Applied Probability and Statistics

Module-3, Lecture-1

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

Hidden Markov models

- Markov models
- Markov properties

Applied Probability and Statistics

Module-3, Lecture-2

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

Hidden Markov models

- Markov models
- Markov process
- Markov properties
- Markov Chains
- Hidden Markov models

Markov model

- A Markov model is a Stochastic method for randomly changing systems where it is assumed that future states do not depend on past states.
- These models show all possible states as well as the transitions, rate of transitions and probabilities between them.
- Markov models are often used to model the probabilities of different states and the rates of transitions among them.
- The method is generally used to model systems.
- Markov models can also be used to recognize patterns, make predictions and to learn the statistics of sequential data.

- There are four types of Markov models that are used situationally:
  - ✓Markov chain - used by systems that are autonomous and have fully observable states
  - ✓Hidden Markov model - used by systems that are autonomous where the state is partially observable.
  - ✓Markov decision processes - used by controlled systems with a fully observable state.
  - ✓Partially observable Markov decision processes - used by controlled systems where the state is partially observable.

Markov chain

- A Markov chain is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event
- A Markov chain is simplest type of Markov model, where all states are observable, and probabilities converge over time.
- Markov chains is that, as you move along the chain, the state where you are at any given time matters.
- The transitions between states are conditioned, or dependent, on the state you are in before the transition occurs.

Markov process

- A Markov process is a random process in which the future is independent of the past, given the present.
- Thus, Markov processes are the natural stochastic analogs of the deterministic processes described by differential and difference equations
- They form one of the most important classes of random processes.

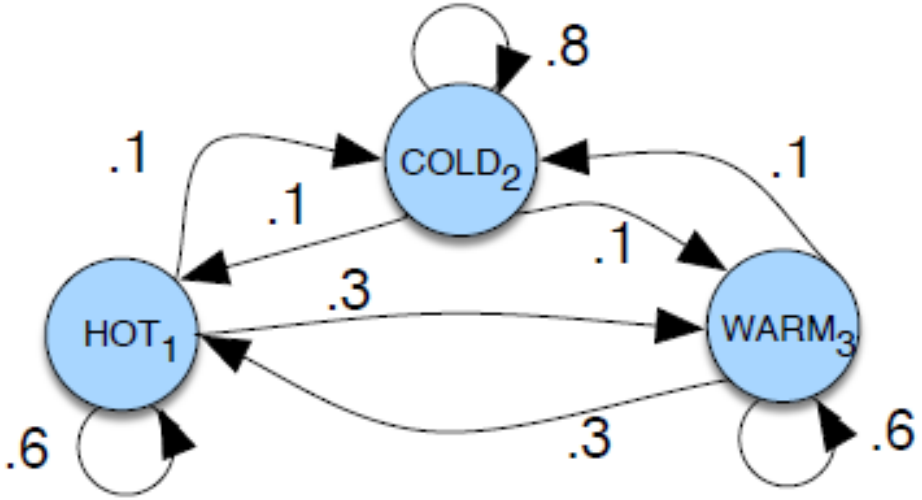
Applied Probability and Statistics

Module-3, Lecture-3

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

Forward Backward Algorithm

- A Markov chain for weather



- A start distribution π is required e.g $\pi = [0.1, 0.7, 0.2]$
- Markov assumption: $P(q_i = a | q_1 \ldots q_{i-1}) = P(q_i = a | q_{i-1})$

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$\pi = \pi_1, \pi_2, \ldots, \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

