Machine Learning

Nirdosh Bhatnagar

Graphical Models

1. Introduction

- Graphical models are diagrammatic representations of probability distributions.
- It is a marriage between probability theory and graph theory.
- These are also called probabilistic graphical models.
- Augments / improves analysis instead of using pure algebra.
- Inference and learning are treated together.
- Supervised and unsupervised learning are merged seamlessly.
- Missing data handled nicely.
- Interpretability is available if necessary.
- Focus on conditional independence and computational issues.
- Key insights are provided from graphical models
- Alternative names for graphical models as per Michael I. Jordan.
  - Belief networks.
  - Bayesian networks
  - Probabilistic independence networks
  - Markov random fields
  - Loglinear models
  - Influence diagrams

It is not necessary to learn that which can be inferred.

2. Topics

- Applications
- What is a Graph?
- Probability Theory
- Graphical Models in Engineering
- Graph Directionality
- Other Types of Special Graphical Models
- Informal Definition
- Conditional Independence

3. Applications

Some possible applications of graphical modelling techniques are:

- Image analysis / segmentation / resaturation
- Medical and technical diagnosis
- Maximum likelihood decoding, error correcting codes

4. What is a Graph?

– A graph is a mathematical structure which consists of:

  – Vertices

  – Edges (links, arcs).

  The edges can be either directed (with arrows, with direction), or undirected (without arrows, without direction).

Example In a graph $G = (V, E)$, V is the set of vertices, and E is the set of edges. In this graph

$$V = \{v_1, v_2, v_3\}$$

$$E = \{a, b, c\}$$

where $(v_1, v_2) = a$, $(v_1, v_3) = b$, and $(v_2, v_3) = c$. This is an example of directed graph.

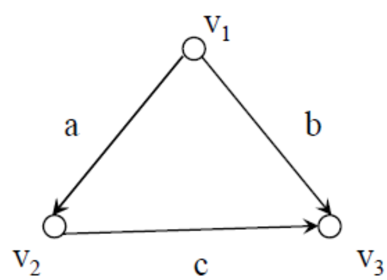


Figure 1. Graph $G = (V, E)$.

5. **Probability Theory**

- Probability theory is a natural tool for handling uncertainty and complexity.
- Consider a set of random variables

$$\{X_1, X_2, \ldots, X_n\}$$

We are interested in:

  - Which variables are independent?
  - Which variables are conditionally independent given others?
  - What are the marginal distributions of subsets of the random variables?

- These questions are answered with the joint distribution

$$P\left(X_1, X_2, \ldots, X_n\right)$$

  - Marginalization is answered by summing over the joint distribution.
  - Independence is answered by checking factorization.

- Assume that the variables are discrete. The joint distribution is a table $p\left(x_1, x_2, \ldots, x_n\right)$. If there are $r \in \mathbb{P}$ possible values for each random variable, then a naive representation of this table contains r^n elements. When n is large, this is expensive to store and use.
- Graphical models provide a more economic representation of the joint distribution by taking advantage of local relationships between random variables.
- In a some what simplistic manner, probability theory can be expressed in terms of two simple equations.
  - *Sum rule*: Let X and Y be random variables; and $a, b \in \mathbb{R}$.

$$P\left(X=a\right)=\sum_{b} P\left(X=a, Y=b\right)$$

  - *Product rule*: Let A and B be events.

$$P\left(AB\right)=P\left(B \mid A\right) P\left(A\right), \quad \text{where } P\left(A\right)>0$$

  All probabilistic inference and learning amounts to repeated application of the sum and product rules.

6. Graphical Models in Engineering

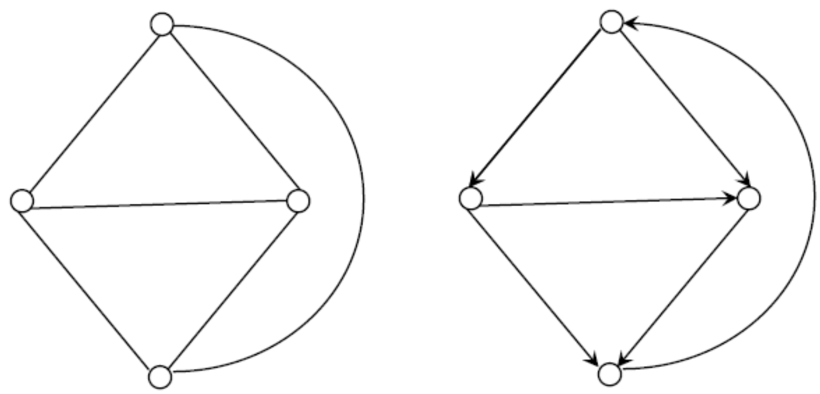
- Fundamental to the idea of a graphical model is the notion of modularity.
  - A complex system is built by combining simpler parts (termed modules).
- Probability theory provides the glue between a mathematical model and data.
- Graph theory provides:
  - An intuitively appealing interface.
  - Data structure. This in turn, lends itself naturally to designing efficient general-purpose algorithms.
- Role of graphical models in machine learning.
  - Simple way to visualize structure of probabilistic models.
  - Provides insights into properties of models. Conditional independence properties can be visualized by inspecting the underlying graph.
  - Complex computations, which are required to perform inference and learning can be expressed as graphical manipulations.
- Graphical description is used to indicate conditional dependencies among various parameters.
- The graph encodes dependencies among interacting variables.

