Bayesian Networks 1: Introduction

Bayesian Networks

Goal: represent a joint distribution P over random variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$

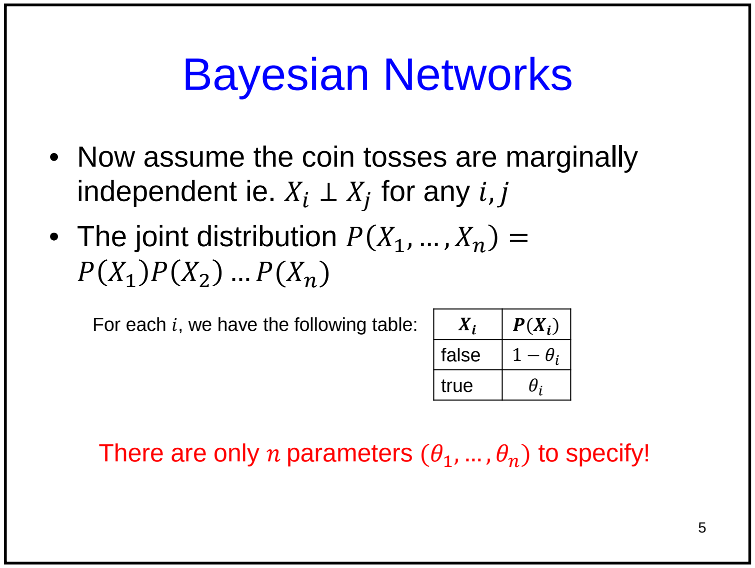
X_1	X_2	$P(X_1, X_2)$
false	false	0.1
false	true	0.2
true	false	0.3
true	true	0.4

Bayesian Networks

- If variables are binary, the joint distribution has $2^n - 1$ probabilities
  - Expensive space usage
  - Human expert has hard time determining these numbers
  - Need large amounts of data to estimate these numbers accurately
- How do we represent a joint probability distribution compactly?
  - Solution: Exploit independence properties

Bayesian Networks

- Suppose we toss n coins and let X_i be the outcome of coin toss i
- The joint distribution $P(X_1, \ldots, X_n)$ has $2^n - 1$ parameters



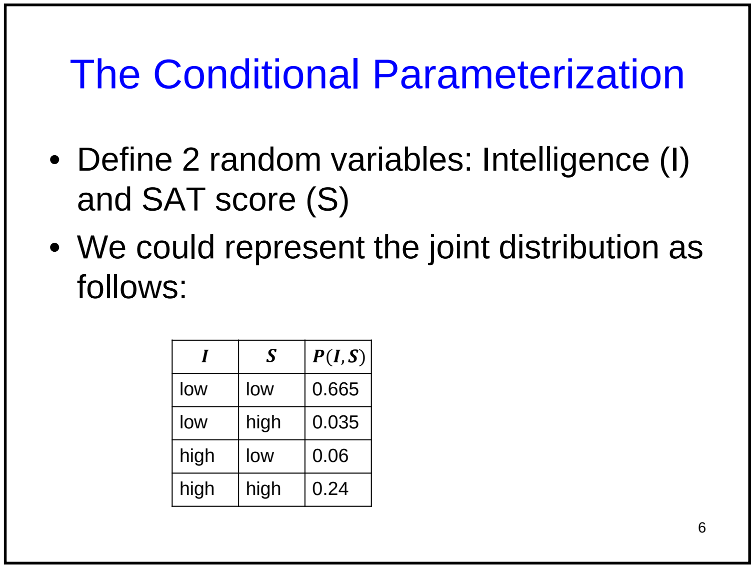
Bayesian Networks

- Now assume the coin tosses are marginally independent ie. $X_i \perp X_j$ for any i, j
- The joint distribution $P(X_1, \dots, X_n) = P(X_1)P(X_2) \dots P(X_n)$

For each i, we have the following table:

X_i	$P(X_i)$
false	$1 - \theta_i$
true	θ_i

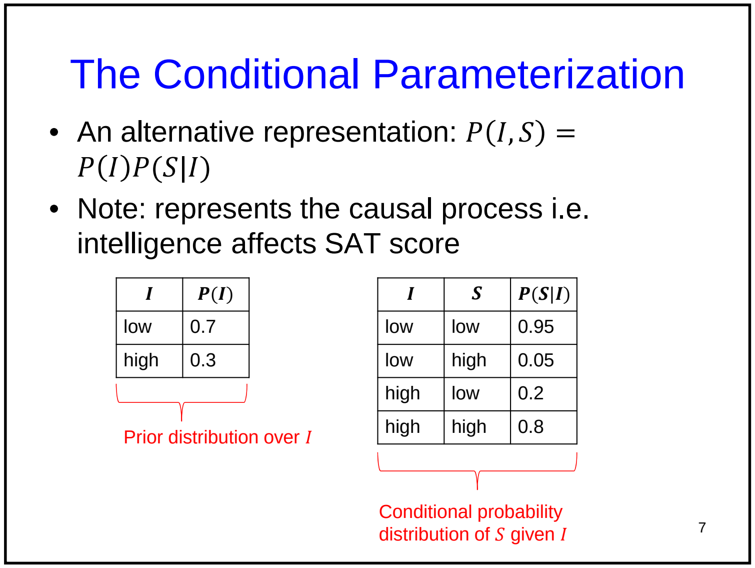
There are only n parameters $(\theta_1, \dots, \theta_n)$ to specify!