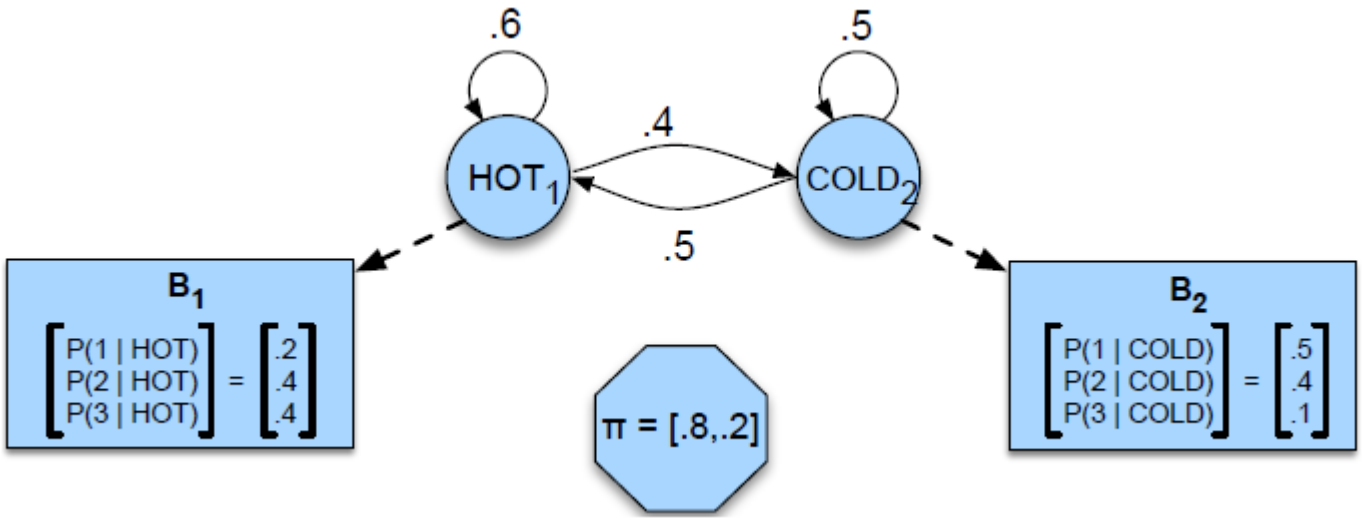
- Using the sample probabilities above with $\pi = [0.1, 0.7, 0.2]$, we can compute the probability of each of the following sequences:
  - hot hot hot hot
  - cold hot cold hot

- A hidden Markov model (HMM) allows us to talk about both observed events Markov model and hidden events that we think of as causal factors in our probabilistic model. An HMM is specified by the following components:

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} \ldots a_{ij} \ldots a_{NN}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{N} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, ..., v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i
$\pi = \pi_1, \pi_2, ..., \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

- A first-order hidden Markov model instantiates two simplifying assumptions
  - First, as with a first-order Markov chain, the probability of a particular state depends only on the previous state:
  - Markov assumption: $P(q_i = a|q_1 \ldots q_{i-1}) = P(q_i = a|q_{i-1})$
  - Second, the probability of an output observation o_i depends only on the state that produced the observation q_i and not on any other states or any other observations:
  - Output Independence: $P(o_i|q_1 \ldots, q_i \ldots, q_T, o_1, \ldots, o_i, \ldots, o_T) = P(o_i|q_i)$

- A hidden Markov model for relating numbers of ice creams eaten (the observations) to the weather (H or C, the hidden variables).



- Find out the state sequence if observation sequence is given.
- Find S^* given O
- $S^* = \frac{argmax}{s} P(S|O)$

- Hidden Markov models should be characterized by three fundamental problems:

Problem 1 (Likelihood):	Given an HMM $\lambda = (A, B)$ and an observation sequence O, determine the likelihood $P(O\|\lambda)$.
Problem 2 (Decoding):	Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the best hidden state sequence Q.
Problem 3 (Learning):	Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B.

function FORWARD-BACKWARD(*observations* of len T, *output vocabulary* V, *hidden state set* Q) **returns** *HMM=(A,B)*

initialize A and B

iterate until convergence

E-step

$$\gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{\alpha_T(q_F)} \quad \forall\, t \text{ and } j$$

$$\xi_t(i,j) = \frac{\alpha_t(i)\, a_{ij} b_j(o_{t+1})\beta_{t+1}(j)}{\alpha_T(q_F)} \quad \forall\, t,\ i, \text{ and } j$$

M-step

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1}\sum_{k=1}^{N} \xi_t(i,k)}$$

$$\hat{b}_j(v_k) = \frac{\sum_{t=1 s.t.\ O_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

