Probabilistic Methods (PM - 330725)

TOPIC 3: Probabilistic Models

Lecture 5

May 2025

Roadmap

Learning Outcomes

By the end of this topic, you will be able to:

- update probabilities and make predictions using Bayesian methods and prior distributions.
- model and analyze processes where future states depend on current states, using Markov chains and their properties.
- construct and apply Bayesian networks and Hidden Markov Models for complex system representation.
- use probabilistic models to handle and predict outcomes in sequential data and time series.

Introduction to Probabilistic Models

All the previous examples:

- Operate in environments where large amounts of data are available
- However, data don't cover all the possible scenarios $\Longrightarrow$ **UNCERTAINTY**
- Use a probabilistic model, typically learnt from data
- Use inference algorithms to carry out prediction and structure analysis

Probabilistic models offer:

- Principled quantification of uncertainty
- Natural way of dealing with missing data
- Interpretability

You have probably heard about two types of uncertainty:

- **Aleatoric**: Due to (pure) randomness, i.e. the variability in the outcome of an experiment due to random effects
- **Epistemic**: Due to lack of knowledge

Example

- Assume we want to predict Y from X
- We estimate a joint distribution $p(x, y)$ [EPISTEMIC][REDUCIBLE]
- We predict Y using $p(y \mid x) = p(x, y)/p(x)$
- We observe $X = x$, what is our model prediction for Y? [ALEATORIC][IRREDUCIBLE]

Introduction to Probabilistic Models

What we need from probabilistic models:

- Ability to operate in high dimensional spaces
- Support efficient inference and learning

Probabilistic graphical models offer:

- Structured specification of high dimensional distributions in terms of low dimensional factors
- Efficient inference and learning taking advantage of the structure
- Graphical representation interpretable by humans

Recap: Conditional probabilities and Bayes' theorem

If we have a probability space and two events A and B, the probability of A given B is called conditional probability, and it's defined as:

$$P(A \mid B) = \frac{P(A,B)}{P(B)}$$

As the joint probability is commutative, that is, $P(A, B) = P(B, A)$, it's possible to derive Bayes' theorem:

$$\begin{cases} P(A, B) = P(A \mid B) \cdot P(B) \\ P(B, A) = P(B \mid A) \cdot P(A) \end{cases} \Rightarrow P(A \mid B) = \frac{P(A,B)}{P(B)}$$

Remark: This theorem allows expressing a conditional probability as a function of the opposite one and the two marginal probabilities $P(A)$ and $P(B)$ and the general form of this theorem can be expressed as:

$$P(A \mid B) \propto P(B \mid A) \cdot P(A)$$

Therefore, we can summarize the relation as:

$$posteriorprobability \propto likelihood \cdot priorprobability$$

The equation of knowledge

Consider two propositions A, B.

A = "it will rain tomorrow", B = "the sky is cloudy"

A = "the Universe is flat", B = "observed CMB temperature map"

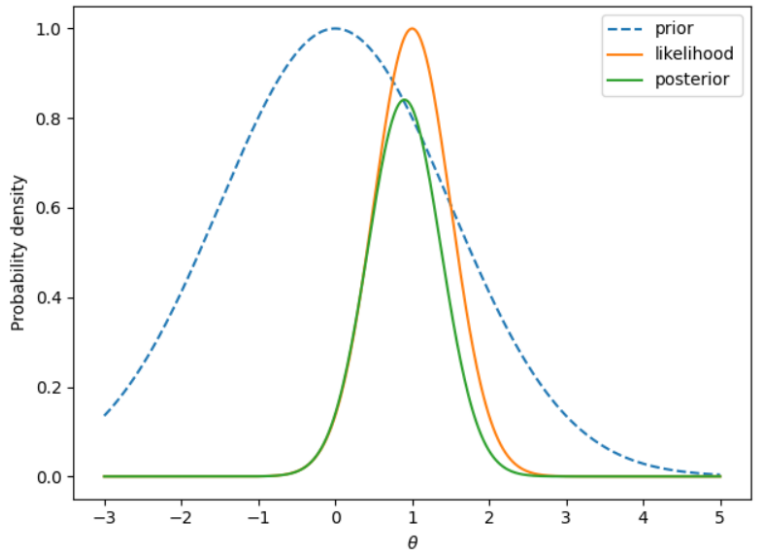
$$P(A \mid B)\, P(B) \;=\; P(A, B) \;=\; P(B \mid A)\, P(A)$$

Replace $A \longrightarrow \theta$ (the parameters of model M) and $B \longrightarrow d$ (the data):

$$P(\theta \mid d, M) \;=\; \frac{P(d \mid \theta, M)\; P(\theta \mid M)}{P(d \mid M)}.$$

$$P(\theta|d,M) = \frac{P(d|\theta,M)P(\theta|M)}{P(d|M)}$$

state of knowledge after ↓ posterior = information from the data ↓ likelihood x state of knowledge before ↓ prior / evidence



Why does Bayes matter?

