Lecture 3, Directed Graphical Models

DS-GA 1005 Inference and Representation, Fall 2023

Yoav Wald

09/20/2023

Today's Plan

- Conditional Independence
- Directed Graphical Models (a.k.a Bayesian Networks)

Previous Lectures

In the previous episodes we learned that:

- Brute-force estimation of probabilistic models (i.e. assigning a parameter for each state) is intractable in high-dimensions
- *Possible solution*: Estimate first and second moments, and "complete the rest" by an inductive rule (e.g. max-entropy)
    - Excellent computational and statistical complexity
    - Bad approximation when higher moments do not adhere to the maximum-entropy principle (i.e. non-Gaussian distributions)
- Natural question: what are other *useful* assumptions that make learning tractable?

Previous Lectures

In the previous episodes we learned that:

- Brute-force estimation of probabilistic models (i.e. assigning a parameter for each state) is intractable in high-dimensions
- *Possible solution*: Estimate first and second moments, and "complete the rest" by an inductive rule (e.g. max-entropy)

Today's lesson: Statistical independence assumptions

What can be Gained from Independence?

- For convenience, focus on binary variables $X_1, \ldots, X_d$, where $X_i \in \{0, 1\} \; \forall i \in [d]$
- Example: return to our medical diagnosis motivation
    - Medical event $X_1 =$ Pneumonia, $X_{d/2+1} =$ Ear Infection
    - Symptoms of pneumonia, $X_2 =$ Cough$, \ldots, X_{d/2} =$ Chest Pain
    - Symptoms of infection, $X_{d/2+2} =$ Ear Ache$, \ldots, X_d =$ Nausea
- We wish to learn a model $P_{\boldsymbol{\theta}}(X_1, \ldots, X_d) = P_{\boldsymbol{\theta}}(\mathsf{P}, \mathsf{EI}, \mathsf{C}, \mathsf{CP}, \mathsf{EA}, \mathsf{N}, \ldots)$

What can be Gained from Independence?

Task: learn a model $P_{\boldsymbol{\theta}}(\mathsf{P}, \mathsf{EI}, \mathsf{C}, \ldots, \mathsf{CP}, \mathsf{EA}, \ldots, \mathsf{N})$
Strategy: Assume independence to break $P_{\boldsymbol{\theta}}$ into a product of smaller chunks; learn each small model separately

- Example: assume pneumonia and ear infection are independent (symptoms included)
  - Medical event $X_1 =$ Pneumonia, $X_{d/2+1} =$ Ear Infection
  - Symptoms of pneumonia, $X_2 =$ Cough, $\ldots, X_{d/2} =$ Chest Pain
  - Symptoms of infection, $X_{d/2+2} =$ Ear Ache, $\ldots, X_d =$ Nausea
- Formally: $X_{[d/2]} \perp\!\!\!\perp X_{[d/2+1,\, d]}$
  Recall that $X_i \perp\!\!\!\perp X_j$ if $P(X_i, X_j) = P(X_i)P(X_j)$

What can be Gained from Independence?

Task: learn a model $P_{\boldsymbol{\theta}}(\mathsf{P}, \mathsf{EI}, \mathsf{C}, \ldots, \mathsf{CP}, \mathsf{EA}, \ldots, \mathsf{N})$
Assumption: $X_{[d/2]} \perp\!\!\!\perp X_{[d/2+1,\, d]}$
Result: $P_{\boldsymbol{\theta}}(X_1, \ldots, X_d) = P_{\boldsymbol{\theta}}(\mathsf{P}, \mathsf{C}, \ldots, \mathsf{CP}) \cdot P_{\boldsymbol{\theta}}(\mathsf{EI}, \mathsf{EA}, \ldots, \mathsf{N})$

- How many parameters do we need to estimate?
- How many samples are (approximately) required for learning?
- Can we further break down $P_{\boldsymbol{\theta}}$?

Marginal and Conditional Independence

- Marginal independence, i.e. $P(X_i, X_j) = P(X_i)P(X_j)$, is a special case of conditional independence
- Conditional independence, $X_i \perp\!\!\!\perp X_j \mid X_k$:

$$P(X_i, X_j \mid X_k = x_k) = P(X_i \mid X_k = x_k)P(X_j \mid X_k = x_k)$$

  for any x_k such that $P(X_k = x_k) > 0$
- **Claim:** Conditional independence can also be defined as $P(X_i \mid X_j, X_k = x_k) = P(X_i \mid X_k = x_k)$

Conditional Independence Examples: Naïve Bayes

Example 1: What if we assume that symptoms are independent conditioned on medical event? Cough $\perp\!\!\!\perp$ Fever | Pneumonia, etc.