The Conditional Parameterization

- Define 2 random variables: Intelligence (I) and SAT score (S)
- We could represent the joint distribution as follows:

I	S	$P(I,S)$
low	low	0.665
low	high	0.035
high	low	0.06
high	high	0.24



The Conditional Parameterization

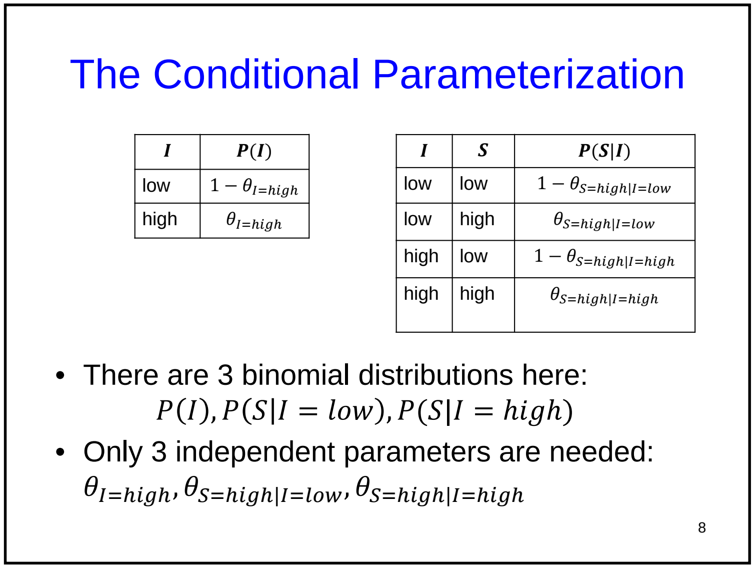
- An alternative representation: $P(I,S) = P(I)P(S|I)$
- Note: represents the causal process i.e. intelligence affects SAT score

I	$P(I)$
low	0.7
high	0.3

Prior distribution over I

I	S	$P(S\|I)$
low	low	0.95
low	high	0.05
high	low	0.2
high	high	0.8

Conditional probability distribution of S given I



The Conditional Parameterization

I	$P(I)$
low	$1 - \theta_{I=high}$
high	$\theta_{I=high}$

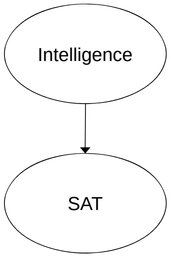
I	S	$P(S\|I)$
low	low	$1 - \theta_{S=high\|I=low}$
low	high	$\theta_{S=high\|I=low}$
high	low	$1 - \theta_{S=high\|I=high}$
high	high	$\theta_{S=high\|I=high}$

- There are 3 binomial distributions here:
  $P(I), P(S|I = low), P(S|I = high)$
- Only 3 independent parameters are needed: $\theta_{I=high}, \theta_{S=high|I=low}, \theta_{S=high|I=high}$

The Conditional Parameterization

The joint distribution (conditional parameterization version) drawn as a Bayesian network looks like:

$P(I, S) = P(I)P(S|I)$



Naïve Bayes

Naïve Bayes

- Now assume we have 3 random variables:
  - Intelligence: low, high
  - SAT score: low, high
  - Grade: A, B, C
- No independencies that hold:
  - Intelligence correlated with SAT score and grade
  - SAT score and grade not independent

Naïve Bayes

- But conditional independencies hold!
- If a student has high intelligence, a high SAT score no longer gives us information about the student's grade
- Formally: $\underbrace{(S \perp G|I)}$

S and G are conditionally independent given I

Note: This is only true if intelligence is the only reason by his grade and SAT score might be correlated

Naïve Bayes

- This leads to the following factored representation:

$P(I,S,G) = P(S,G|I)P(I)$ [As before]

$\quad = P(S|I)P(G|I)P(I)$ [Conditional independence: $P(S,G|I) = P(S|I)P(G|I)$]

There are 3 binomial distributions:

- $P(I)$ with parameter: $\theta_{I=high}$
- $P(S|I = low)$ with parameter: $\theta_{S=high|I=low}$
- $P(S|I = high)$ with parameter: $\theta_{S=high|I=high}$

- And 2 three-valued multinomial distributions:
  - $P(G|I = low)$ with parameters: $\theta_{G=A|I=low}, \theta_{G=B|I=low}$
  - $P(G|I = high)$with parameters: $\theta_{G=A|I=high}, \theta_{G=B|I=high}$