This is what our scientific questions are about (the posterior)

This is what classical statistics is stuck with (the likelihood)

$$P(\text{hypothesis} \mid \text{data}) \neq P(\text{data} \mid \text{hypothesis})$$

Example: is a randomly selected person female? (Hypothesis)

Data: the person is pregnant ($d = \text{pregnant}$)

$$P(\text{female} \mid \text{pregnant}) = 1 \qquad P(\text{pregnant} \mid \text{female}) = 0.03$$

"Bayesians address the question everyone is interested in by using assumptions no–one believes,
while frequentists use impeccable logic to deal with an issue of no interest to anyone"

- Louis Lyons

Prior probability

Initial belief before observing the data

1. The proportion is not a limitation, because the term $P(B)$ is always a normalizing constant that can be omitted.
2. We must remember to normalize $P(A, B)$ so that its terms always sum up to one as we don't directly trust the prior probability, but we reweight it using the likelihood of our observations.
3. To achieve this goal, we need to introduce the prior probability, which represents the initial knowledge (before observing the data).

Non-informative vs Informative priors

Initial belief before observing the data

1. If the domain knowledge is consolidated, a precise prior distribution allows us to achieve a more accurate posterior distribution.

2. if the prior knowledge is limited, we avoid specific distributions, also called non-informative priors.

3. In general, distributions that concentrate the probability in a restricted region are very informative and their entropy is low because the uncertainty is capped by the variance..

Conjugate priors

1. important family of prior distributions are the conjugate priors with respect to a specific likelihood.
2. A distribution P is said to be conjugate prior to Q with respect to the likelihood L if, using the Bayes' formula, $Q \propto L \cdot P$, Q and P belong to same family.

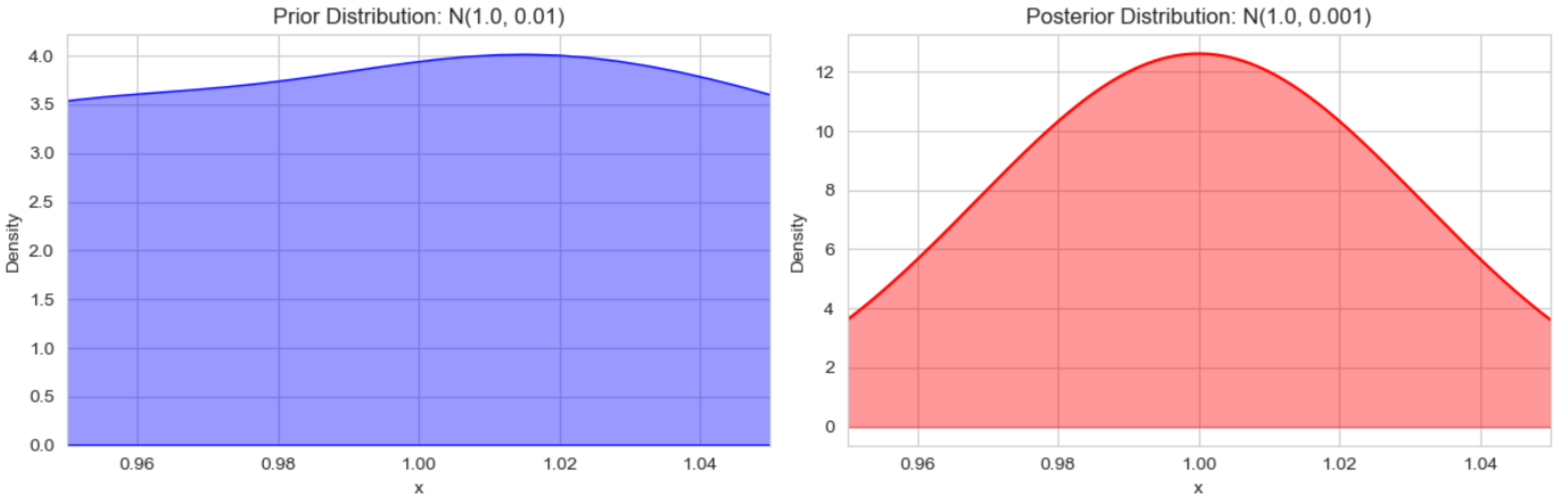
Conjugate priors are helpful for many reasons

1. First, they simplify the calculations, because, given a likelihood, it's possible to find the posterior without any integration
2. Second, in some domains, the posterior is naturally expected to belong to same family of the prior distribution.

Examples

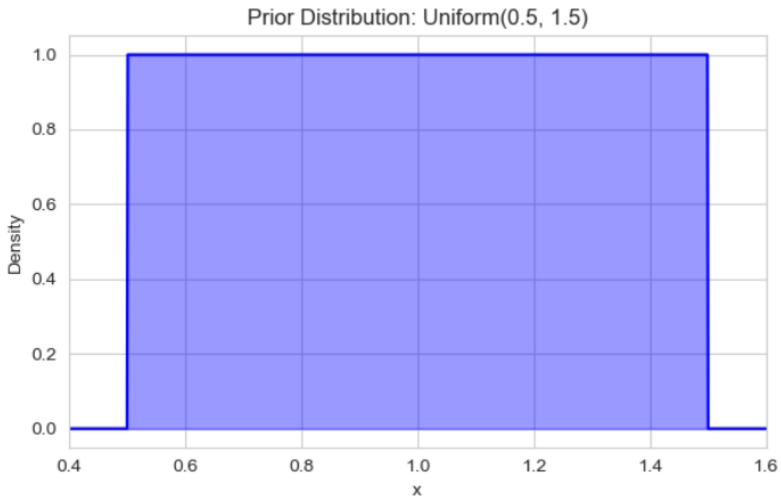
- For example, if the likelihood $L \approx N(\mu, \sigma)$ with known σ^2, the normal distribution is conjugate to itself, that is, the role of the likelihood is only to shift the Gaussian without altering the variance.

