completely discarded from consideration, they will not be developed further, we will not give new states to them and transitions, we will advance these. So, this is the computational will produce and whichever leaf has the highest probability will give us the most probable state sequence. So, these are very simple idea, we keep on taking product of probabilities and keep on letting those leaf survive, which are the best amongst pairs, that means all of them are having the same leaf state. So, when we do, so we achieve an immense amount of efficiency in

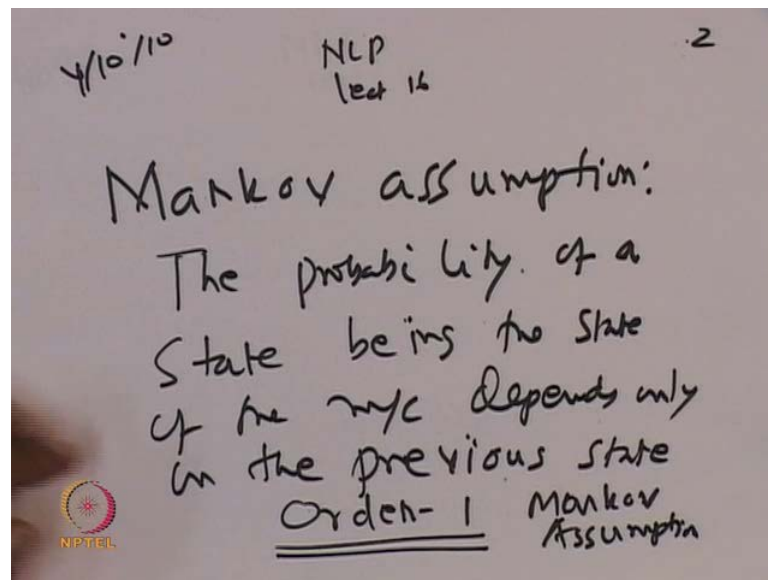
computation as we will see, now the question that arises is what is the rationale behind, letting this survive and not the other 2, what is the reason for that. The reason is that, we have made here an assumption, the assumption is that any state is the probability of reaching any state, depends only on the previous state of machine was in.

So, now, you see the machine was in S 2, before reaching this 2 states, the machine was in S 1 before reaching this 2 states, now if we advance all this states, the we are computing, namely taking the probabilities this effectively means, that we are giving importance to calculate the probability at a leaf only to the previous state. This probability here is the accumulated probability coming from S 2 and since S ones influence is absorbed in S 2, we need not worry about, what is happening in S 1, that has been accounted for.

So, later when, we advance the states from this 2 states, this 2 states will have no chance of winning, nothing no children coming from this will ever be winners and that is because we are always giving importance to the previous state. So, that this S 2 is winner compared to this S 2 and all its children will remain winner compared to all the children coming from S 2.

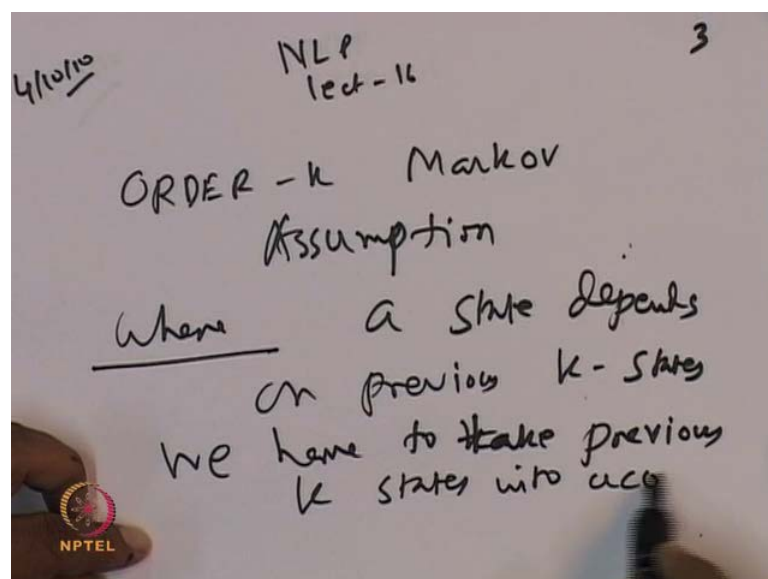
So, there is no point advancing S 2 anymore, similarly S ones children will continue to remain winner, compared to the children from this S 1 and there is no point advancing this S 1. So, this is due to a very important assumption called the Markov assumption, this is order 1 Markov assumption, we are saying that a state's behavior depends completely on the previous state it was in. So, let me write this down.

