Technische Hochschule Rosenheim

Modul - Introduction to AI (AI1)

Bachelor Programme AAI

09 - Probability and Bayes Theorem

Prof. Dr. Marcel Tilly

Faculty of Computer Science, Cloud Computing

Plan X-Mas break

- Mock exams on 23.12 in LC
- First exercise and mock exam discussion 14.01.2022 ab 13:45
- First lecture 11.01.2022
- Visitor in lecture 18.01.2022

Goals (formal)

- Students know about uncertainty and probabilities.
- Students can explain Bayes Theorem.
- Students understand Markov Models.
- Students know about hidden states in Markov Models.

Uncertainty

- Last lecture: How can AI represent and derive new knowledge?
- However, often, in reality, the AI has only partial knowledge of the world, leaving space for **uncertainty**.
- GOAL: We would like our AI to make the best possible decision in these situations.

> For example, when predicting weather, the AI has information about the weather today, but there is no way to predict with 100% accuracy the weather tomorrow.

We want an AI that makes optimal decisions given limited information and uncertainty.

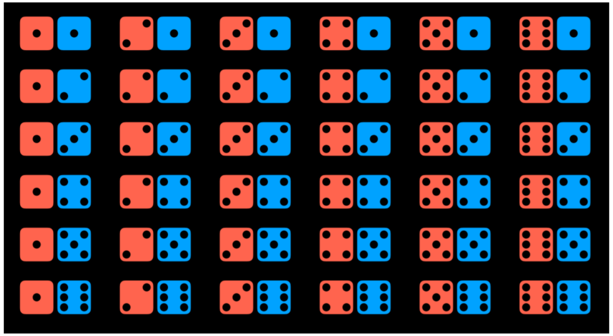
Probability

Uncertainty can be represented as a number of events and the likelihood, or probability, of each of them happening.

- **Possible Worlds**: Every possible situation can be thought of as a world, represented by the lowercase Greek letter omega ω.
  - For example, rolling a dice can result in six possible worlds (a world where the die yields a 1, a world where the die yields a 2, and so on
  - To represent the probability of a certain world, we write $P(\omega)$.
- **Axioms in Probability**
  - $0 < P(\omega) < 1$: every value representing probability must range between 0 and 1.
  - 0 is an impossible event and 1 is an event that is certain to happen
  - The higher the value, the more likely the event is to happen.
  - The probabilities of every possible event, when summed together, are equal to 1: $\sum_w P(w) = 1$

Summing Probabilities

- The probability of rolling a number R with a standard die can be represented as P(R)
- P(R) = 1/6
  - because there are six possible worlds (rolling any number from 1 through 6)
  - each is equally likely to happen
- Now, consider the event of rolling two dice -> there are 36 possible events, which are, again, equally as likely.



Question?

What happens if we try to predict the sum of the two dice? What is P(6,6)?

Sum of Two Dice

- To get the probability of an event, we divide the number of worlds in which it occurs by the number of total possible worlds.
  - For example, there are 36 possible worlds when rolling two dice.
  - Only in one of these worlds, when both dice yield a 6, do we get the sum of 12.
  - Thus, P(12) = 1/36, or, in words, the probability of rolling two dice and getting two numbers whose sum is 12 is 1/36.

Question?

What is P(7)?

Question?

Technische Hochschule Rosenheim

What is P(7)?

We count and see that the sum 7 occurs in 6 worlds. Thus, P(7) = 6/36 = 1/6.

Unconditional Probability

Technische Hochschule Rosenheim

Unconditional probability is the degree of belief in a proposition in the absence of any other evidence.

- All the questions that we have asked so far were questions of unconditional probability, because the result of rolling a die is not dependent on previous events.

Conditional Probability

- Conditional probability is the degree of belief in a proposition given some evidence that has already been revealed.
- AI can use partial information to make educated guesses about the future. -To use this information, which affects the probability that the event occurs in the future, we rely on conditional probability.

Conditional Probability

Conditional probability is expressed using the following notation: P(a | b)

- meaning

  the probability of event a occurring given that we know event b to have occurred

  or

  the probability of a given b

Now we can ask questions like

- What is the probability of rain today given that it rained yesterday:

  P(rain today | rain yesterday)?

- What is the probability of the patient having the disease given their test results

  P(disease | test results)?