7. Graph Directionality

There are two types of graphical models.

1. Undirected graphical model.
2. Directed graphical model.

Examples of undirected and directed graphs.



(a) Undirected graph. (b) Directed graph.

Figures 2. Different types of graphs.

1. *Undirected graphical models*
   (a) Graphs have links without direction (no arrows).
   (b) Can be used for describing Markov random fields (*Markov networks*).
      i. Better suited to express soft constraints between variables.
      ii. More popular in applications dealing with vision and physics.
2. *Directed graphical models*
   (a) Graphs have links with direction (have arrows).
   (b) Can be used for describing *Bayesian networks*.
   Expresses causal relationships between random variables.
   (c) More popular in artificial intelligence (AI) and statistics.
   (d) Markovian models are a special case.

Acyclic graph

– An undirected or directed graph with no cycles (closed loops) is called an acyclic graph.

– A tree graph has no cycles.

– A digraph (directed graph) is acyclic if and only if its vertices can be sequenced (ordered) such that its adjacency matrix is an upper triangular matrix. Note that in an upper triangular matrix, all of the matrix elements in the lower triangle are all zeros.

8. **Other Types of Special Graphical Models**

Some other types of special graphical models are as follows.

– *Factor graph*: Undirected bipartite graph connecting variables and factors.

Each factor represents a function over the variables it is connected to.

– *Clique tree / Junction tree:* This is a tree of cliques used in junction tree algorithms.

– *Chain graph*: Such graphs have both directed and undirected edges.

Directed acyclic graphs and undirected graphs are special cases of chain graphs.

Used to describe generalized Bayesian and Markov networks.

– *Ancestral graphs*: Have undirected, directed, and bidirected edges.

– *Conditional random fields*: Discriminative model specified over an undirected graph.

– *Restricted Boltzman machine*: This is a bipartite generative model specified over an undirected graph.

9. Informal Definitions

Graphical model is a graphical representation of (conditional) independence relationships in a joint probability distribution.

This in turn leads to a convenient evaluation of the joint probability distribution itself.

- The graphical model utilizes a graph as its base structure.
- And the parametrization depends upon the graph itself, where the parameters themselves are *local.*
- *Independence* in the joint distribution $\Leftrightarrow$ *Separation* in the graph.
- This mapping is *neither unique* nor *perfect*.

A directed graphical model utilizes a:

- Directed acyclic graph. *A missing edge in the graph encodes an independence relationship.*
- The vertices are random variables: $X_1, X_2, ..., X_n$ (sometimes events also).
- Edges denote the *parent of* relationship, where Θ_i is the set of parents of X_i.

We have

$$P\left(X_1, X_2, ..., X_n\right) = \prod_{i=1}^{n} P\left(X_i \mid \Theta_i\right)$$

This joint distribution is defined in terms of *local probability tables*. Each table contains the conditional probabilities of a variable for each value of the conditioning set. Advantages of local probability tables are:

- Suppose $X_i,\ i = 1, 2, \ldots, n$ are binary (Bernoulli) random variables. The full joint distribution requires 2^n entries in the table.
- The total number of entries in the tables which are necessary to compute the joint distribution via the graphical model requires

$$\sum_{i=1}^{n} 2^{|\Theta_i|+1} \text{ entries}$$

- This number can be reduced to simply

$$\sum_{i=1}^{n} 2^{|\Theta_i|} \text{ entries}$$

  This is possible because, only half the entries in each table are necessary to convey full information. So for example, if $P\left(X = T\right) \triangleq a$, then $P\left(X = F\right) = (1 - a)$ can be computed easily. Hence the probability $P\left(X = F\right)$ need not be stored. This leads to a reduction in storage.

10. Conditional Independence

Notation:

– The symbol $\Leftrightarrow$ means *if and only if*. "If and only if" is also abbreviated as "iff."

– Also $A \cap B$ is abbreviated as AB. □

10.1. Independence of Events.

Definition 1. *Independence of events.*

(a) *Let events A and B be independent of each other. This statement is abbreviated as $A \perp B$.*

(b) $A \perp B \Leftrightarrow P(AB) = P(A)P(B)$. □

Fact 1. $A \perp B \Leftrightarrow P(A \mid B) = P(A)$, where $P(B) > 0$.
Proof:

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

□

Fact 2. Product rule. $P(AB) = P(A)P(B \mid A)$. □

Fact 3. Repeated application of product rule give

$$P(ABC) = P(A)P(B \mid A)P(C \mid A, B)$$

□

10.2. Conditional Independence of Events.

Definition 2. *Conditional independence of events. Given event C, events A and B are conditionally independent iff*

$$P(AB \mid C) = P(A \mid C)P(B \mid C), \quad \textit{where } P(C) > 0$$

□

Fact 4: Events A and B are conditionally independent, given event C. Then

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

Proof: Note that

$$\begin{aligned} P(A \mid B, C) &= \frac{P(A, B \mid C)}{P(B \mid C)} \\ &= \frac{P(A \mid C)P(B \mid C)}{P(B \mid C)} \\ &= P(A \mid C) \end{aligned}$$

$\square$

Concept of conditional independence can be similarly extended to random variables.