(Refer Slide Time: 41:13)



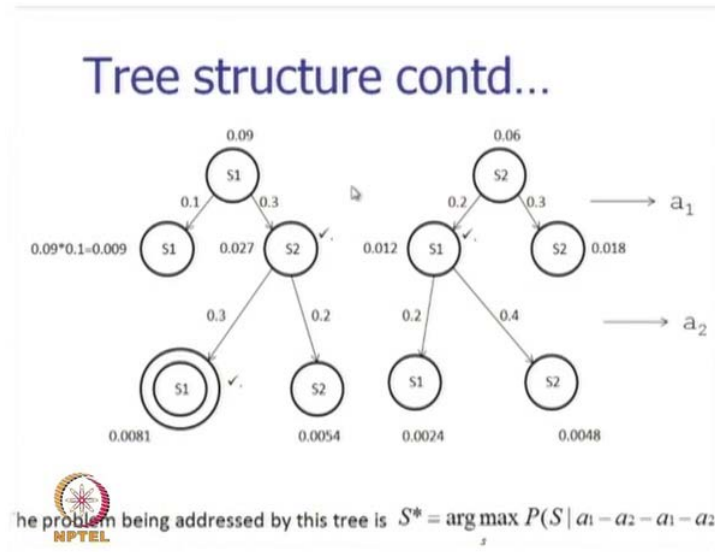
Markov assumption, the probability of a state being reached, the probability of a state being, the state of the machine depends only on the previous state. So, this is the Markov assumption, this is order 1 Markov assumption, this is order 1 alright.

(Refer Slide Time: 41:55)



So, similarly we can have order K Markov assumption where, a state depends on previous K states alright. So, in this case, we have to take previous K states into account alright.

(Refer Slide Time: 42:22)



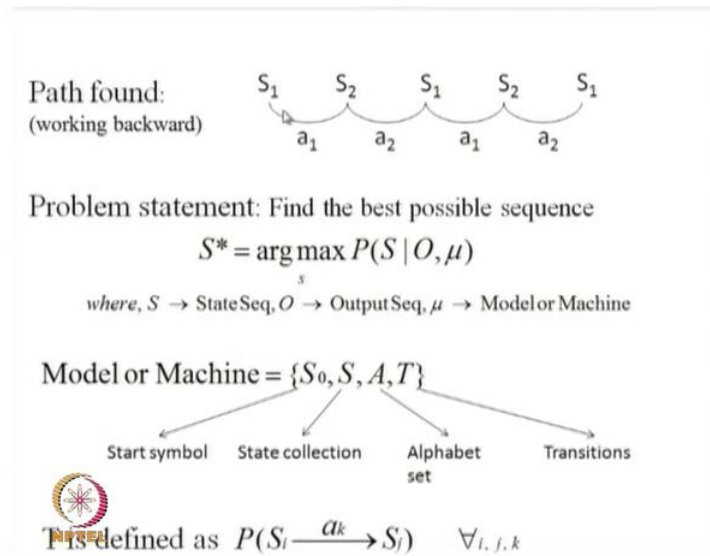
Proceeding further, the tree structure as you see is continued here, our sequence of observations were a 1 a 2 then again a 1 a 2. So, we got the surviving states as S 1 and S 2 from the right side of the tree and this next state is a 1 on a 1 output S 1 can go to the state S 1 with probability 0.1, that accumulated probability was 0.09 multiplied by 0.1, this gives me 0.009 and S 1 can go to S 2 with output a 1 probability 0.3.