Conditional Probability

Mathematically, to compute the conditional probability of a given b, we use the following formula:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

The probability that *a given b* is true, is equal to the probability of a and b being true, divided by the probability of b.

> An intuitive way of reasoning about this is the thought "we are interested in the events where both a and b are true (the numerator), but only from the worlds where we know b to be true (the denominator)."

Equivalent Formulas

$$P(a \wedge b) = P(b)P(a|b)$$

$$P(a \wedge b) = P(a)P(b|a)$$

For example:

- Consider P(sum 12 | roll six on one dice) - *or the probability of rolling two dice and getting a sum of twelve, given that we have already rolled one die and got a six.*
  - To calculate this, we first restrict our worlds to the ones where the value of the first dice is six: P(6) = 1/6
  - Now we ask how many times does the event a (the sum being 12) occur in the worlds that we restricted the question to (dividing by P(b), or the probability of the first die yielding 6): P(sum12) = 1/36 and P(sum12|6)= 1/6

Random Variables

A random variable is a variable in probability theory with a domain of possible values that it can take on.

- For example, to represent possible outcomes when rolling a die, we can define a random variable *Roll*, that can take on the values {0, 1, 2, 3, 4, 5, 6}.
- To represent the status of a flight, we can define a variable *Flight* that takes on the values {on time, delayed, canceled}.
- Often, we are interested in the probability with which each value occurs. We represent this using a probability distribution.
- For example,

```
P(Flight = on time) = 0.6
P(Flight = delayed) = 0.3
P(Flight = canceled) = 0.1
```

Random Variables

To interpret the probability distribution with words, this means that there is a 60% chance that the flight is on time, 30% chance that it is delayed, and 10% chance that it is canceled.

> Note that, as shown previously, the sum the probabilities of all possible outcomes is 1.

A probability distribution can be represented more succinctly as a vector.

For example:

```
P(Flight) = <0.6, 0.3, 0.1>
```

For this notation to be interpretable, the values have a set order (in our case, on time, delayed, canceled).

Independence

Independence is the knowledge that the occurrence of one event does not affect the probability of the other event.

- For example, when rolling two dice, the result of each die is independent from the other.
- This is opposed to dependent events, like clouds in the morning and rain in the afternoon. If it is cloudy in the morning, it is more likely that it will rain in the morning, so these events are dependent.

Independence can be defined mathematically:

Events a and b are independent if and only if the probability of a and b is equal to the probability of a times the probability of b:

$P(a \wedge b) = P(a)P(b)$.

Bayes' Rule

- Bayes' rule is commonly used in probability theory to compute conditional probability.
- In words, Bayes' rule says that the probability of b given a is equal to the probability of a given b, times the probability of b, divided by the probability of a.

$$P(a \wedge b) = P(b)P(a|b)$$

$$P(a \wedge b) = P(a)P(b|a)$$

$$P(a)P(b|a) = P(b)P(a|b)$$

$$P(b|a) = \frac{P(b)P(a|b)}{P(a)}$$

Bayes' Rule

Example

- We would like to compute the probability of it raining in the afternoon if there are clouds in the morning, or P(rain | clouds).
- We start with the following information:
  - 80% of rainy afternoons start with cloudy mornings, or P(clouds | rain).
  - 40% of days have cloudy mornings, or P(clouds).
  - 10% of days have rainy afternoons, or P(rain).

Applying Bayes' rule, we compute (0.1)(0.8)/(0.4) = 0.2.

That is, the probability that it rains in the afternoon given that it was cloudy in the morning is 20%.

Bayes' Rule

- Knowing P(a | b), in addition to P(a) and P(b), allows us to calculate P(b | a).
- This is helpful, because knowing the conditional probability of a visible effect given an unknown cause, P(visible effect | unknown cause), allows us to calculate the probability of the unknown cause given the visible effect, P(unknown cause | visible effect).

For example

- we can learn P(medical test results | disease) through medical trials, where we test people with the disease and see how often the test picks up on that.
- Knowing this, we can calculate P(disease | medical test results), which is valuable diagnostic information.

Joint Probability

- Joint probability is the likelihood of multiple events all occurring.
- Let us consider the following example, concerning the probabilities of clouds in the morning and rain in the afternoon.