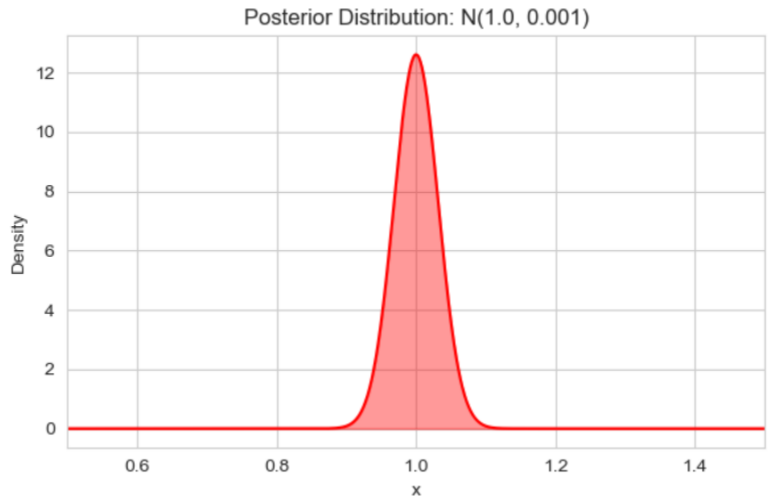
- Sampling of prior Gaussian distribution $N(\mu = 1.0, \sigma^2 = 0.01)$, the posterior to be very peaked around the mean. (1000 samples are drawn from this distribution with Standard deviation $\sigma = 0.0316$)



Examples

- Now, an example if we know that the posterior mean can be found in the range $(0.5, 1.5)$ but we are not sure about the true value, it's preferable to employ a distribution with a larger entropy, like a uniform one





Recap on conditional independence

conditional independence

consider two variables A and B, which are conditioned to a third one, C. We say that A and B are conditionally independent given C if:

$$P(A, B|C) = P(A|C) \cdot P(B|C)$$

IF an event A that is conditioned to a series of causes $C_1, C_2, \ldots, C_n$. The conditional probability is, therefore, $P(A|C_1, C_2, \ldots, C_n$. Applying Bayes' theorem, we get:

$$P(A|C_1, C_2, \ldots, C_n) \propto P(C_1, C_2, \ldots, C_n) \cdot P(A)$$

Now, ff there is conditional independence, the previous expression can be simplified as follows:

$$P(A|C_1, C_2, \ldots, C_n) \propto P(C_1|A) \cdot P(C_2|A) \ldots P(C_n|A) \cdot P(A) = \prod_{i=1}^{n} P(C_i|A) \cdot P(A)$$

Recap on joint probability

Chain rule of probabilities

suppose we have the joint probability $P(X_1, X_2, \ldots, X_n)$. It can be expressed as:

$$P(X_1, X_2, \ldots, X_n) = P(X_1|X_2, \ldots, X_n) \cdot P(X_2, \ldots, X_n) \ldots P(X_n)$$

We finally get:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_{i+1}, \ldots, X_n)$$

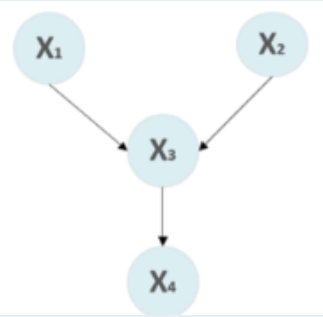
Lastly, we can express the full joint probability as the product of hierarchical conditional probabilities, until the last term, which is a marginal distribution

Introducing Bayesian Networks from Graphs

1. A graph consists of nodes and edges
2. Nodes: $X = X_1, X_2, \ldots, X_n)$
3. Undirected Edge: $X_i \longrightarrow X_j$
4. Directed Edge: $X_i \longrightarrow X_j$
5. Between a pair of nodes, at most one type of edge exists
   - We cannot have $X_i \longrightarrow X_j$ and $X_j \longrightarrow X_i$ at the same time, and
   - We cannot have $X_i \longrightarrow X_j$ and $X_j \longrightarrow X_i$ at the same time
6. Some edge: $X_i \rightleftarrows X_j$

Directed and undirected

- A graph is directed if its all edges are directed
- A graph is undirected if its all edges are undirected



Relationships

- $X_i \rightarrow X_j$
  - X_i is the parent and X_j is the child
- X_i and X_j are neighbors
- $X_i \rightarrow X_j$
  - X_i and X_j are adjacent

