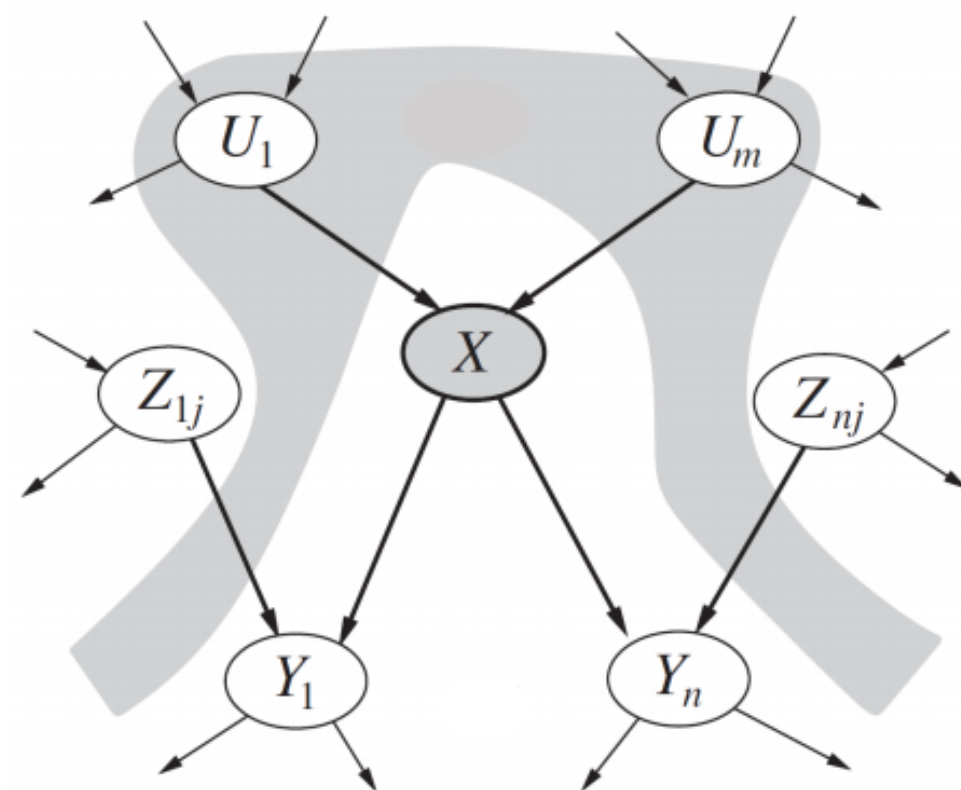


Bayesian Learning

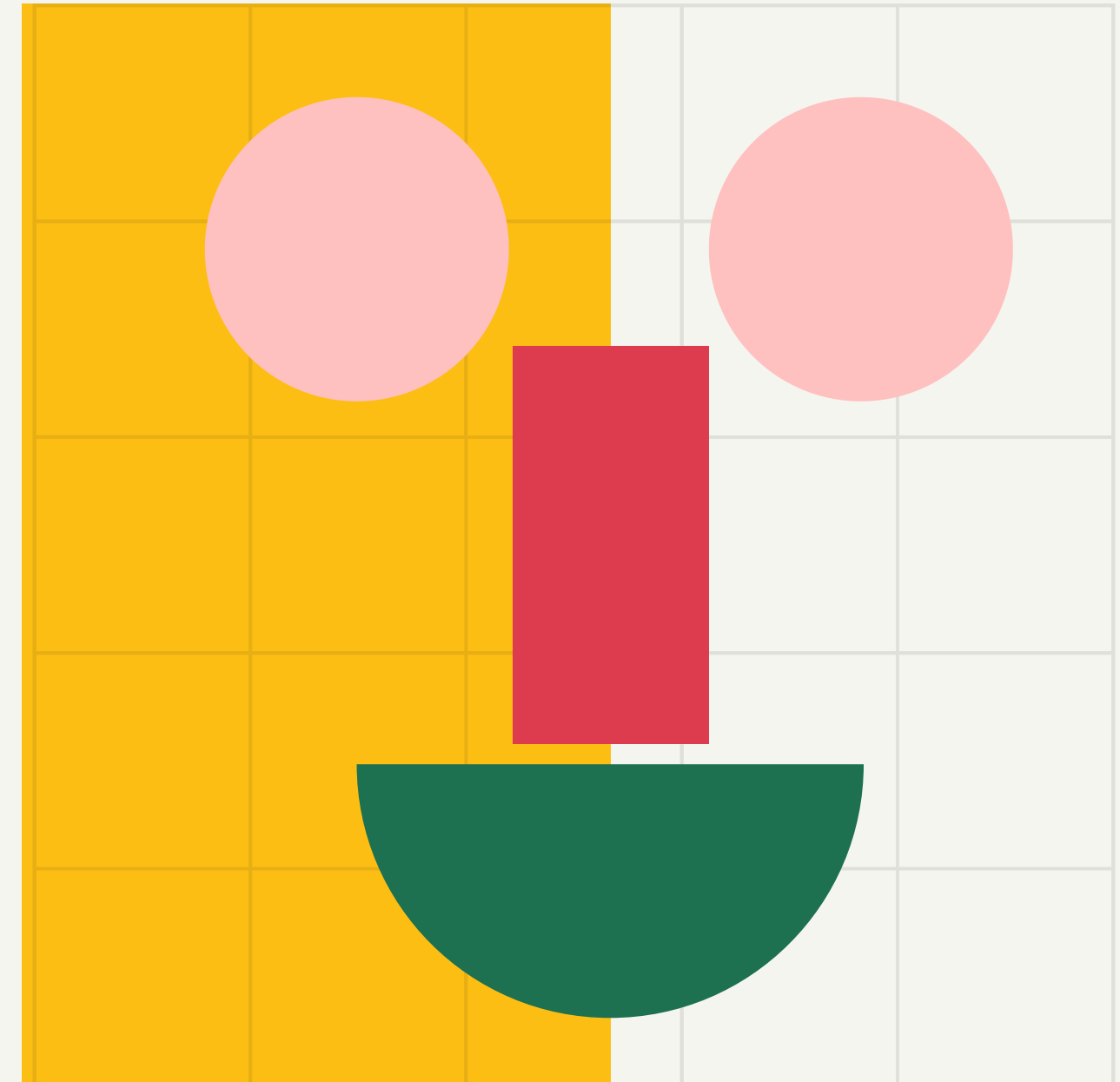


Sandesh Pokhrel - 075BCT076
Sanjay Bhandari - 075BCT079
Santosh Pangeneni - 075BCT082

Bayesian Learning

Today's Topics