Cloud:

C = cloud	C = ¬cloud
0.4	0.6

Rain:

R = rain	R = ¬rain
0.1	0.9

Joint Probability

- Looking at these data, we can't say whether clouds in the morning are related to the likelihood of rain in the afternoon.
- To be able to do so, we need to look at the *joint probabilities* of all the possible outcomes of the two variables.
- We can represent this in a table as follows:

	R = rain	**R = ¬rain**
C = cloud	0.08	0.32
C = ¬cloud	0.02	0.58

- Now we are able to know information about the co-occurrence of the events. For example, we know that the probability of a certain day having clouds in the morning and rain in the afternoon is 0.08. The probability of no clouds in the morning and no rain in the afternoon is 0.58.

Joint Probability

Using joint probabilities, we can deduce conditional probability.

For example:

- if we are interested in the probability distribution of clouds in the morning given rain in the afternoon:

  $P(C \mid rain) = P(C, rain)/P(rain)$ (a side note: in probability, commas and $\wedge$ are used interchangeably.
- Thus, $P(C, rain) = P(C \wedge rain)$). In words, we divide the joint probability of rain and clouds by the probability of rain.

Joint Probability

It is possible to view P(rain) as some constant by which P(C, rain) is multiplied.

- Thus, we can rewrite P(C, rain)/P(rain) = αP(C, rain), or α<0.08, 0.02>.
- Factoring out α leaves us with the proportions of the probabilities of the possible values of C given that there is rain in the afternoon.
- Namely, if there is rain in the afternoon, the proportion of the probabilities of clouds in the morning and no clouds in the morning is $0.08 : 0.02$.

> Note that 0.08 and 0.02 don't sum up to 1; however, since this is the probability distribution for the random variable C, we know that they should sum up to 1. Therefore, we need to normalize the values by computing α such that $\alpha 0.08 + \alpha 0.02 = 1$.

- Finally, we can say that P(C | rain) = <0.8, 0.2>.

Probability Rules

- **Negation: $P(\neg a) = 1 - P(a)$**: This stems from the fact that the sum of the probabilities of all the possible worlds is 1, and the complementary literals a and ¬a include all the possible worlds.
- **Inclusion-Exclusion**: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$. The worlds in which a or b are true are equal to all the worlds where a is true, plus the worlds where b is true. However, in this case, some worlds are counted twice (the worlds where both a and b are true)). To get rid of this overlap, we subtract once the worlds where both a and b are true (since they were counted twice).
- **Marginalization**: $P(a) = P(a, b) + P(a, \neg b)$. The idea here is that b and ¬b are disjoint probabilities. That is, the probability of b and ¬b occurring at the same time is 0. We also know b and ¬b sum up to 1. Thus, when a happens, b can either happen or not. When we take the probability of both a and b happening in addition to the probability of a and ¬b, we end up with simply the probability of a.

Marginalization

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$

- The left side of the equation means “The probability of random variable X having the value x_i.”
- For example, for the variable C we mentioned earlier, the two possible values are clouds in the morning and no clouds in the morning. The right part of the equation is the idea of marginalization.
- $P(X = x_i)$ is equal to the sum of all the joint probabilities of x_i and every single value of the random variable Y.
- For example:

  P(C = cloud) = P(C = cloud, R = rain) + P(C = cloud, R = ¬rain) = 0.08 + 0.32 = 0.4.

Conditioning

$$P(X = x_i) = \sum_j P(X = x_i | Y = y_j) P(Y = y_j)$$

In this formula, the random variable X takes the value x_i with probability that is equal to the sum of the probabilities of x_i given each value of the random variable Y multiplied by the probability of variable Y taking that value.

This makes sense if we remember that P(a | b) = P(a, b)/P(b). If we multiply this expression by P(b), we end up with P(a, b), and from here we do the same as we did with marginalization.