- We have $X_i \perp\!\!\!\perp X_j \mid X_1 \quad \forall 1 < i, j \leq d/2$
  where $X_1 =$ Pneumonia, $X_2 =$ Cough, $\ldots, X_{d/2} =$ Chest Pain
- This lets us further break down our model

$$P_{\boldsymbol{\theta}}(X_1, X_2, \ldots, X_{d/2}) = P(X_1) \cdot P(X_2 \mid X_1) P(X_3 \mid X_1, X_2) \cdot \ldots \cdot P(X_{d/2} \mid X_1, \ldots X_{d/2-1})$$

$$\begin{aligned} P_{\boldsymbol{\theta}}(X_1, X_2, \ldots, X_{d/2}) =& P(X_1) \cdot P(X_2 \mid X_1) P(X_3 \mid X_1, X_2) \\ & \cdot \ldots \cdot P(X_{d/2} \mid X_1, \ldots X_{d/2-1}) \\ =& P(X_1) \prod_{i=2}^{d/2} P(X_i \mid X_1) \end{aligned}$$

$$P_{\boldsymbol{\theta}}(X_1, X_2, \ldots, X_{d/2}) = P(X_1) \prod^{d/2} P(X_i \mid X_1)$$

Conditional Independence Examples: Naïve Bayes

Example 2: Spam filter

- Y = Spam/Not Spam,
  X_1 = Does the word "prince" appear in the email?
  X_2 = Does the word "heritage" appear in the email?
  ...
- A Naïve Bayes model assumes
  $P(Y, X_1, \ldots, X_d) = P(Y) \prod_{i=1}^{d} P(X_i \mid Y)$

Conditional Independence to Graphical Models

- More examples of conditonal independence: Markov models $X_{[t-1]} \perp\!\!\!\perp X_{t+1} \mid X_t$.
- Notice that in all these examples we used Bayes rule + independence to rewrite P as a product of smaller distributions
- Q: if $X_i \perp\!\!\!\perp X_j$, does it hold that $X_i \perp\!\!\!\perp X_j \mid X_k$?
    - Maybe the other way around?

Probabilistic Graphical Models

- Goal: A mathematical language to relate factorizations of probability distributions, and independence properties
  - Given such a language, maybe we can come up with learning and inference algorithms that work for many types of models
- Most natural mathematical object to use for this language is a graph $G = (V, E)$
- Today we will talk about directed graphs

Writing Distributions in a Factorized Form

- We can always write a given distribution $P(X_1, \ldots, X_d)$ as a product of conditional distributions (factors)
  1. Choose some ordering of the variables, and write
  $$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i \mid X_{[i-1]})$$
  2. We may obtain additional factorizations if for some set $Pa(i) \subseteq [i-1]$, we have $X_i \perp\!\!\!\perp X_{[i-1] \setminus Pa(i)} \mid X_{Pa(i)}$:
  $$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i \mid X_{Pa(i)})$$
- Let us associate these factorizations with graphs

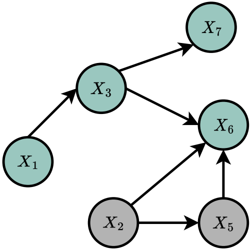
Directed Acyclic Graphs (DAGs)

Definition

A directed graph is a data structure $G = (V, E)$ where $E = \{(i,j), i,j \in V\}$ are **ordered** tuples (also $i \to j$). G is acyclic if it has no directed paths from any node $i \in V$ to itself ($i \not\rightsquigarrow i$)

We will usually consider $V = \{X_i\}_{i=1}^d$, where each random variable corresponds to a vertex

- *Topological ordering*: An ordering $\sigma_1 < \sigma_2 < \ldots < \sigma_d$ of $V = \{X_{\sigma_k}\}_{k=1}^d$ such that $\sigma_i < \sigma_j$ for all $(X_i, X_j) \in E$
- For $i \in V$, we define its parents $Pa(i) = \{j : (j,i) \in E\}$, and non-descendants $ND(i) = \{j : i \not\rightsquigarrow j\}$



Correspondence between DAGs and Factorizations

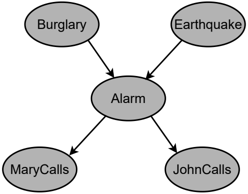
There seems to be a direct association between a probabilistic model P and a DAG G

- $(P \to G)$ a factorization of P defines a DAG G, why?
  - We wrote down P as $\prod_{i=1}^{d} P(X_i \mid X_{Pa(i)})$ and $Pa(i) \subseteq [i-1]$
- $(G \to P)$ a DAG G can describe properties of distributions that "has the same structure" as the graph

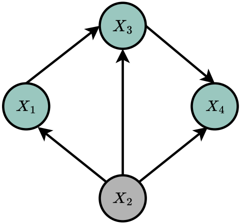
Q: What is the exact correspondence? How is it related to conditional independence?

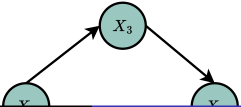
Graphical Models: Some Examples

- Alarm example, write down the factorized distribution



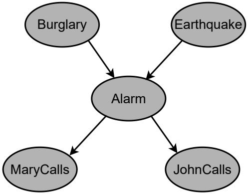
- Is a distribution P always associated with some DAG?