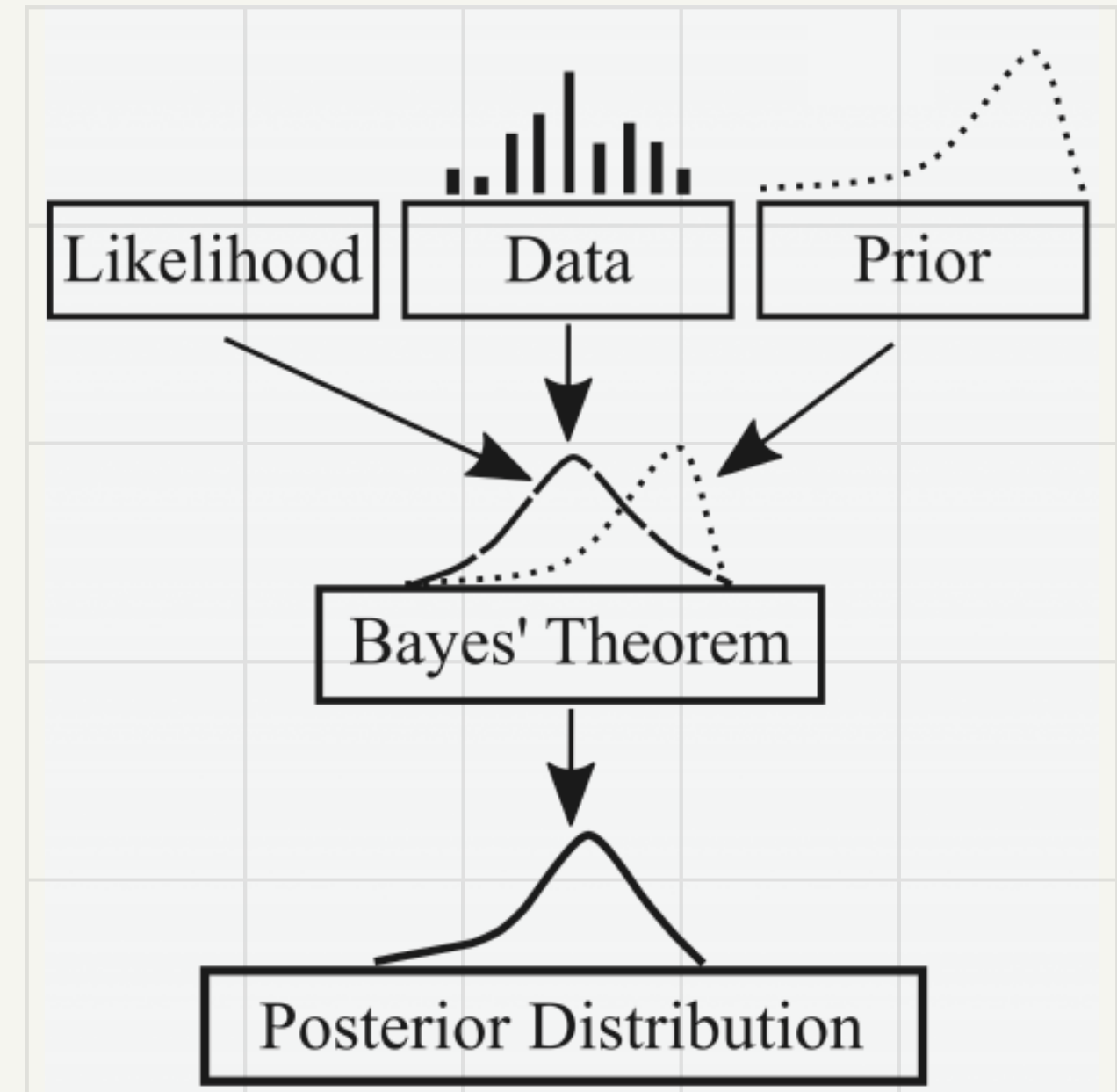
- Probability for learning
- Bayes Theorem
- MAP and ML Hypothesis
- Bayesian Optimal Classifier
- Gibbs Sampling
- Bayesian Network
- Inference and Applications
- Naive Bayes
- Project Demo