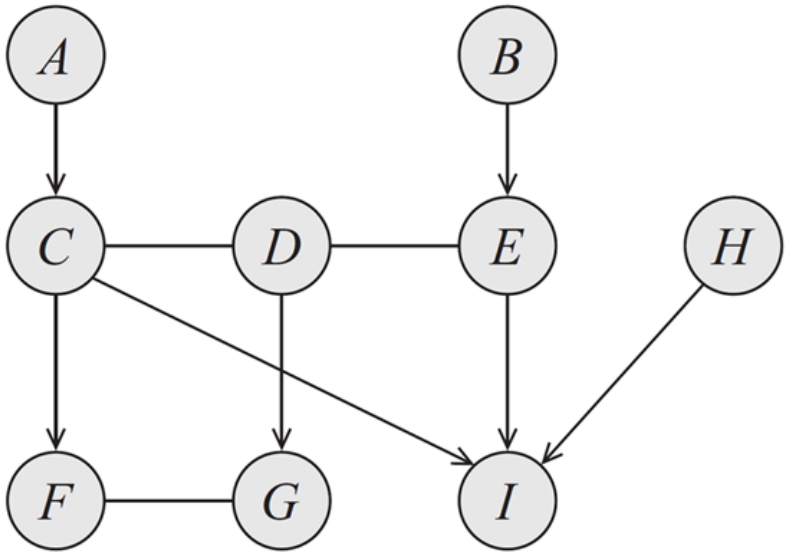
1. Degree of X_i: The number of edges Xi is part of
2. Indegree of X_i: The number of directed edges pointing to X_i
3. Degree of a graph: The maximal degree of a node in the graph

1. X_i is an ancestor of X_j if there is a directed path from X_i to X_j
2. X_i is a descendant of X_j if there is a directed path from X_j to X_i

Cycles and Loops

- A cycle is a directed path from a node to itself
- A graph is acyclic if it contains no cycles
- A directed acyclic graph is the one where all edges are directed and there are no cycles

- A loop is a trail from a node to itself
- A graph is singly-connected if it contains no loops



