Independence in probabilities, Bayes' nets

Some material on these slides are derived from the slides created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley, available at http://ai.berkeley.edu.

Models

- A **model** is a way to describe how the world works (or at least, the portion of interest to us)
- You have learned about many models in physics, chemistry, biology, economics
  - The so-called "laws" are very often better described as "models"
- Models are not reality!
  - Even in the best case, they only model the aspects of reality we are interested in
  - We ignore some variables because (a) we are not interested in them or (b) we cannot measure them or (c) we cannot handle the computation...

Probabilistic models

- Models that describe the world through random variables and their relationships
- Sources of randomness:
  - **Intrinsic randomness** in the world we model: eg. our returns at the casino
  - **Lack of knowledge** in predictions: eg. predicting tracks of hurricanes.
  - **Populations** we look at measurements averaged over populations (or repeated samples). Example: predicting
  - **Noisy observations** the world is deterministic, but our observations are affected by random noise
  - **Preference for probabilistic models** sometimes we prefer a probabilistic model because the math works out better. Example: diffusion models for image generation.

How do we use probabilistic models

- The world is described by a set of random variables
- We have some information about the distributions of these variables and the relations between them
  - Some individual probabilities, some joints, some conditionals
- Some variables are known (**evidence**)
- Using the relations between the variables, we answer questions about the world such as:
  - **explanation** - diagnostic reasoning
  - **prediction** - what is the likely value of a variable?
  - **value of information** - what other information do we need to better answer a question?

The best probabilistic model, and why we cannot have it

- Let us say that we describe the world with random variables $X_1, X_2, \ldots X_n$
- If we have the joint probability table $P(X_1, X_2, \ldots P_n)$ all our problems are solved!
  - We can answer any question, for any evidence by simply applying the reasoning process described in the last class: select rows compatible with evidence + marginalize!
- The problem:
  - The joint table is huge! If $n = 10$ and each random variable has 10 possible values, we need 10^{10} probabilities.
    - And this is nothing compared to, eg. the stock market
  - How are we going to build a model this big? Who is going to give us the probabilities?

Two fundamentally different ideas to build a probabilistic model

- Idea 1: Introduce structure
- Idea 2: Use learning

Structured probabilistic models

- We assume that the joint probability table does not contain all independent values. For instance, some of the values can be described by mathematical operations on a smaller number of probabilities
- This can happen, for instance, if some of the variables are **independent** of each other.
- These relationships can be described through some kind of **structure**, often described as a graph
- Examples:
  - Bayesian Networks (directed graphs)
  - Markov Random Fields (undirected graphs)
  - Factor Graphs
  - Probabilistic Circuits, etc.

Learned probabilistic models

- We assume that the probabilities are described by a certain parameterized function $P(x_1, x_2, \ldots x_n) = f(x_1, x_2, \ldots x_n; \Theta)$
    - We use data to learn the parameter Θ
- Examples:
    - Classical parametric models: eg Gaussian Mixture Models
    - Latent variable models
    - Deep generative models: use deep neural networks for f
        - Variational autoencoders, Normalizing flows, Diffusion models
    - Energy based models
        - Define probability via an unnormalized energy function (we don't need to calculate the normalizing factor Z)

What we will do today?

- One of the simplest ways to simplify probabilistic models:
    - Introduce the concept of independence and conditional independence
    - Create a directed graph (the bayesian network) what describes the set of dependencies between the variables of the model.

Independence

- Two variables are independent if:

$$\forall x, y : P(x, y) = P(x)P(y)$$

- This says that their joint distribution factors into a product of two simplest distributions.
- Why is this "independence"? Another way to write this (direct consequence of the definition of a conditional)

$$\forall x, y : P(x|y) = P(x)$$

We write

$$X \perp\!\!\!\perp Y$$

Independence is a simplifying assumption

- If we have 10 variables with 10 possible values
  - If they are **not independent** we need 10^10 probability values to describe the joint
  - If they are **independent** we need 100 probability values
- Are there many independent variables in the world?
  - Sure: throwing a dice in Las Vegas and an airport delay in Orlando are usually independent!
  - N fair, independent coin flips are independent
- But any set of variables chosen for a given problem, they are rarely fully independent!