Probability for learning

Probability for classification and modelling concept

- **Bayesian Probability**
 - Notion of probability interpreted as partial belief
- **Bayesian Estimation**
 - Calculates the validity of proposition based on :
 - i. Prior estimates of its probability
 - ii. New relevant evidence



Bayes Theorem

- It describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Bayes Rule:
$$P(h | D) = \frac{P(D | h)P(h)}{P(D)}$$

- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h | D)$ = probability of h given D (posterior density)
- $P(D | h)$ = probability of D given h (likelihood of D given h)

Baes Analogy:

(pun intended)

 $P(\text{girl likes you} | \text{she smiled at you})$

$$= \frac{P(\text{she smiles at you} | \text{she likes you}) \times P(\text{she likes you})}{P(\text{she just smiles in general})}$$



Maximum A Posteriori (MAP) Hypothesis



Bayesian-based approach for estimating a distribution and model parameters that best explain an observed dataset. Mathematically,

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h | D) \\ &= \arg \max_{h \in H} \frac{P(D | h)P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D | h)P(h) \end{aligned}$$

Maximum Likelihood Hypothesis

- If every hypothesis in H is equally probable a priori, we only need to consider the likelihood of the data D given h , $P(D|h)$. Then, hMAP becomes the Maximum Likelihood Hypothesis. Mathematically,

$$\mathbf{h(ml)} = \mathbf{argmax} \mathbf{P(D \mid h_i)} \text{ where } \mathbf{h_i \in H}$$