Bayesian Networks: Definition

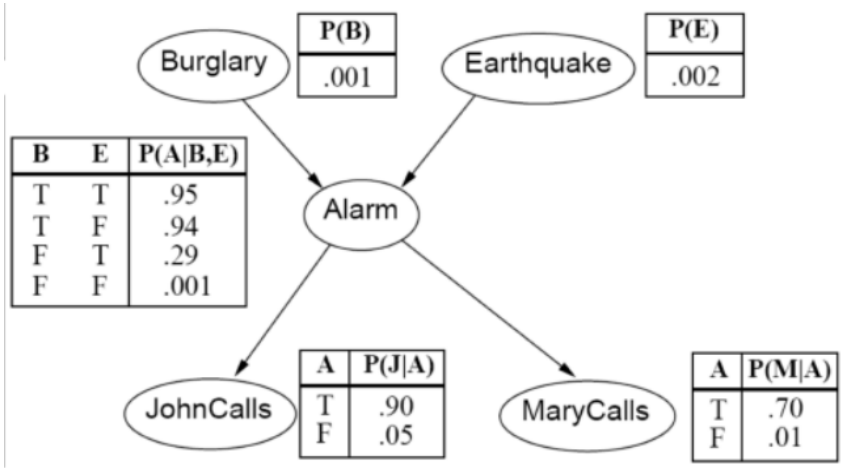
Definition: Bayesian Network

A Bayesian network over random variables $X_1, \ldots, X_n$ consists of

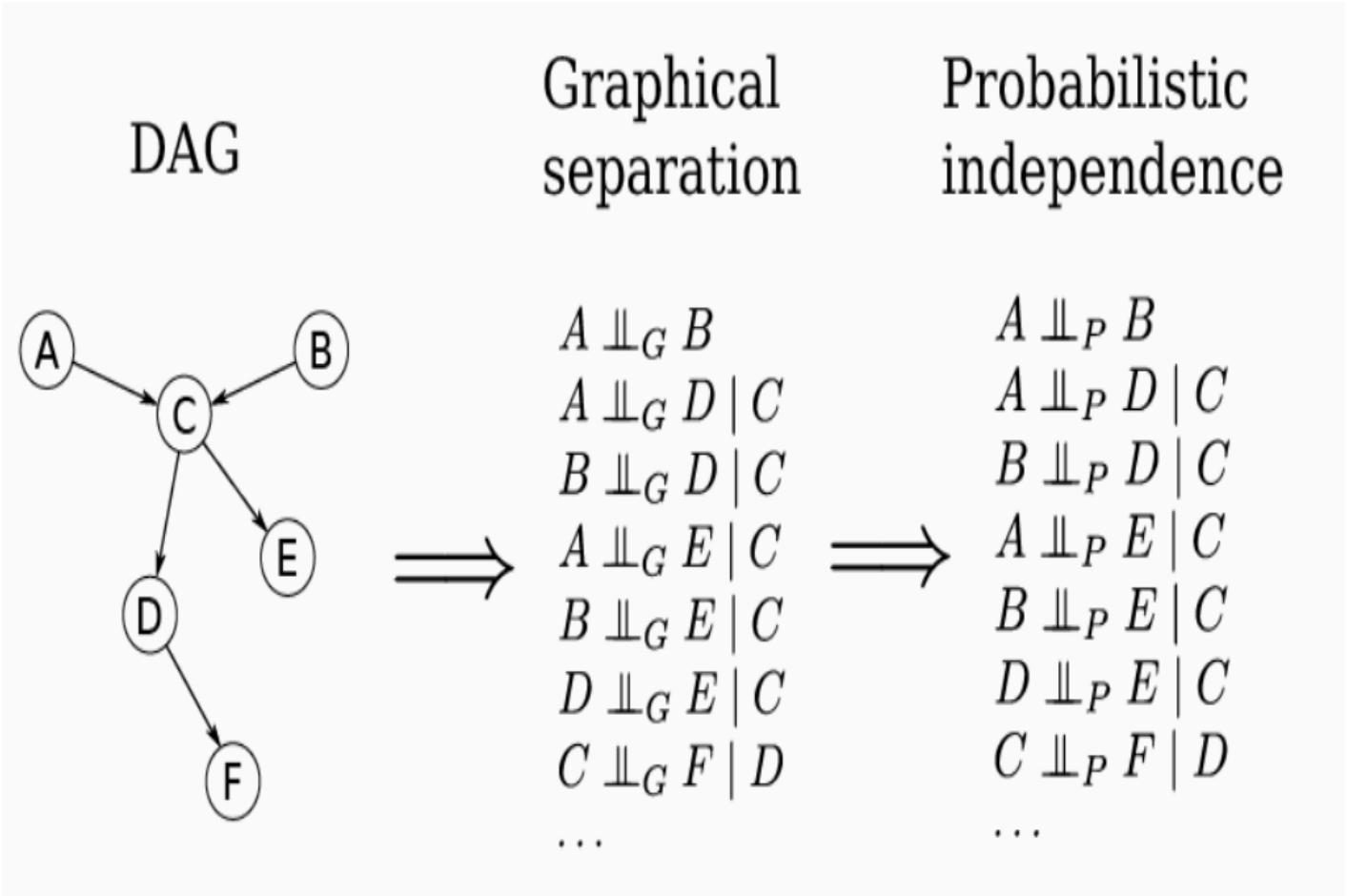
- A DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{X_1, \ldots, X_n\}$
- A set of local conditional distributions

$$\mathcal{P} = \left\{ p(X_i \mid \mathrm{pa}(X_i)) \,,\ X_i \in \mathcal{V} \right\},$$

where $\mathrm{pa}(X_i)$ denotes the parents of X_i according to $\mathcal{E}$.



How the DAG maps to the Probability Distribution

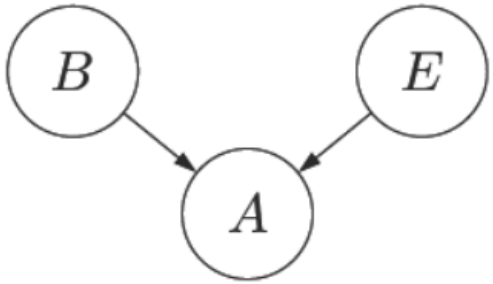


Formally, the DAG is an independence map of the probability distribution of $(\mathbf{X})$: graphical separation $(\perp\!\!\!\perp_G)$ implies probabilistic independence $(\perp\!\!\!\perp_P)$. **We'll see later!!!**

Explaining away

Key idea: explaining away

Suppose two causes positively influence an effect. Conditioned on the effect, conditioning on one cause reduces the probability of other cause.

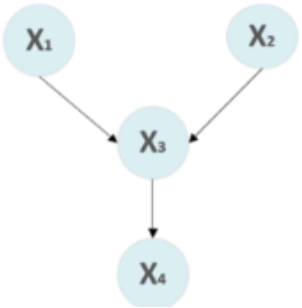


1. This last phenomenon has a special name: explaining away. Suppose we have two cause variables B and E, which are parents of an effect variable A. Assume the causes influence the effect positively (e.g., through the OR function).
2. Conditioned on the effect **A=1**, there is some posterior probability of B. Conditioned on the effect **A=1**and the other cause **E=1**, the new posterior probability is reduced. We then say that **the other cause E has explained away B**.

Bayesian Networks

- A Bayesian Network is a directed acyclic graph whose nodes are random variables and edges represent, intuitively, the direct influence of one node on another.
- Naive Bayes is a special Bayesian network Bayesian networks is
  - A data structure that provides the skeleton for representing a joint distribution compactly in a factorized way
  - A compact representation for a set of conditional independence assumptions about a distribution

Example of Bayesian network



1. The variable X_4 is dependent on X_3, which is dependent on X_1 and X_2.
2. for the network, we need the marginal probabilities $P(X_1)$ and $P(X_2)$
3. and the conditional probabilities $P(X_3|X_1, X_2)$ and $P(X_4|X_3)$

Using the chain rule, we can derive the full joint probability as:

$$P(X_1, X_2, X_3, X_4) = P(X_4|X_3) \cdot P(X_3|X_2, X_1) \cdot P(X_4) \cdot P(X_1)$$

Example of Bayesian network

- For example, if X_4 is caused indirectly by both X_1 and X_2, adding the edges $X_1 \longrightarrow X_4$ and $X_2 \longrightarrow X_4$ might seem good (potential caviat),
- we know that the final influence on X_4 is determined by the value of X_3 only, whose probability is conditional on X_1 and X_2.
- As such, we can say with confidence that $X_1 \longrightarrow X_4$ and $X_2 \longrightarrow X_4$ are spurious edges, and they don't need to be added.