So, the probability at S 2 become 0.027, similarly on this part of the tree from S 2 to S 1, we have the probability 0.2, which when multiplied with 0.06 give me 0.012, similarly here at S 2, we get the probability 0.06 into 0.3, which is 0.018. So, again by using the Markov assumption, we look at these leaves S 1 S 2, S 1 S 2 and we find that, this S 1 is bigger than, this S 1, so this survives, this S 2 is bigger than this S 2, so this survives and then after that, this 2 states are advanced. The next symbol, which is produced in a 2, now a 2 says that, S 2 can output a 2 and go to S 1 with probability 0.3. And again by multiplying the probabilities the probability at this leaf level is 0.0081, the probability at S 2 here is 0.0054, the probability here is 0.0024, the probability here is 0.0048.

So, a 1 a 2 a 1 a 2 being the output sequence, we know that the most probable or the best possible state sequence would be going backward S 1 S 2 S 1, S 1 S 2 S 1, then S 2 and then S 1, so the state sequence is S 1 S 2 S 1, S 2 S 2, then again S 1. So, this is the best possible most probable state sequence and the problem, which is being addressed by this tree is what is the best possible sequence S star where, we do an arg max based

computation of S with probability of we compute, the arg max of probability of S given a 1 a 2 a 1 a 2 as the output sequence and the machine mu.

(Refer Slide Time: 45:32)



So, this is the probability, we would like to calculate, now one can mathematical see how, why this works, after we have discussed this state sequence, the system is initially the machine is initially in S 1 on a 1, it go to S 2 then a 2, it goes to S 1 again on a 1, it goes to S 2 then a 2, it go to S 1. So, the states alternate and this is the best possible state sequence and our task as, we have stated is to find out the best possible state sequence S star given the observation sequence O and the machine mu, the machine mu is actually the set of probability values and this arg max is over all possible S S. So, a very simple mathematics will show us that, Markov assumption is indeed sound as, we write down now.

(Refer Slide Time: 46:29)

4/10/10 NLP - lect 16 4

$$S^* = \underset{S}{\operatorname{argmax}} (P(S | a_1 a_2 a_1 a_2, \mu))$$

$$P(s_{i_1}^{a_1} s_{i_2}^{a_2} s_{i_3}^{a_1} s_{i_4}^{a_2} s_{i_5} | a_1 a_2 a_1 a_2)$$

By chain rule

$$= P(s_{i_1}) \cdot P(s_{i_2} | s_{i_1}) \cdot P(s_{i_3} | s_{i_1} s_{i_2})$$

Markov assumption

$$= P(s_{i_1}) \cdot P(s_{i_2} | s_{i_1}) \cdot P(s_{i_3} | s_{i_2}) \dots$$

So, if I take this particular problem, we are handling, so S^* is $\operatorname{argmax}_S P(S)$ given $a_1 a_2 a_1 a_2$ and the machine μ . So, this will be equal to probability of let us say $s_{i_1} s_{i_2} s_{i_3} s_{i_4} s_{i_5}$ given $a_1 a_2 a_1 a_2$, so we will neglect μ for the moment, we will neglect new for the moment and by applying by chain rule, we have this is equal to $P(s_{i_1})$ into $P(s_{i_2} | s_{i_1})$ into $P(s_{i_3} | s_{i_1} s_{i_2})$ and so on. Then when we make Markov assumption, this will be equal to $P(s_{i_1})$ into $P(s_{i_2} | s_{i_1})$ into $P(s_{i_3} | s_{i_2})$ given $s_{i_1} s_{i_2}$ and so on, this we will continue in the next class.