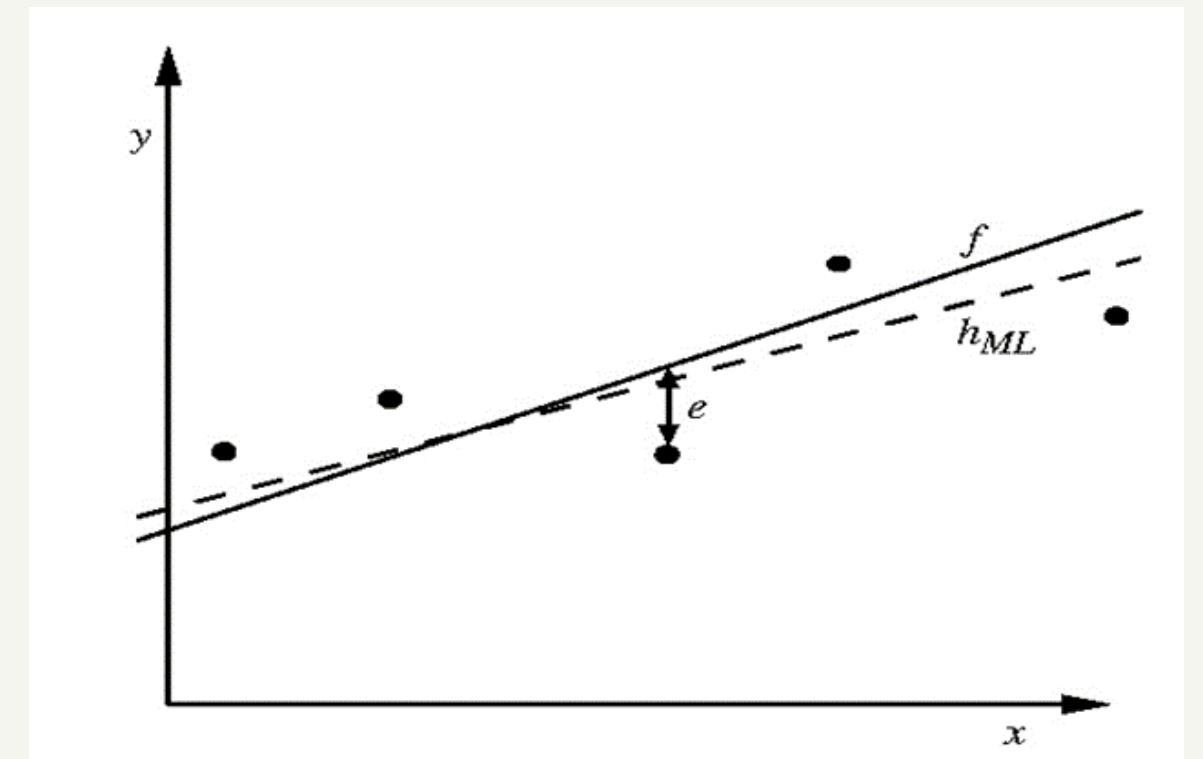
Example : ML Hypothesis

Learn a Real-Valued Function:

- Consider any real-valued target function f .
- Training examples (x_i, d_i) are assumed to have Normally distributed noise e_i with zero mean and variance σ^2 , added to the true target value $f(x_i)$
- d_i satisfies $N(f(x_i), \sigma^2)$

Assume that e_i is drawn independently for each x_i .

$$\begin{aligned} h_{ML} &= \arg \max_{h \in H} p(D | h) \\ &= \arg \max_{h \in H} \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2} \\ &= \arg \max_{h \in H} \sum_{i=1}^m -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\ &= \arg \min_{h \in H} \sum_{i=1}^m (d_i - h(x_i))^2 \end{aligned}$$

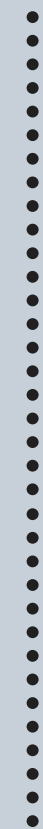


MAP Learner

Features

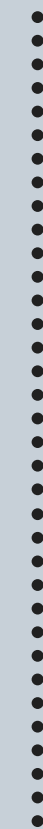
Prior Knowledge of Hypothesis

Choosing $P(h)$ and $P(D|h)$ reflects our prior knowledge about the learning task



Measure of performance

It provides a standard for judging the performance of learning algorithms



Computationally Intensive

Since prior probabilities of all hypothesis and probabilities of D for all h is to be computed for estimating the MAP hypothesis.

Bayesian Optimal Classifier

- It computes the posterior probabilities for every hypothesis and combines the predictions of each hypothesis to classify each new instance.

BAYES OPTIMAL
CLASSIFICATION:

$$\arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j | h_i) P(h_i | D)$$

Why optimal ?

No other classifier using the same H and prior knowledge can outperform it on average.

Given new instance x , what is its most probable classification?

- $h_{\text{MAP}}(x)$ is not the most probable classification!
- It is given by the Bayesian Optimal Classifier.

Example: Let $P(h_1 | D) = .4$, $P(h_2 | D) = .3$, $P(h_3 | D) = .3$

Given new data x , we have $h_1(x)=+$, $h_2(x)=-$, $h_3(x)=-$

What is the most probable classification of x ?

Solution:

$$P(h_1 | D) = .4,$$

$$P(h_2 | D) = .3,$$

$$P(h_3 | D) = .3,$$

$$P(- | h_1) = 0,$$

$$P(- | h_2) = 1,$$

$$P(- | h_3) = 1,$$

$$P(+ | h_1) = 1$$

$$P(+ | h_2) = 0$$

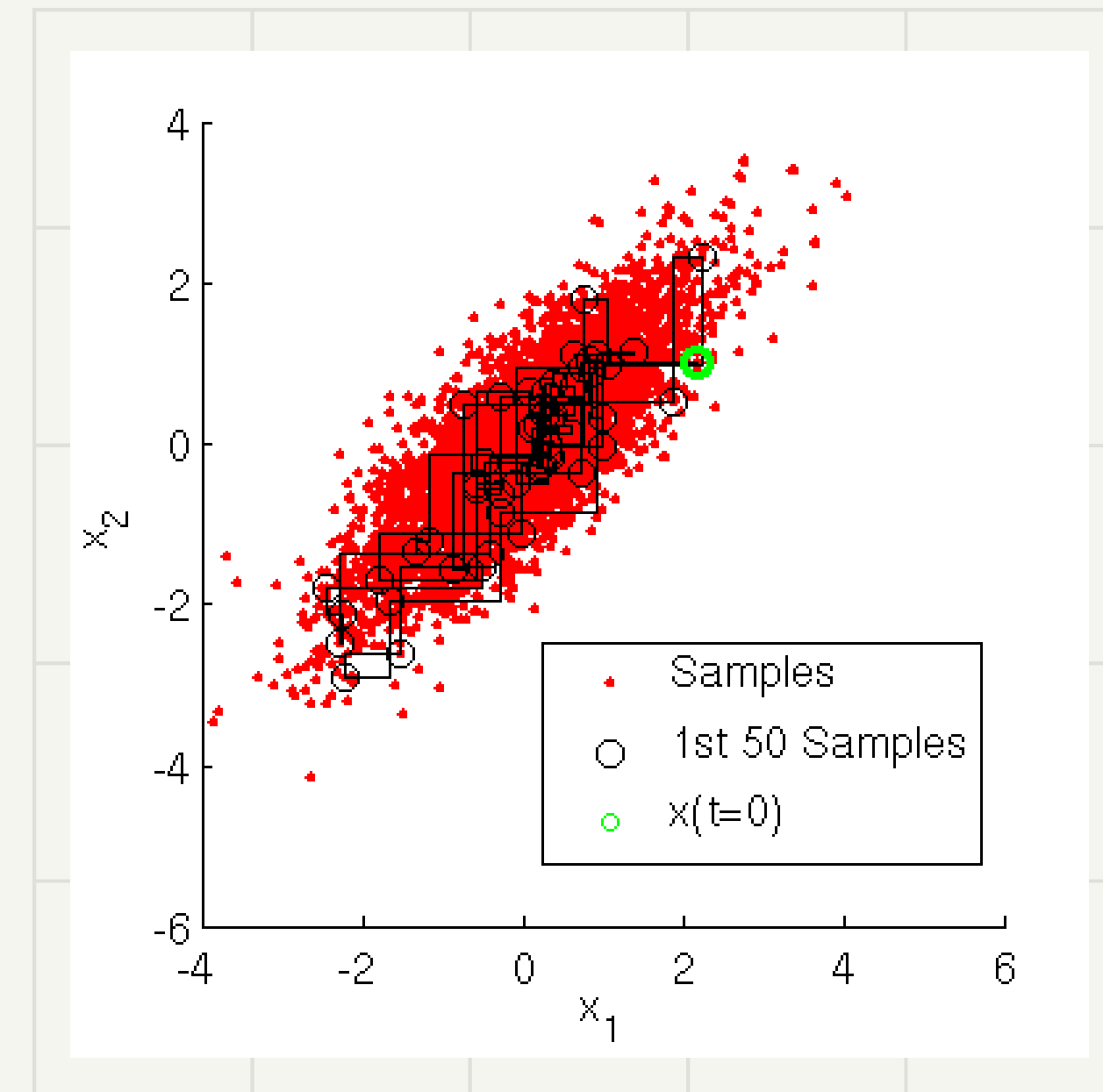
$$P(+ | h_3) = 0$$

$$\sum_{h_i \in H} P(+ | h_i) P(h_i | D) = .4$$

$$\sum_{h_i \in H} P(- | h_i) P(h_i | D) = .6$$

Gibbs Algorithm

- Bayes optimal classifier is quite computationally expensive, if H contains a large number of hypotheses.
- An alternative, less optimal classifier Gibbs algorithm, defined as follows:
 1. Choose a hypothesis randomly according to $P(h|D)$, where D is the posterior probability distribution over H .
 2. Use it to classify new instance



Error in Gibbs Algorithm

Assuming the expected value is taken over target concepts drawn at random, according to the prior probability distribution assumed by the learner, then

$$E_f[\text{error}_{X,f} \text{GibbsClassifier}] \leq 2E_f[\text{error}_{X,f} \text{BayesOptimal}],$$

where f denotes a target function, X denotes the instance space.

BAYESIAN NETWORK

A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph.