Technische Hochschule Rosenheim

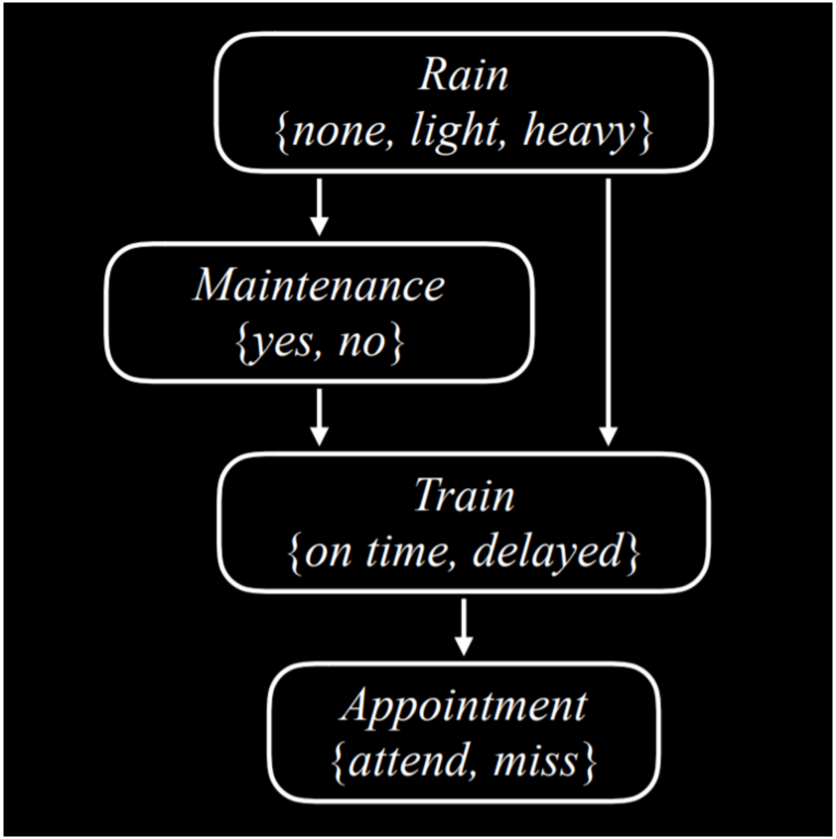
Bayesian *Networks*

Bayesian Networks

A Bayesian network is a data structure that represents the dependencies among random variables. Bayesian networks have the following properties:

- They are directed graphs.
- Each node on the graph represent a random variable.
- An arrow from X to Y represents that X is a parent of Y. That is, the probability distribution of Y depends on the value of X.
- Each node X has probability distribution P(X | Parents(X)).

Bayesian Network

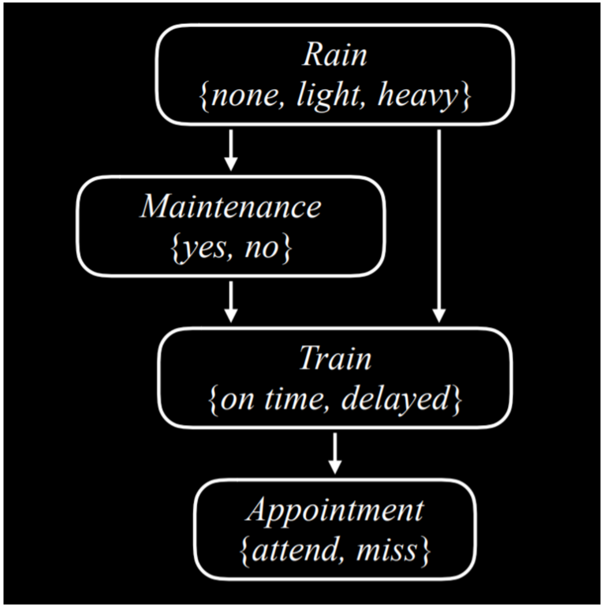


Bayesian Network

Let's describe this Bayesian network from the top down:

- **Rain** is the root node in this network. This means that its probability distribution is not reliant on any prior event. In our example, Rain is a random variable that can take the values {none, light, heavy} with the following probability distribution:

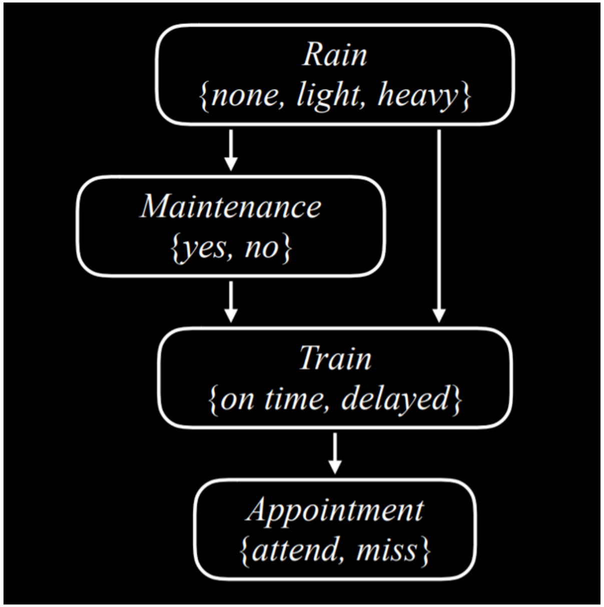
none	light	heavy
0.7	0.2	0.1



Bayesian Network

- **Maintenance**, in our example, encodes whether there is train track maintenance, taking the values {yes, no}. Rain is a parent node of Maintenance, which means that the probability distribution of Maintenance is affected by Rain.

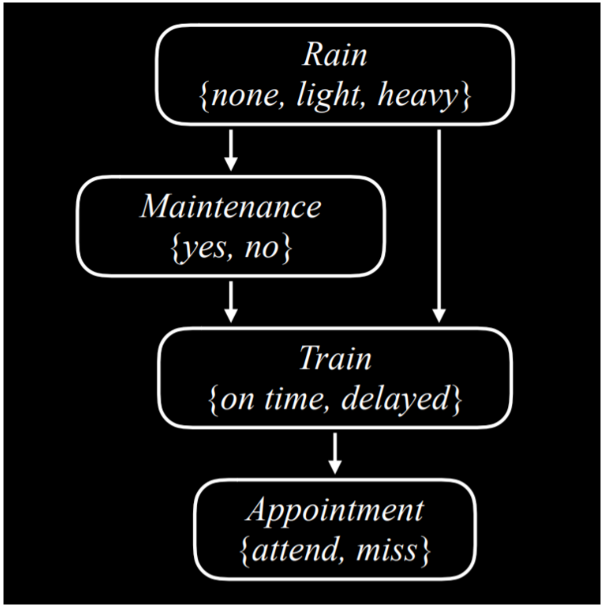
R	**yes**	**no**
none	0.4	0.6
light	0.2	0.8
heavy	0.1	0.9



Bayesian Network

- **Train** is the variable that encodes whether the train is on time or delayed, taking the values {on time, delayed}. Train has two parents, and their values affect the probability distribution of Train.

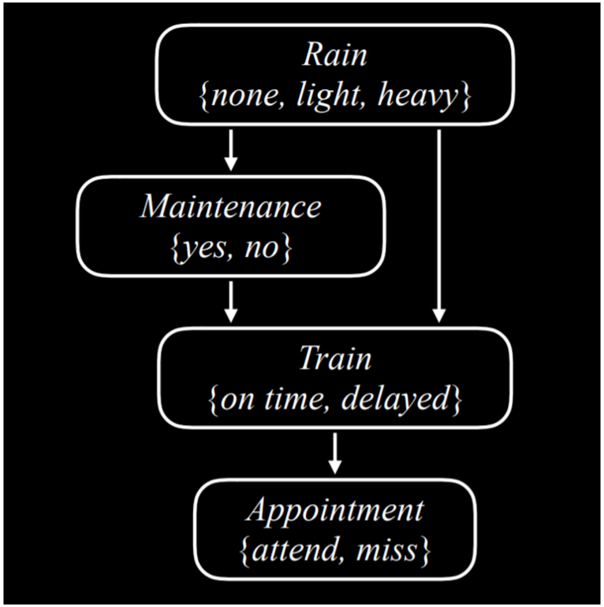
R	M	on time	delayed
none	yes	0.8	0.2
none	no	0.9	0.1
light	yes	0.6	0.4
light	no	0.7	0.3
heavy	yes	0.4	0.6
heavy	no	0.5	0.5



Bayesian Network

- **Appointment** is a random variable that represents whether we attend our appointment, taking the values {attend, miss}. It is true that maintenance affects whether the train is on time, and whether the train is on time affects whether we attend the appointment. However, in the end, what directly affects our chances of attending the appointment is whether the train came on time, and this is what is represented in the Bayesian network.