Conditional independence

- X is conditionally independent of Y given Z

$$X \perp\!\!\!\perp Y | Z$$

if and only if

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently

$$\forall x, y, z : P(x|y, z) = P(x|z)$$

Importance of conditional independence

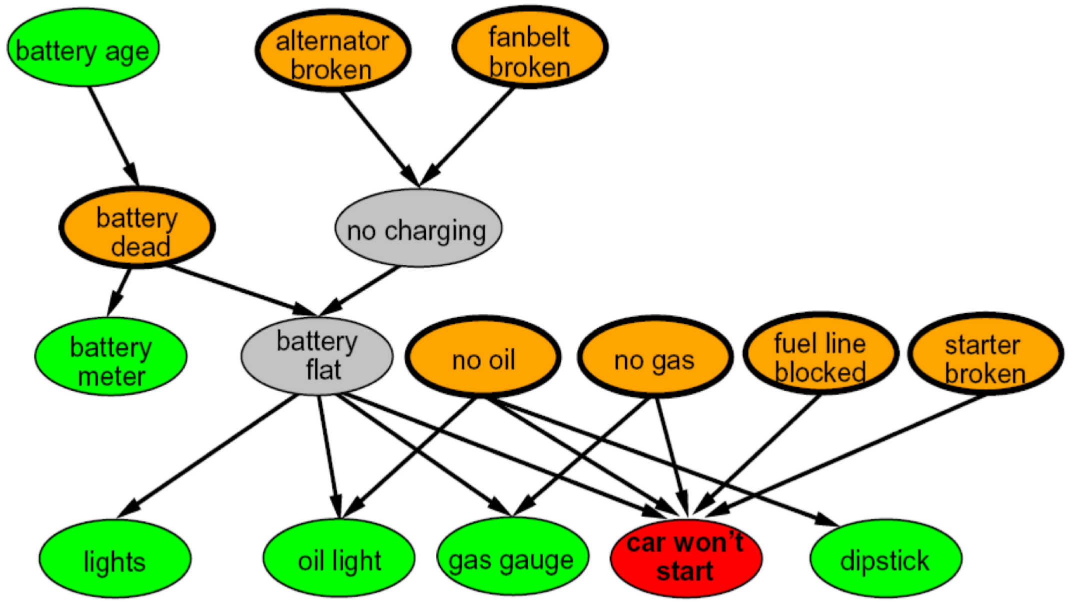
- It is much more likely that we find conditional independence in the variables of our problem.

Example with conditional independence here

Bayes Nets

- We have a probabilistic model, which would require us to have the full joint table.
- We will describe the joint table using simple, local, conditional distributions
- We describe how the variables interact locally using a graph
    - part of a family called **probabilistic graphical models**
- Local interactions chain together to give global interactions

Example Bayes Net for diagnosing a car that won't start



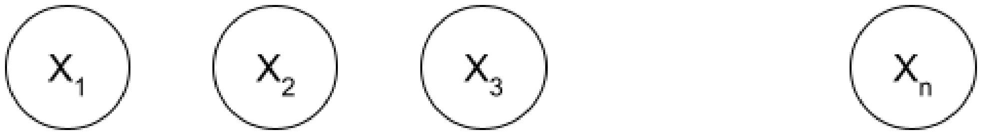
Notice that there are relatively few dependencies between the nodes!

Notations

- Nodes: variables with domains
  - Can be assigned (observed) or unassigned (unobserved)
- Directed edges
  - Indicate **direct influence** between variables
  - Formally: encode conditional independence
- Although this is not strictly correct, you can imagine that the edges mean direct causation.