Bayesian Network

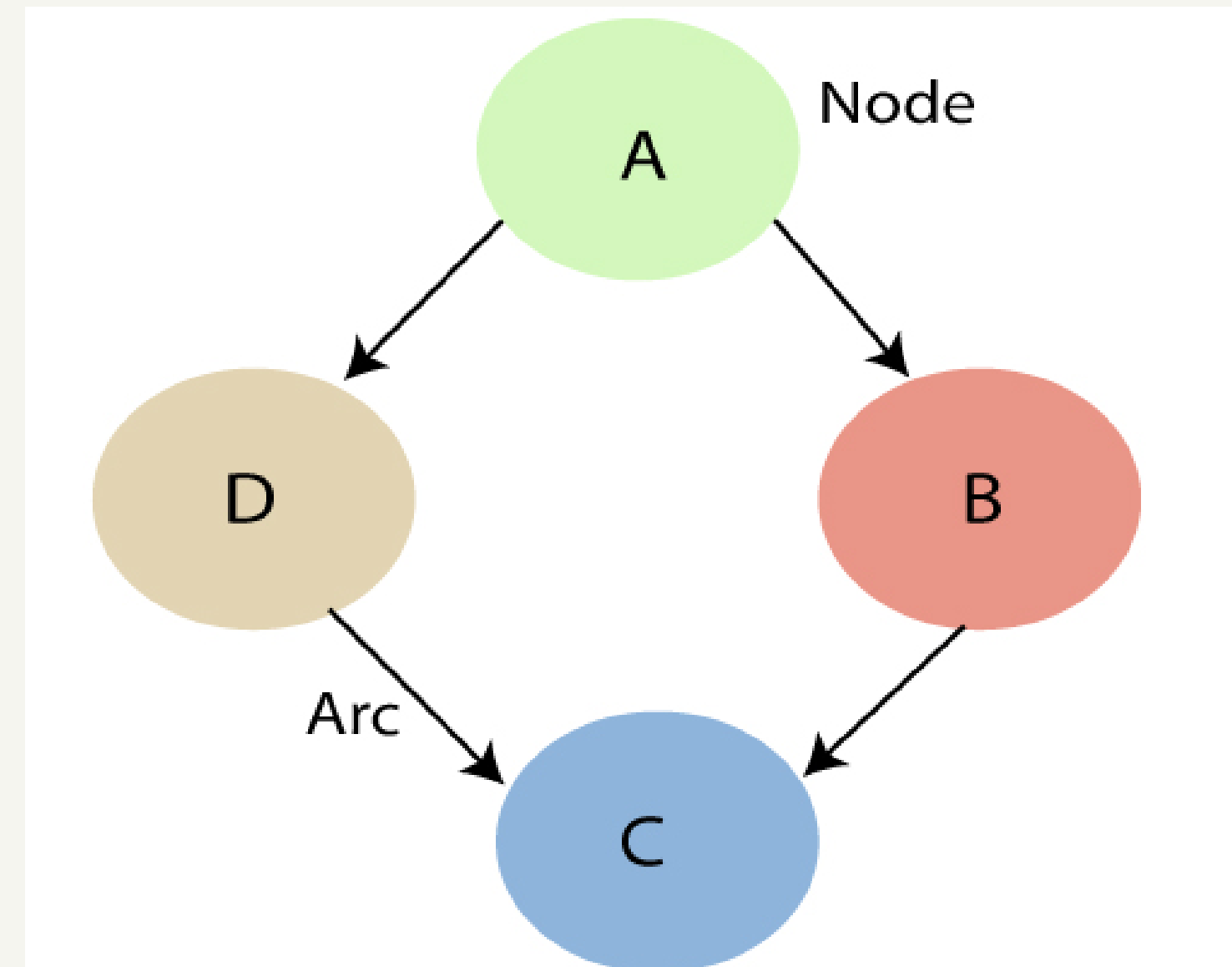
Bayesian Network is an efficient representation of the joint probability distribution of the variables that can be used for building models from data and experts opinions.

It consists of two parts:

- Directed acyclic graph
- Table of conditional probabilities

Directed Acyclic Graph



- Each node corresponds to the random variables, and a variable can be continuous or discrete.
 $X = \{ X_1, \dots, X_n \}$
- Arcs represent probabilistic dependence among variables. Lack of an arc denotes a conditional independence.