Sampling from a Bayesian network

Some considerations and drawbacks

- Direct inference on a Bayesian network can be quite a difficult
- high number of variables and edges, because of the full joint probability, can become extremely complex!
- we need to compute the normalization constant to obtain the posterior probability

Solution!

How to determine the full joint probability sampling from a network? We'll use

1. a direct approach
2. MCMC algorithms.

Direct sampling Example

- The approach is to approximate the full joint probability through a sequence of samples drawn from each conditional distribution.

- From a mathematical viewpoint, we just create a frequency vector $F_{samples}(x_1, x_2, \ldots, x_N; N_{samples})$ and then approximating the full joint probability considering:

$$P(x_1, x_2, \ldots, x_N) = \lim_{N_{samples} \to \infty} F_{samples}(x_1, x_2, \ldots, x_N; N_{samples})$$

Direct sampling Example

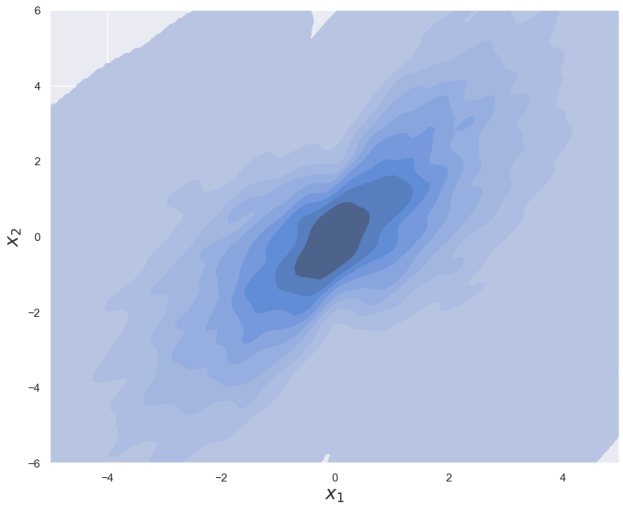
Algorithm 1 Sampling-based Estimation of P_{sampled}

Initialize the variable N_{samples}.
Initialize a matrix S of size (N, N_{samples}).
Initialize a dictionary or array F_{samples} for frequencies.
for $t = 1$ **to** N_{samples} **do**
    for $i = 1$ **to** N **do**
        Sample $x_i \sim P\big(X_i \mid \text{Predecessor}(X_i)\big)$.
        Store the sample: $S[i, t] \leftarrow x_i$.
    end for
    if F_{samples} contains key $S[:, t]$ **then**
        $F_{\text{samples}}[S[:, t]]$ += 1
    else
        $F_{\text{samples}}[S[:, t]] \leftarrow 1$
    end if
end for
Create a vector P_{sampled} of size $(N, 1)$.
for $i = 1$ **to** N **do**
    $P_{\text{samples}}[i, 0] \leftarrow \dfrac{F_{\text{sampled}}[i]}{N}$
end for

Example of direct sampling

Examples_Lecture5.ipynb

```
import numpy as np
def X1\_sample():
return np.random.normal(0.1, 2.0)
%def X2\_sample(x1):
%return np.random.normal(x1, 0.5 + np.
    sqrt(np.abs(x1)))
%
%Nsamples = 10000
%X = np.zeros((Nsamples, ))
%Y = np.zeros((Nsamples, ))
%for i, t in enumerate(range(Nsamples)
    ):
%x1 = X1\_sample()
%x2 = X2\_sample(x1)
%
%X[i] = x1
%Y[i] = x2
```



Recap on Discrete-time Markov chains

- State-transition systems augmented with probabilities
- States
  - set of states representing possible configurations of the system being modelled
- Transitions
  - transitions between states model evolution of system's state; occur in discrete time-steps
- Probabilities
  - probabilities of making transitions between states are given by discrete probability distributions

Markov property

If the current state is known, then the future states of the system are independent of its past states

1. i.e. the current state of the model contains all information that can influence the future evolution of the system
2. also known as "memorylessness"

Recap for Continuous-time Markov chains

- Continuous-time Markov chains (CTMCs)
  - labelled transition systems augmented with rates
  - discrete states
  - continuous time-steps
  - delays exponentially distributed
- Suited to modelling:
  - reliability models
  - control systems
  - queueing networks
  - biological pathways
  - chemical reactions

Idea:

We want strategies to draw a sequence of samples according to a precise transition probability from a sample to the following one.

1. Let's consider a time-dependent random variable $X(t)$,
2. and let's assume a discrete time sequence $X_1, X_2, \ldots, X_t, X_{t+1}, \ldots$ where X_t represents the value assumed at time t.