T	**attend**	**miss**
on time	0.9	0.1
delayed	0.6	0.4



Bayesian Network

Example

- If we want to find the probability of missing the meeting when the train was delayed on a day with no maintenance and light rain, or P(light, no, delayed, miss)
- We will compute the following:

```
P(light)P(no | light)P(delayed | light, no)P(miss | delayed).
```

- The value of each of the individual probabilities can be found in the probability distributions above, and then these values are multiplied to produce

```
P(no, light, delayed, miss)
```

Inference

- Inference is about entailment.
- This means that we could definitively conclude new information based on the information that we already had.
- We can also infer new information based on probabilities.
- While this does not allow us to know new information for certain, it allows us to figure out the probability distributions for some values. Inference has multiple properties.

Inference

- **Query X**: the variable for which we want to compute the probability distribution.
- **Evidence variables E**: one or more variables that have been observed for event e. For example, we might have observed that there is light rain, and this observation helps us compute the probability that the train is delayed.
- **Hidden variables Y**: variables that aren't the query and also haven't been observed. For example, standing at the train station, we can observe whether there is rain, but we can't know if there is maintenance on the track further down the road. Thus, Maintenance would be a hidden variable in this situation.
- **The goal**: calculate P(X | e). For example, compute the probability distribution of the Train variable (the query) based on the evidence e that we know there is light rain.

Inference Example

Technische Hochschule **Rosenheim**

- We want to compute the probability distribution of the Appointment variable given the evidence that there is light rain and no track maintenance.
- That is, we know that there is light rain and no track maintenance, and we want to figure out what are the probabilities that we attend the appointment and that we miss the appointment, P(Appointment | light, no).
- From the joint probability section, we know that we can express the possible values of the Appointment random variable as a proportion, rewriting

```
P(Appointment | light, no) as αP(Appointment, light, no)
```

- How can we calculate the probability distribution of Appointment if its parent is only the Train variable, and not Rain or Maintenance? Here, we will use marginalization. The value of

```
P(Appointment, light, no) = α[P(Appointment, light, no, delayed)
                             + P(Appointment, light, no, on time)]
```

Inference by *Enumeration*

Inference by enumeration is a process of finding the probability distribution of variable X given observed evidence e and some hidden variables Y.

Inference by Enumeration

Technische Hochschule Rosenheim

$$P(X|e) = \alpha P(X, e) = \alpha \sum_{y} P(X, e, y)$$

- In this equation, X stand for the query variable, e for the observed evidence, y for all the values of the hidden variables, and α normalizes the result such that we end up with probabilities that add up to 1.
- To explain the equation in words, it is saying that the probability distribution of X given e is equal to a normalized probability distribution of X and e. To get to this distribution, we sum the normalized probability of X, e, and y, where y takes each time a different value of the hidden variables Y.

Markov *Models*

Markov Models

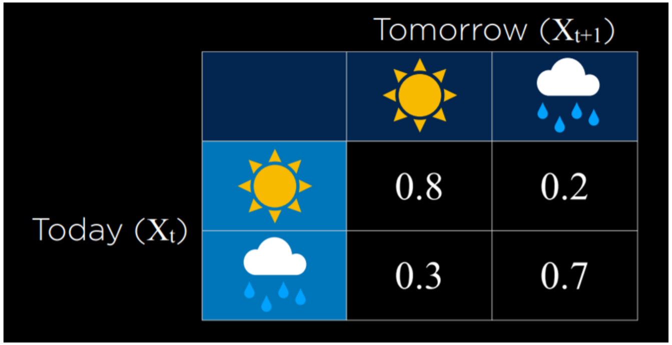
- So far, we have looked at questions of probability given some information that we observed.
- In this kind of paradigm, the dimension of time is not represented in any way.
- However, many tasks do rely on the dimension of time, such as prediction.
- To represent the variable of time we will create a new variable, X, and change it based on the event of interest, such that X_t is the current event, X_{t+1} is the next event, and so on. To be able to predict events in the future, we will use **Markov Models.**

The Markov Assumption

- The Markov assumption is an assumption that the current state depends on only a finite fixed number of previous states.
  - Think of the task of predicting weather. We could use all the data from the past year to predict tomorrow’s weather. However, it is infeasible, both because of the computational power this would require and because there is probably no information about the conditional probability of tomorrow’s weather based on the weather 365 days ago.
  - Using the Markov assumption, we restrict our previous states (e.g. how many previous days we are going to consider when predicting tomorrow’s weather), thereby making the task manageable. This means that we might get a more rough approximation of the probabilities of interest, but this is often good enough for our needs.