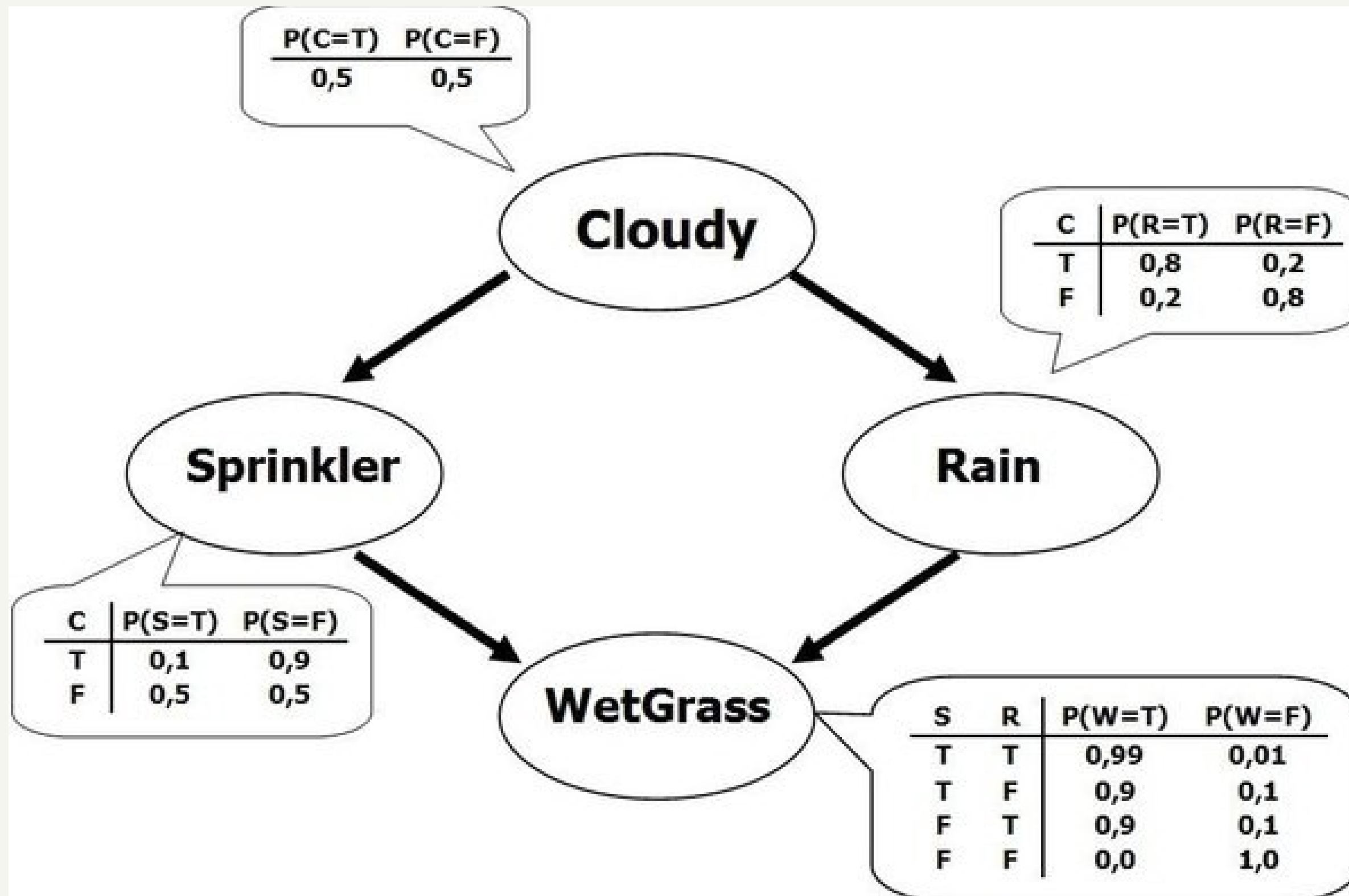
Conditional Probabilities Table

- Each node in the Bayesian network has conditional probability distribution $P(X_i | \text{Parent}(X_i))$, which determines the effect of the parent on that node.
- Each node is asserted to be conditionally independent of its non-descendants, given its immediate parents.
i.e, Node C is independent of node A but dependent to B and D.

Inference In Bayesian Networks

- Computes posterior probabilities given evidence about some nodes
 - Exploits probabilistic independence for efficient computation.
 - Though exact inference is known to be NP-hard, approximation techniques, in practice are shown to be useful.
- 
- 

Bayesian Network Example



Bayesian Network Inferences

i. All True,

$$\begin{aligned}P(W,R,S,C) &= P(W \mid R,S) * P(R|C) * P(S|C) * P(C) \\&= 0.5 * 0.1 * 0.8 * 0.99 \\&= 0.396\end{aligned}$$

ii. Grass is wet, when the day is cloudy and sprinkler is on.

$$\begin{aligned}P(W, C, S) &= P(W \mid S,R) * P(S|C) * P(C) + P(W|S,\neg R) * P(S,C) * P(C) \\&= 0.99*0.1*0.8 + 0.9*0.1 *0.8 \\&= 0.1512\end{aligned}$$