The following diagram, there's a schematic representation of this sequence:

Markov Chains

- Suppose we have N different states s_i for all $i = (1, N)$, in that case, it's possible to consider the probability $P(X_t = s_i | X_{t-1} = s_j, \ldots, X_1 = s_p)$.
- (X_t) is defined as a first-order Markov process if:

$$P(X_t = s_i | X_{t-1} = s_j, \ldots, X_1 = s_p) = P(X_t = s_i | X_{t-1}).$$

$$(X_t)$$

1. the probability that $X(t)$ is in a certain state depends only on the state assumed in the previous time instant. Therefore, we can define a transition probability for every couple (i, j):

$$P(j \rightarrow i) = P(X_t = s_i | X_{t-1} = s_j).$$

- Now, considering all the couples (i, j), it's also possible to build a transition probability matrix.
- the marginal probability that $X_t = s_i$ using a standard notation is defined as:

$$\pi_i(t) = P(X_t = s_i | X_{t-1} = s_j).$$

- using the Chapman-Kolmogorov equation

$$\pi_i(t+1) = \sum_k P(k \to i)\pi_k(t) \Rightarrow \bar{\pi}(t+1) = T^T \bar{\pi}(t).$$

- in order to compute , we need to sum over all possible previous states, considering the relative transition probability.
- rewritten in matrix form, using a vector containing all states and the transition probability matrix T^T (the uppercase superscript T means that the matrix is transposed)

$$\pi_i(t+1) = T^T \bar{\pi}(t) = T^T(T^T \bar{\pi}(t-1)) = \ldots = (T^T)^t \bar{\pi}(1).$$

Markov Chains

- it's really important to consider Markov chains that are able to reach a stationary distribution:

$$\bar{\pi_s} = T^T \bar{\pi_s}.$$

  Remark: state does not depend on the initial condition $\bar{\pi}(1)$.

- Stationarity: process of ergodicity for Markov
- the process has the same properties if averaged over time (which is often impossible) or averaged vertically (freezing the time) over the states following these conditions:
  1. aperiodicity for all states
  2. all states must be positive recurrent
- Finally, the existence of a unique stationary distribution, is that we are considering the sampling processes modeled as Markov chains
  1. It's possible to prove that a chain always reaches a stationary distribution if:

$$\text{For all } i, j \Rightarrow P(i \to j)\pi_{s_i} = P(j \to i)\pi_{s_j}.$$

Example

Now, consider the following Markov chain X_t, $t = 0, 1, 2, \ldots$ with the transition probability matrix:

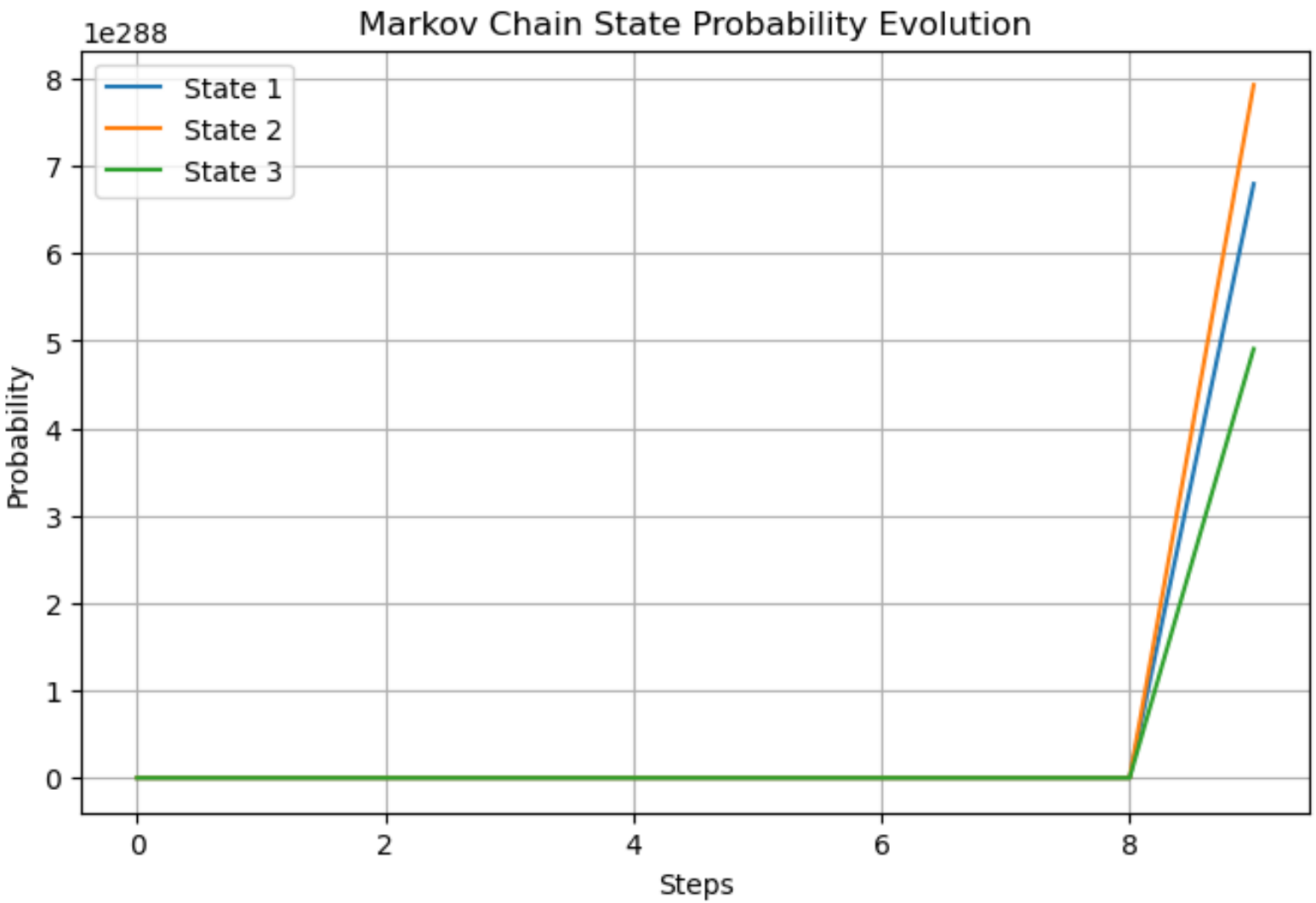
```
# Define the transition matrix for a 3-state Markov chain
P = np.array([
[0.2, 0.6, 0.2],
[0.5, 0.3, 0.2],
[0.3, 0.3, 0.4]
])
```