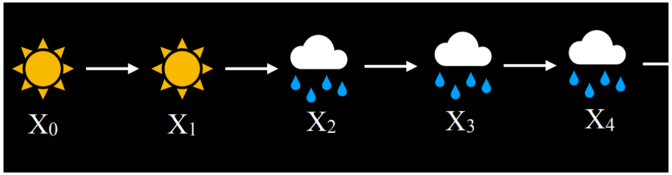
Markov Chain

- A Markov chain is a sequence of random variables where the distribution of each variable follows the Markov assumption.
- That is, each event in the chain occurs based on the probability of the event before it.
- To start constructing a Markov chain, we need a **transition model** that will specify the the probability distributions of the next event based on the possible values of the current event.



Transition Model

- In this example, the probability of tomorrow being sunny based on today being sunny is 0.8.
- This is reasonable, because it is more likely than not that a sunny day will follow a sunny day.
- However, if it is rainy today, the probability of rain tomorrow is 0.7, since rainy days are more likely to follow each other.
- Using this transition model, it is possible to sample a Markov chain.
- Start with a day being either rainy or sunny, and then sample the next day based on the probability of it being sunny or rainy given the weather today.



Markov Chain

Given this Markov chain, we can now answer questions such as “what is the probability of having four rainy days in a row?”

Here is an example of how a Markov chain can be implemented in code:

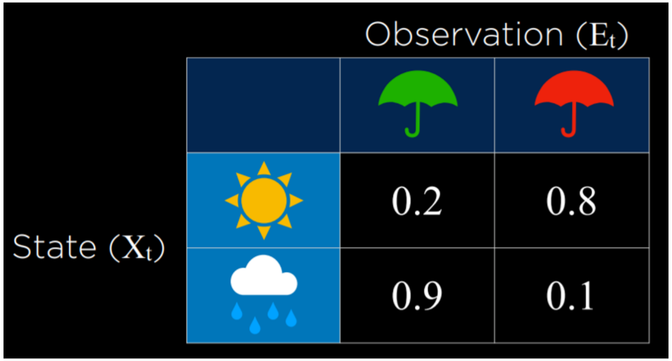
```
from pomegranate import *

# Define starting probabilities
start = DiscreteDistribution({
    "sun": 0.5,
    "rain": 0.5
})

# Define transition model
transitions = ConditionalProbabilityTable([
    ["sun", "sun", 0.8],
    ["sun", "rain", 0.2],
    ["rain", "sun", 0.3],
    ["rain", "rain", 0.7]
], [start])

# Create Markov chain
model = MarkovChain([start, transitions])

# Sample 50 states from chain
print(model.sample(50))
```

Hidden Markov Models

- A hidden Markov model is a type of a Markov model for a system with hidden states that generate some observed event.
- This means that sometimes, the AI has some measurement of the world but no access to the precise state of the world.
- In these cases, the state of the world is called the hidden state and whatever data the AI has access to are the observations. Here are a few examples for this:
  - For a robot exploring uncharted territory, the hidden state is its position, and the observation is the data recorded by the robot's sensors.
  - In speech recognition, the hidden state is the words that were spoken, and the observation is the audio waveforms.
  - When measuring user engagement on websites, the hidden state is how engaged the user is, and the observation is the website or app analytics.

Example

- Our AI wants to infer the weather (the hidden state), but it only has access to an indoor camera that records how many people brought umbrellas with them.
- Here is our sensor model (also called emission model) that represents these