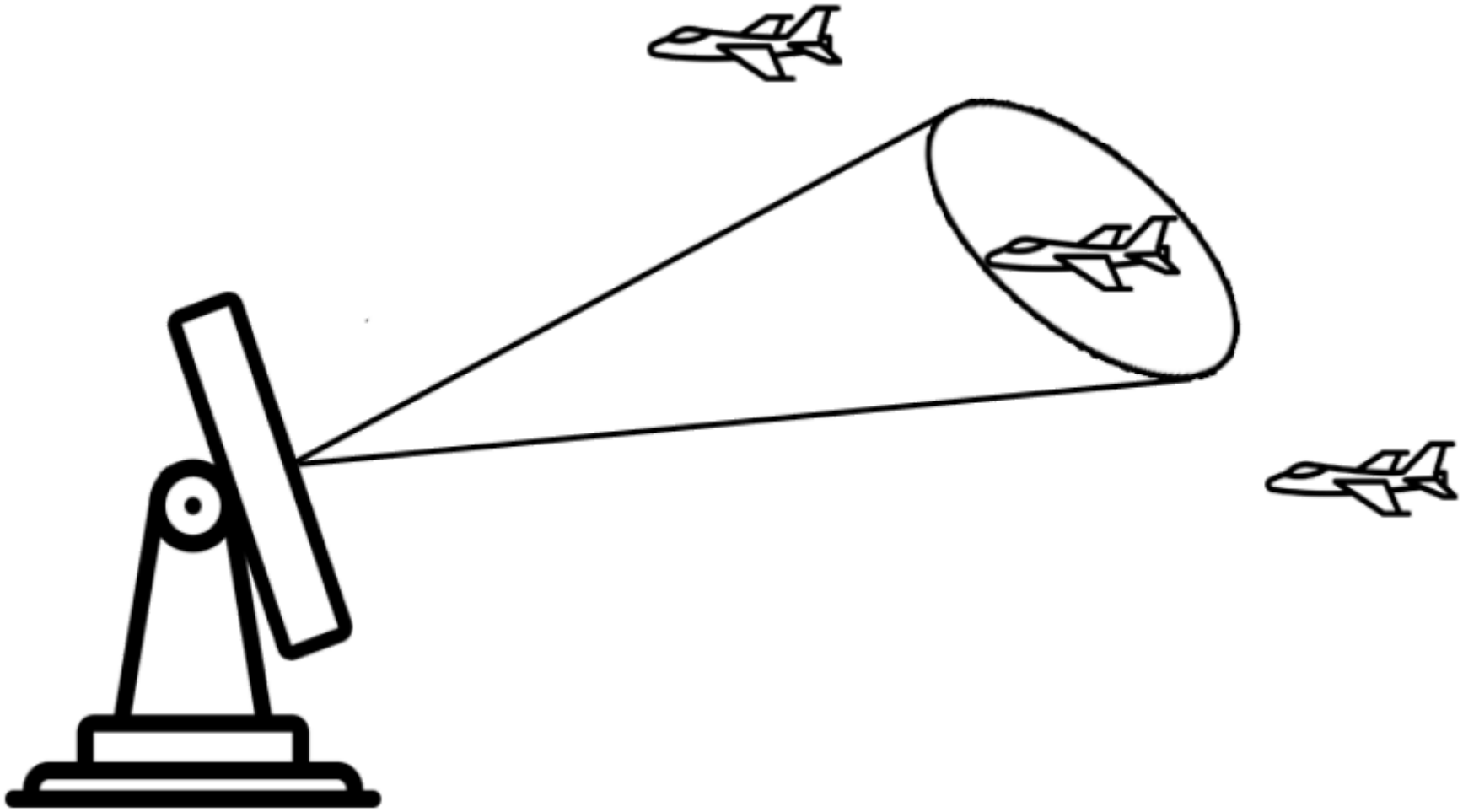
return A, B

Applied Probability and Statistics

Module-3, Lecture-4

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

Kalman filter

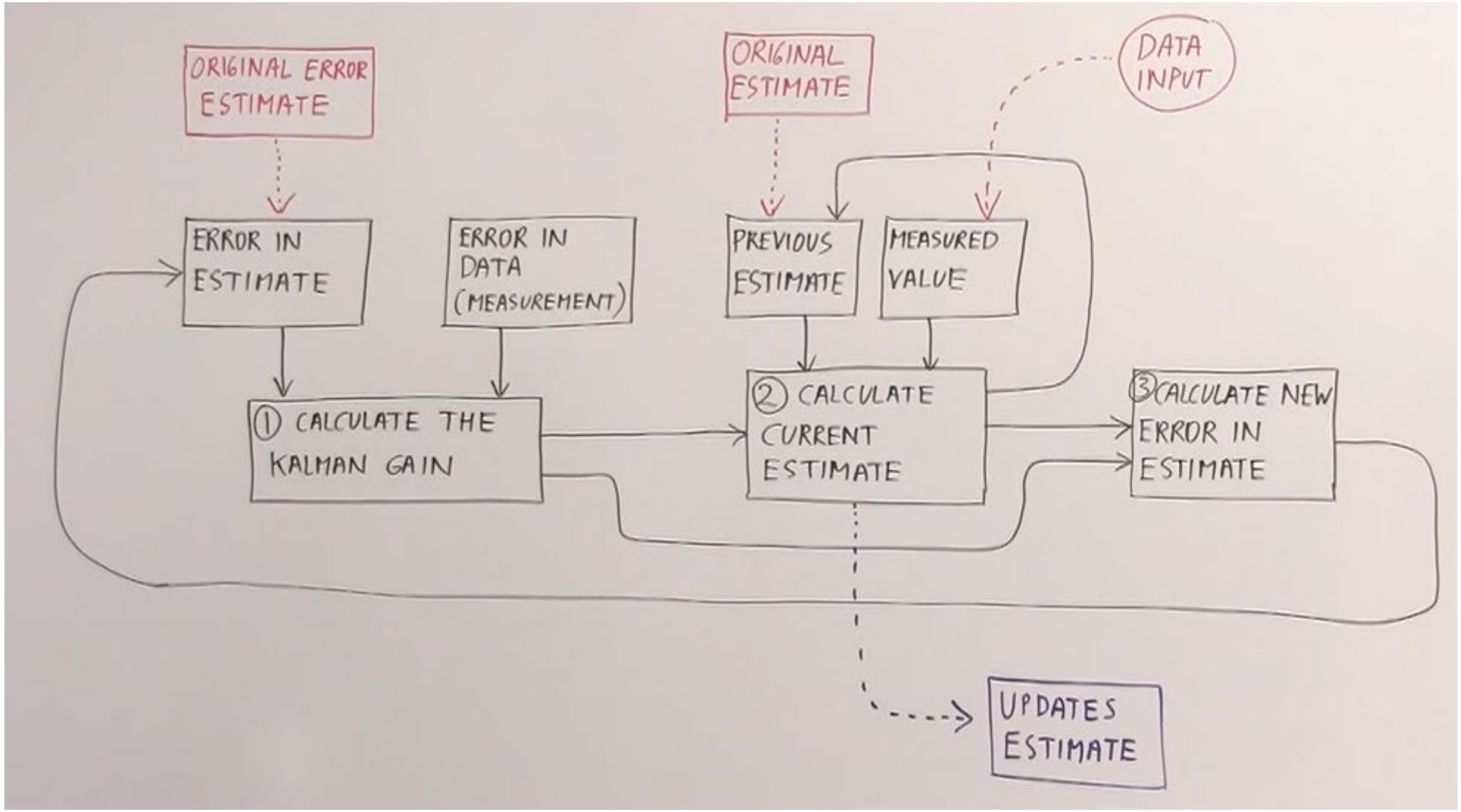


Kalman filter

- It is an optimal estimation algorithm
- It is an iterative mathematical process that uses a set of equations and consecutive data inputs to quickly estimate the true value, position, velocity, etc. of the object being measured, when the measured values contains unpredicted or random error, uncertainty or variation.
- The filter is named after Rudolf E. Kalman, who was one of the primary developers of its theory.

- Kalman filtering , also known as linear quadratic estimation (LQE), is an algorithm that uses a series of measurements observed over time, including statistical noise and other inaccuracies, and produces estimates of unknown variables that tend to be more accurate than those based on a single measurement alone, by estimating a joint probability distribution over the variables for each timeframe.

- Kalman filtering has numerous technological applications.
- A common application is for guidance, navigation, and control of vehicles, particularly aircraft, spacecraft and ships positioned dynamically .
- The concept much is applied in time series analysis used for topics such as signal processing and econometrics .
- Kalman filtering is also one of the main topics of robotic motion planning and control and can be used for trajectory optimization.
- Kalman filtering also works for modeling the central nervous system's control of movement.