Example: coin flips

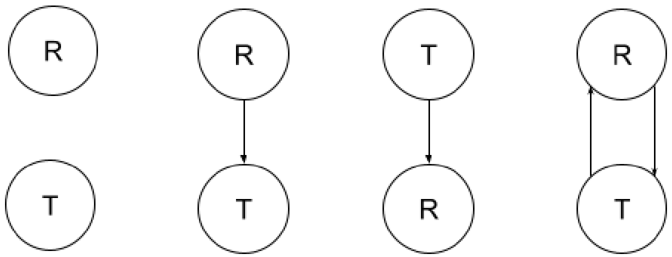
- N independent coin flips



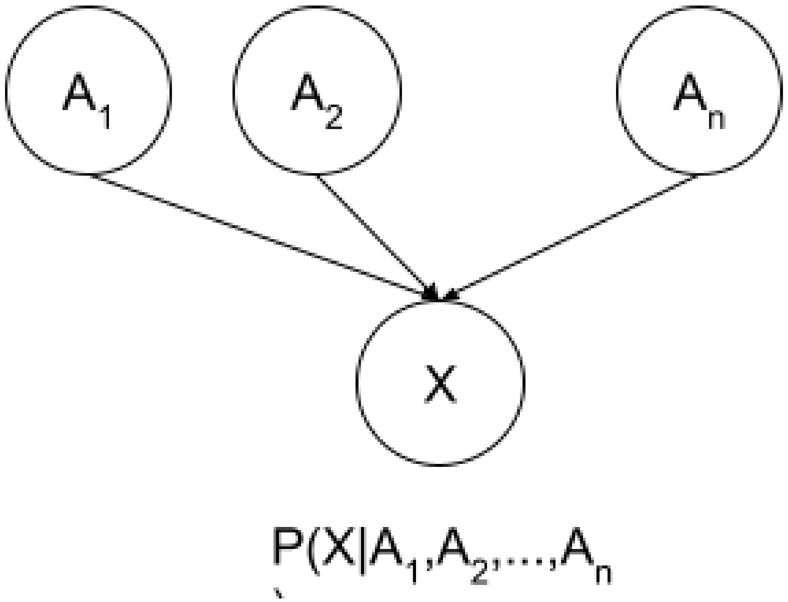
- No interactions between variables: **absolute independence**

Example: traffic

- Variables: R it rains, T there is traffic
- We can model it as independent (model 0)
- Or we can model it that rain causes traffic (model 1) or traffic causes rain (model 2)



- Which is better? What does an agent gain by using model 1?



Bayes' net semantics

- A set of nodes, one per variable X
- A directed, acyclic graph
- A conditional distribution for each node
    - A collection of distributions over X, one for each combination of parent's values
        - $P(X|a_1, a_2, \ldots a_n)$
- **Bayes Net = Topology (graph) + Local Conditional Probabilities**

Probabilities in Bayes' nets

- Bayes' nets implicitly encode joint distributions
    - As a product of local conditional distributions
    - To see what probability a Bayes Net gives to a full assignment, multiply all the relevant conditionals together

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

- This means that a Bayes net cannot express any possible joint distribution, only those that can be written this way.
    - The smaller the number of arrows (parent/child relationships) the less number of values we need to describe the joint distro described by the Bayes net.

Probabilities in Bayes' nets

- But, wait, will this result in proper joint distribution? Meaning, does it add up to 1.0?
- Chain rule (valid for all distributions)

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | x_1, \ldots x_{i-1})$$

- If we assume the conditional independence:

$$P(x_i | x_1, \ldots x_{i-1}) = P(x_i | parents(X_i))$$

- What this means that the node only depends on the parents!
  - If we know how the parents turned out, the probability of the node doesn't depend on the rest of the network ("**d-separation**")

Example Bayes net - independent nodes

Example Bayes net - Naive Bayes

Example Bayes net - Markov chain

Example Bayes net - Hidden markov model

Causality?

- Bayes' nets are just conditional probabilities
  - Arrows just encode conditional independence
  - The arrows only reflect correlation, not causation.
  - This is good, because we can build a Bayes net even if we don't understand causation
- Sometimes we can build the Bayes net such that it really reflects causality
  - Often simpler (nodes have fewer patterns)
  - Often easier to think about
  - Often easier to elicit from experts

What do we think about Bayes nets?

- The whole area of probabilistic graphical models was the last great AI idea before the deep learning revolution
- It is still in the spirit of expert systems:
  - Interview an expert to deduct the structure of the Bayes net
  - Ask the expert for the probabilities - or learn the probabilities from data.
- They still have benefits:
  - Explainability: they can be understood by humans
  - Often learning can be more efficient if we first restrict what we are learning by recognizing conditional independencies.