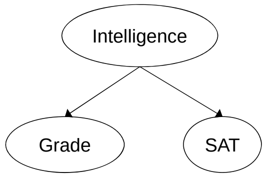
7 vs 11 independent parameters for a full joint distribution

Naïve Bayes

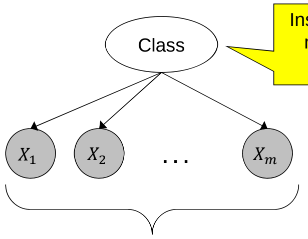
I	$P(I)$
low	$1-\theta_{I=high}$
high	$\theta_{I=high}$

I	S	$P(S|I)$
low	low	$1-\theta_{S=high|I=low}$
low	high	$\theta_{S=high|I=low}$
high	low	$1-\theta_{S=high|I=high}$
high	high	$\theta_{S=high|I=high}$

I	G	$P(G|I)$
low	C	$1-\theta_{G=B|I=low}-\theta_{G=A|I=low}$
low	B	$\theta_{G=B|I=low}$
low	A	$\theta_{G=A|I=low}$
high	C	$1-\theta_{G=B|I=high}-\theta_{G=A|I=high}$
high	B	$\theta_{G=B|I=high}$
high	A	$\theta_{G=A|I=high}$



Naïve Bayes



Instances fall into one of k mutually exclusive and exhaustive classes

Features: characteristics of the instances that help predict the class. These are typically observed.

Naïve Bayes assumption: features are conditionally independent given the instance's class ie.

$(X_i \perp \boldsymbol{X_{-i}}|C)$ for all i

where $\boldsymbol{X_{-i}} = \{X_1, \ldots, X_m\} - \{X_i\}$

Naïve Bayes

Based on these assumptions, the joint distribution factorizes as:

$$P(C, X_1, \ldots, X_m) = P(C)\prod_{i=1}^{m} P(X_i|C)$$

If all the variables are binary, there are a total of $(2m + 1)$ independent parameters needed to specify the naive Bayes model

Bayesian Networks

Bayesian Network

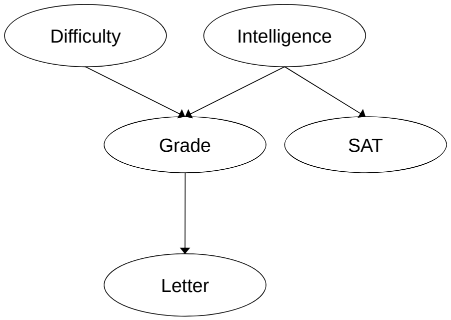
A Bayesian network is composed of:

- The DAG structure
- The conditional probability distributions in each node

Bayesian Networks

- A Bayesian network is represented as a Directed Acyclic Graph (DAG) *G*
  - Nodes are random variables
  - Edges correspond to the direct influence of one random variable on another
- *G* can be viewed in two different ways:
  - The skeleton for a compact, factored representation of a joint distribution
  - A compact representation for a set of conditional independence assumptions about a distribution
- Both are equivalent

Bayesian Networks

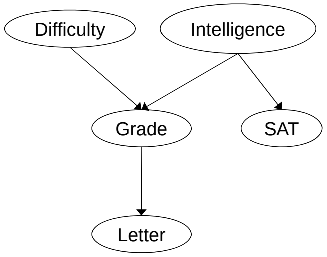


DAG Structure: intuitively, each variable depends directly only on its parents

Bayesian Networks

D	P(D)
low	0.6
high	0.4

D	I	G	P(G|D,I)
low	low	C	0.3
low	low	B	0.4
low	low	A	0.3
low	high	C	0.02
low	high	B	0.08
low	high	A	0.9
high	low	C	0.7
high	low	B	0.25
high	low	A	0.05
high	high	C	0.2
high	high	B	0.3
high	high	A	0.5



I	P(I)
low	0.7
high	0.3

I	S	P(S|I)
low	low	0.95
low	high	0.05
high	low	0.2
high	high	0.8

G	L	P(L|G)
C	weak	0.99
C	strong	0.01
B	weak	0.4
B	strong	0.6
A	weak	0.1
A	strong	0.9

Bayesian Networks

Each node has a local probability model

- Captures the conditional probability distribution of the node given its parents ie. $P(X|Parents(X))$
- Specifies a distribution over each value of X given each possible joint assignment of values to its parents
- A node with no parents eg. $P(I)$ is conditioned on the empty set of variables and is a marginal distribution

Bayesian Networks

With a Bayesian network, you can compute the value of any state of the joint probability distribution

$$P(I = high, D = low, G = B, S = high, L = weak)$$
$$= P(I = high)P(D = low)P(G = B \mid I = high, D = low)*$$
$$P(S = high \mid I = high)P(L = weak \mid G = B)$$
$$= 0.3*0.6*0.08*0.8*0.4 = 0.004608$$

This uses the chain rule for Bayesian networks (more on this later)