ORIGINAL ERROR ESTIMATE
ORIGINAL ESTIMATE
DATA INPUT
ERROR IN ESTIMATE
ERROR IN DATA (MEASUREMENT)
PREVIOUS ESTIMATE
MEASURED VALUE
① CALCULATE THE KALMAN GAIN
② CALCULATE CURRENT ESTIMATE
③ CALCULATE NEW ERROR IN ESTIMATE
UPDATES ESTIMATE

- $KG = Kalman\ Gain$
- $E_{est} = Error\ in\ estimate$
- $E_{mae} = Error\ in\ measurement/observation$
- $EST_t = Current\ estimate$
- $EST_{t-1} = Previous\ estimate$
- $MAE = Measurement/observation$

- $KG = \frac{E_{est}}{E_{est}+E_{mea}}$
- $0 \leq KG \leq 1$
- $EST_t = EST_{t-1} + KG[MAE - EST_{t-1}]$
- $E_{est_t} = \frac{E_{mae}E_{est_{t-1}}}{E_{mae}+E_{est_{t-1}}} = [1 - KG]E_{est_{t-1}}$

- $x_t = Ax_{t-1} + Bu_{t-1} + w_{t-1}$
- $z_t = Hx_t + v_t$

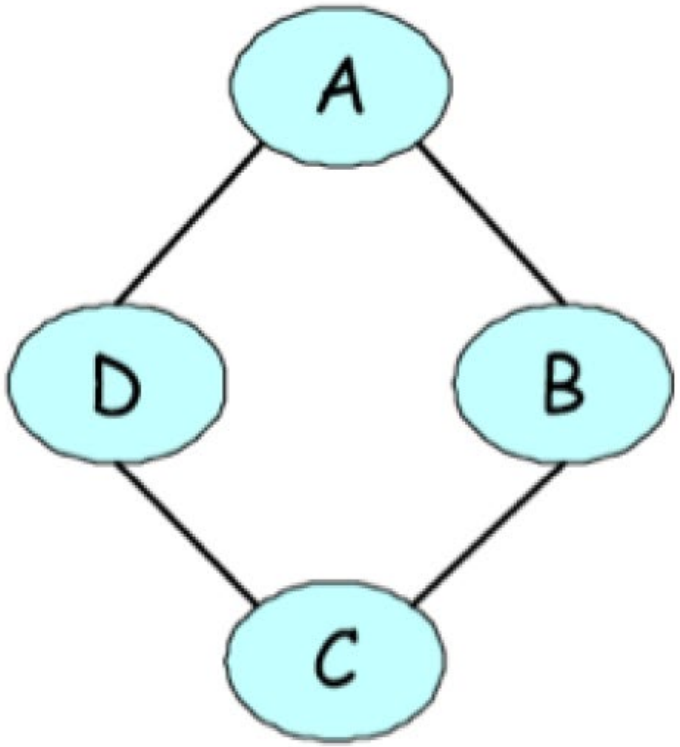
Applied Probability and Statistics

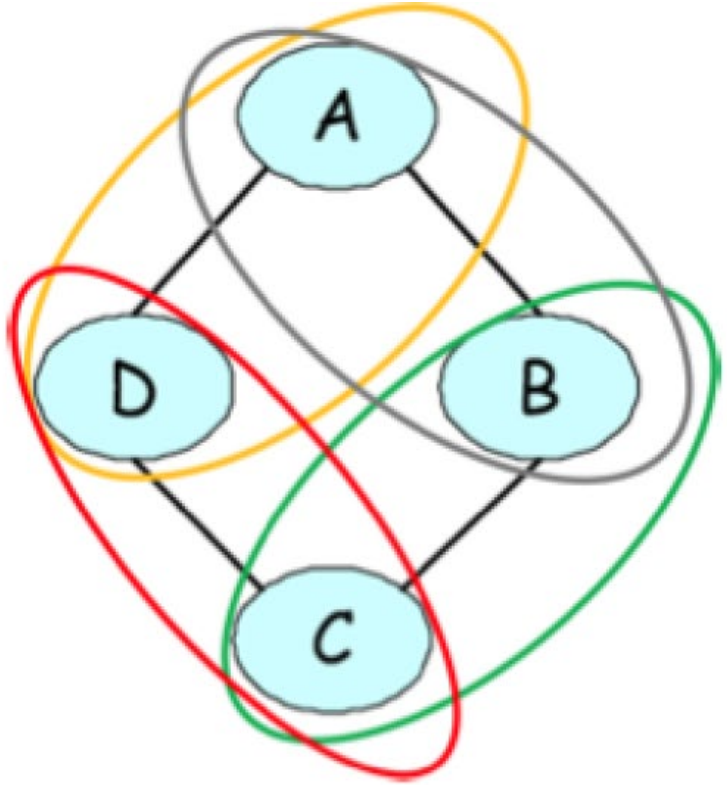
Module-3, Lecture-5

By: Dr. A K Yadav (9911375598)
Dept of CSE, ASET, AUUP

Markov Random Fields(MRF)

- Undirected graphical representation of a joint probability of voting preferences over four individuals.