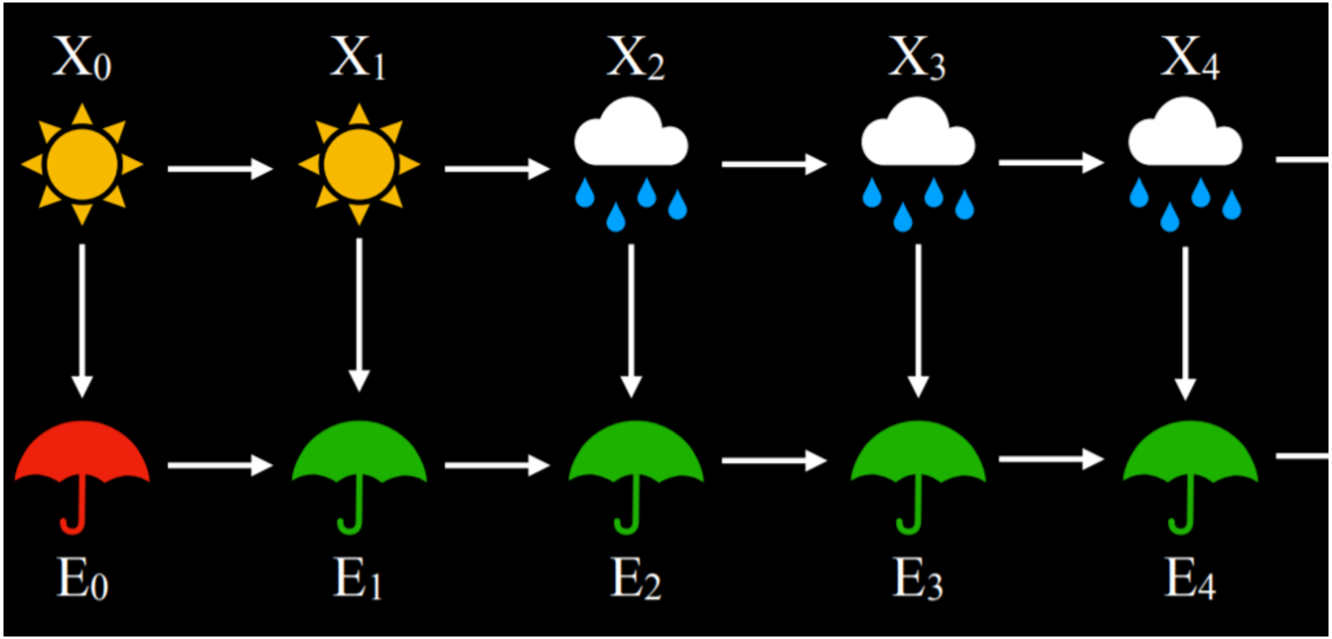


> In this model, if it is sunny, it is most probable that people will not bring umbrellas to the building. If it is rainy, then it is very likely that people bring umbrellas to the building. By using the observation of whether people brought an umbrella or not, we can predict with reasonable likelihood what the weather is outside.

Sensor Markov Assumption

Technische Hochschule **Rosenheim**

- A hidden Markov model can be represented in a Markov chain with two layers. The top layer, variable X, stands for the hidden state. The bottom layer, variable E, stands for the evidence, the observations that we have.



As code

```
from pomegranate import *

# Observation model for each state
sun = DiscreteDistribution({
    "umbrella": 0.2,
    "no umbrella": 0.8
})

rain = DiscreteDistribution({
    "umbrella": 0.9,
    "no umbrella": 0.1
})

states = [sun, rain]

# Transition model
transitions = numpy.array(
    [[0.8, 0.2], # Tomorroα's predictions if today = sun
     [0.3, 0.7]] # Tomorroα's predictions if today = rain
)

# Starting probabilities
starts = numpy.array([0.5, 0.5])

# Create the model
model = HiddenMarkovModel.from_matrix(
    transitions, states, starts,
    state_names=["sun", "rain"]
)
model.bake()
```

Comment

> Note that our model has both the sensor model and the transition model. We need both for the hidden Markov model.

In the following code snippet, we see a sequence of observations of whether people brought umbrellas to the building or not, and based on this sequence we will run the model, which will generate and print the most likely explanation (i.e. the weather sequence that most likely brought to this pattern of observations):

```
from model import model

# Observed data
observations = [
    "umbrella",
    "umbrella",
    "no umbrella",
    "umbrella",
    "umbrella",
    "umbrella",
    "umbrella",
    "no umbrella",
    "no umbrella"
]

# Predict underlying states
predictions = model.predict(observations)
for prediction in predictions:
    print(model.states[prediction].name)
```

Summary

- Uncertainty
- Probabilities
- Bayes Theorem
- Markov Model
- Hidden Markov Models