Bayesian Network

- Structure of the graph \Leftrightarrow Conditional independence relations

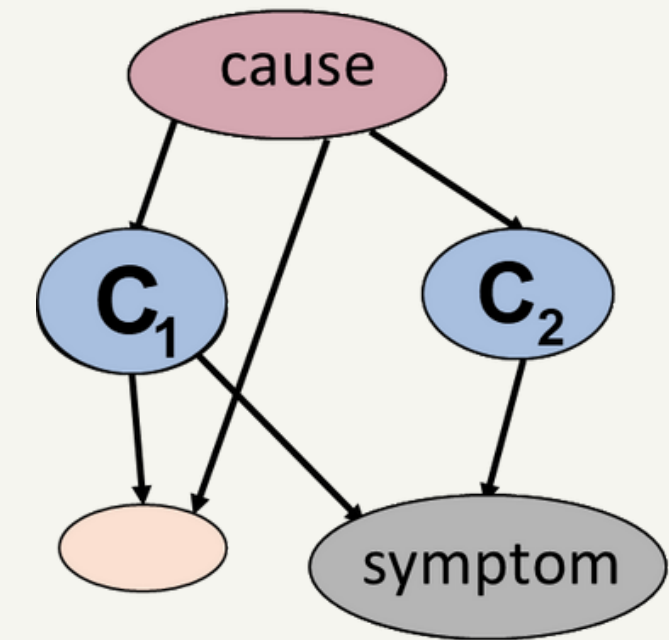
- In general,

$$P(X_1, X_2, \dots, X_N) = \prod P(X_i \mid \text{parents}(X_i))$$

Full-joint distribution

Graph structured
approximation

Applications



Diagnosis

$$P(\text{cause}|\text{symptom}) = ?$$



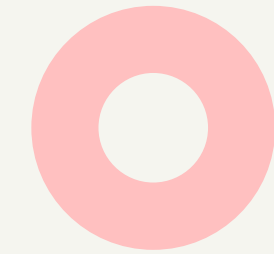
Prediction

$$P(\text{symptom}|\text{cause}) = ?$$



Classification

$$P(\text{class}|\text{data}) = ?$$



Decision Making

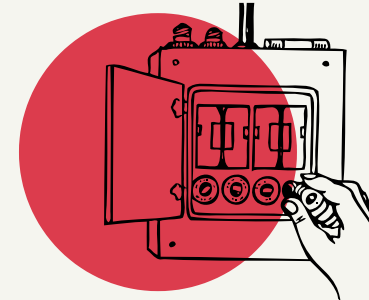
When given a cost function

Real World Uses



Medical Diagnosis

Prevalence of a certain disease can be inferred through a Bayesian Network when there is presence of symptoms.



Fault Diagnosis

Electrical Engineers use Bayesian Networks to find relation between system behavior and faults.

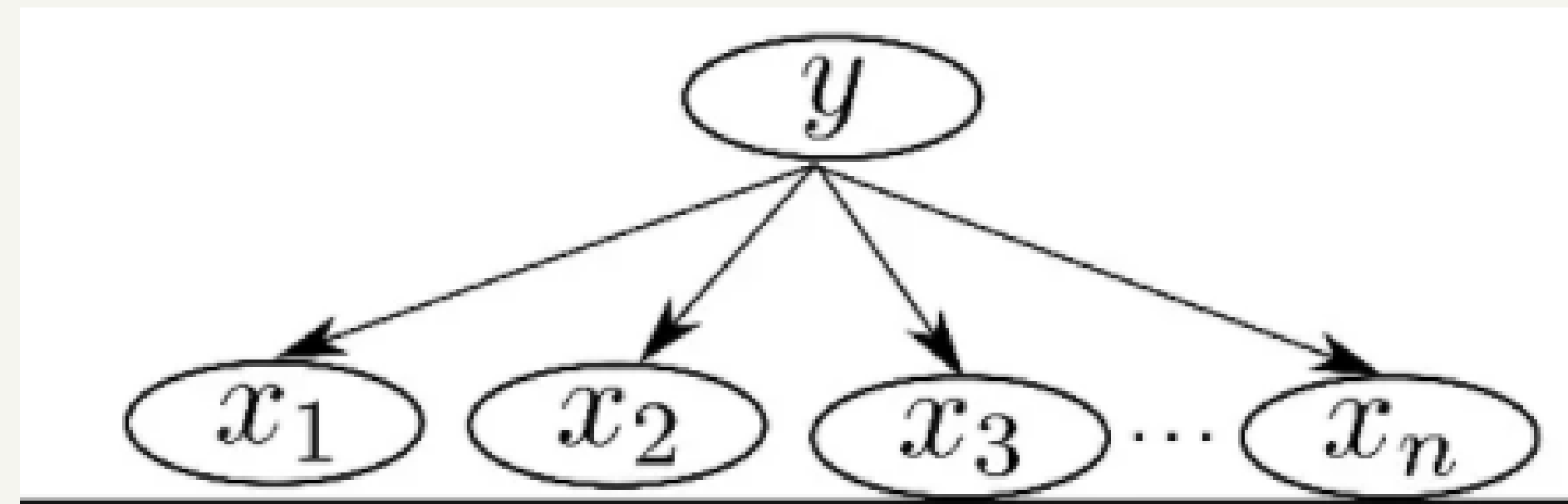


Spam Filter

A particular class of Bayesian Network is used in classification, Naïve Bayes is prominent in this regard.

Naive Bayes

- Naive as it assumes no dependency between input variables
- When to use?
- Applications



Project Demo

SPAM FILTER

Naive Bayes

**And we're done
for the day!**