- One way to define a probability over the joint voting decision of A, B, C, D is to assign scores to each assignment to these variables and then define a probability as a normalized score.
- A score can be any function, but in our case, we will define it to be of the form
- $\tilde{p}(A, B, C, D) = \phi(A, B)\phi(B, C)\phi(C, D)\phi(D, A)$
- Where $\phi(X, Y)$ is a factor that assigns more weight to consistent votes among friends X, Y

$$\phi(X,Y) = \begin{cases} 10 & \text{if } X = Y = 1 \\ 5 & \text{if } X = Y = 0 \\ 1 & \text{otherwise.} \end{cases}$$

- The factors in the unnormalized distribution are often referred to as *factors*
- . The final probability is then defined as

$$p(A,B,C,D) = \frac{1}{Z}\tilde{p}(A,B,C,D)$$

- Where is $Z = \sum_{A,B,C,D} \tilde{p}(A,B,C,D)$ a normalizing constant that ensures that the distribution sums to one.

- When normalized, we can view $\phi(A, B)$ as an interaction that pushes $B's$ vote closer to that of A.
- The term $\phi(B, C)$ pushes $B's$ vote closer to C, and the most likely vote will require reunite these conflicting influences.
- Note that unlike in the directed case, we are not saying anything about how one variable is generated from another set of variables (as a conditional probability distribution would do).

- We simply indicate a level of coupling between dependent variables in the graph.
- In a sense, this requires less prior knowledge, as we no longer have to specify a full generative story of how the vote of B is constructed from the vote of A (which we would need to do if we had a $P(B|A)$ factor).
- Instead, we simply identify dependent variables and define the strength of their interactions

- This in turn defines an energy landscape over the space of possible assignments and we convert this energy to a probability via the normalization constant.

Formal definition

A Markov Random Field (MRF) is a probability distribution p over variables $x_1, \ldots, x_n$ defined by an *undirected* graph G in which nodes correspond to variables x_i. The probability p has the form

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c),$$

where C denotes the set of *cliques* (i.e., fully connected subgraphs) of G, and each *factor* ϕ_c is a non-negative function over the variables in a clique. The *partition function*

$$Z = \sum_{x_1, \ldots, x_n} \prod_{c \in C} \phi_c(x_c)$$

is a normalizing constant that ensures that the distribution sums to one.

- Thus, given a graph G, our probability distribution may contain factors whose scope is any clique in G, which can be a single node, an edge, a triangle, etc.
- Note that we do not need to specify a factor for each clique.
- In our above example, we defined a factor over each edge (which is a clique of two nodes).
- However, we chose not to specify any unary factors, i.e., cliques over single nodes

More generally, MRFs have several advantages over directed models:

- They can be applied to a wider range of problems in which there is no natural directionality associated with variable dependencies.
- Undirected graphs can succinctly express certain dependencies that Bayesian nets cannot easily describe (although the converse is also true)

They also possess several important drawbacks:

- Computing the normalization constant Z requires summing over a potentially exponential number of assignments. We will see that in the general case, this will be NP-hard; thus, many undirected models will be intractable and will require approximation techniques.
- Undirected models may be difficult to interpret.
- It is much easier to generate data from a Bayesian network, which is important in some applications.

- It is not hard to see that Bayesian networks are a special case of MRFs with a very specific type of clique factor (one that corresponds to a conditional probability distribution and implies a directed acyclic structure in the graph), and a normalizing constant of one.
- If we take a directed graph G and add side edges to all parents of a given node (and removing their directionality), then the CPDs (seen as factors over a variable and its ancestors) factorize over the resulting undirected graph.
- The resulting process is called moralization

- A Bayesian network can always be converted into an undirected network with normalization constant one.
- The converse is also possible, but may be computationally intractable, and may produce a very large (e.g., fully connected) directed graph.
- Thus, MRFs have more power than Bayesian networks, but are more difficult to deal with computationally.
- A general rule of thumb is to use Bayesian networks whenever possible, and only switch to MRFs if there is no natural way to model the problem with a directed graph (like in our voting example).