Key Concepts Covered:

- Transition Matrix: defines the probabilities of moving from one state to another.
- Ergodicity Check: ensures the chain is irreducible and aperiodic.
- Convergence to Steady-State: Simulates the Markov process over time.
- Plotting the Evolution: visualizing how state probabilities evolve.

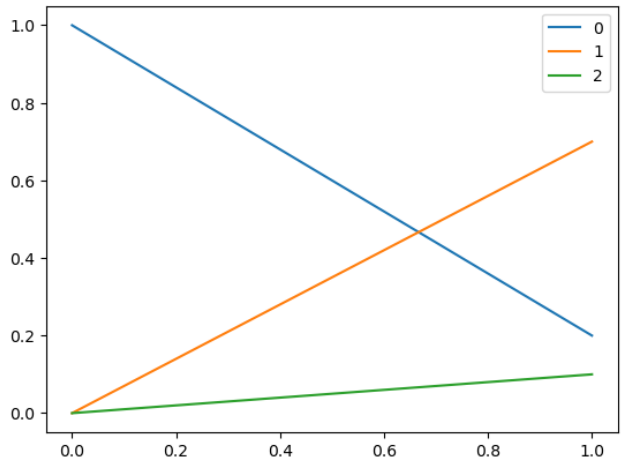
Follow the **Lecture 5 Markov chains basics.ipynb**

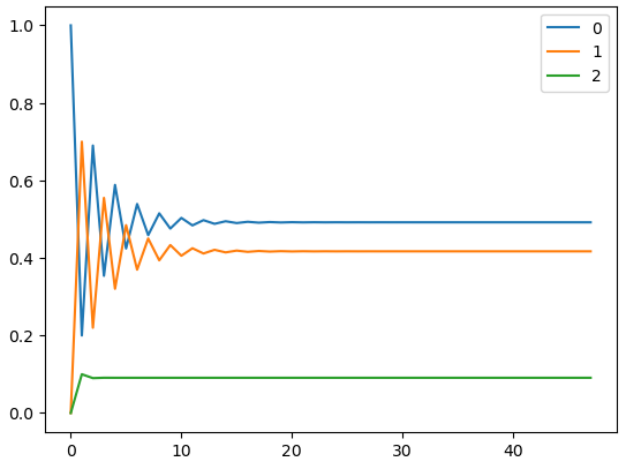
Example


Markov Chain State Probability Evolution
1e288
State 1
State 2
State 3
Probability
Steps
0
1
2
3
4
5
6
7
8
0
2
4
6
8

Example

```
P = np.array([[0.2, 0.7, 0.1],
[0.9, 0.0, 0.1],
[0.2, 0.8, 0.0]])
state=np.array([[1.0, 0.0, 0.0]])
stateHist=state
dfStateHist=pd.DataFrame(state)
distr\_hist = [[0,0,0]]
for x in range(50):
state=np.dot(state,P)
print(state)
stateHist=np.append(stateHist,state,axis=0)
dfDistrHist = pd.DataFrame(stateHist)
dfDistrHist.plot()
plt.show()
```





Wrapping up

Today we discussed:

1. Bayes' theorem and its applications
2. Bayesian networks
3. Sampling from a Bayesian network
4. Markov Chain main properties recap

Some References

- Pratt J., Raiffa H., Schlaifer R., Introduction to Statistical Decision Theory, The MIT Press, 2008
- Hoffmann M. D., Gelman A., The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, arXiv:1111.4246, 2011
- A. Gelman, J. B. Carlin, H. S. Stern, Bayesian Data Analysis, CRC Press, 2013
- Walsh B., Markov Chain Monte Carlo and Gibbs Sampling, Lecture Notes for EEB 596z, 2002
- R. A. Howard, Dynamic Programming and Markov Process, The MIT Press, 1960
- Rabiner L. R., A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77.2, 1989
- W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrik, 57/1, 04/1970
- Kevin B. Korb, Ann E. Nicholson, Bayesian Artificial Intelligence, CRC Press, 2010 Pearl J., Causality, Cambridge University Press, 2009
- L. E. Baum, T. Petrie, Statistical Inference for Probabilistic Functions of Finite State Markov Chains, The Annals of Mathematical Statistics, 37, 1966

Thank you very much!

ANY QUESTIONS OR COMMENTS?