The Separation-Independence Connection: an Intuition

- Intuitively, conditional independence $X_i \perp\!\!\!\perp X_j \mid X_k$ means that observing X_k blocks the flow of information between X_i and X_j
- We can also define separation in G, where the vertex X_k blocks all paths between two vertices X_i, X_j
- Let us explore this correspondence in detail



Independence Sets and I-Maps

Definition (Independence set)

Let P be a distribution over $\mathcal{X} = \{X_1, \ldots, X_d\}$. Then $\mathcal{I}(P)$ is the set of all conditional independence statements of the form $X \perp\!\!\!\perp Z \mid Y$ that hold for P

- Intuitively, for each P we will want to establish the existence of a graph from which we can read off $I(P)$. Why?

Independence Sets and I-Maps

Definition (Factorization)

Let G be a DAG over vertices that correspond to random variables $X_1, \ldots, X_d$. We say that P **factorizes over** G if $P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i \mid X_{Pa(i)})$, where $Pa(i)$ are the parents of X_i in G.
We call the tuple (P, G) a Bayesian network if P is specified as a set of conditional distributions associated with vertices of G

Recall, $\mathcal{I}(P)$ is the set of independence statements that hold in P

Definition (I-map)

A DAG G is an I-map for P if $\mathcal{I}_l(G) \subseteq \mathcal{I}(P)$, where $\mathcal{I}_l(G)$ is the set of local independencies of G,

$$\mathcal{I}_l(G) = \{X_i \perp\!\!\!\perp X_{Nd(i)} \mid Pa(i) \quad \forall i\}$$

Correspondence between $\mathcal{I}_l(G)$ and $\mathcal{I}(P)$

Theorem (Thm 3.1 and 3.2 on Koller & Friedman)

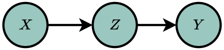
P factorizes according to G if and only if G is I-map for P

- The theorem tells us that if P factorizes over G, it is guaranteed that it satisfies all independence statements in $\mathcal{I}_l(G)$, i.e. $\mathcal{I}_l(G) \subseteq \mathcal{I}(P)$
- Q: Are there additional conditional independence constraints that are encoded by G?

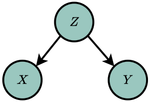
d-separation

- d-separation provides a criterion to check whether G encodes a conditional independence $X_1 \perp\!\!\!\perp X_2 \mid X_3$, where X_1, X_2, X_3 are some disjoint subsets of vertices in G
- It examines whether there is an "active path" in the graph that allows influence to flow. Paths are consisted of 3 building blocks

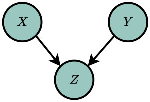
- Cascade



- Common Cause

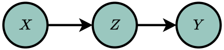


- Common Effect (Z is a collider)

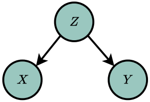


d-separation

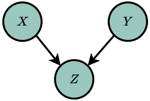
- Cascade



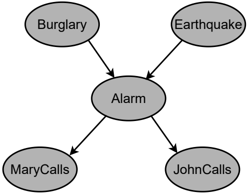
- Common Cause



- Common Effect (Z is a collider)



- *Intuition*: conditioning on colliders and their descendants activates paths, conditioning on other vertices deactivates them



D-Separation

Definition (active trail)

An *undirected* trail between X_1 and X_n is active given a set of vertices $\mathbf{Z}$ if

- For every collider X_i on the trail, either X_i or one of its descendants is in $\mathbf{Z}$
- No other node along the trail is in $\mathbf{Z}$

Definition (d-separation)

Vertices X, Y are d-separated given $\mathbf{Z}$ if there are no active paths between them given $\mathbf{Z}$

Claim

If P factorizes over G then $\mathcal{I}(G) \subseteq \mathcal{I}(P)$

D-Separation

It turns out that $\mathcal{I}(G) = \{X \perp\!\!\!\perp Y \mid Z : X, Y \text{ d-separated given } Z\}$ captures the most possible independence statements that can be read from a DAG.

Claim

If P factorizes over G then $\mathcal{I}(G) \subseteq \mathcal{I}(P)$

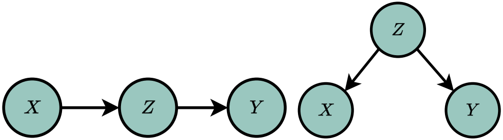
Proof: in future lectures

Claim

For almost all distributions P that factorize over G (all except a measure 0 set) it holds that $\mathcal{I}(G) = \mathcal{I}(P)$

Does Each Distribution P has a canonical graph?

- It is tempting to think that for any P there is a single "true" graph associated with it, in the sense that $\mathcal{I}(P) = \mathcal{I}(G)$
- This cannot hold because there are graphs G_1, G_2 that have $\mathcal{I}(G_1) = \mathcal{I}(G_2)$.



Can we Always Capture $\mathcal{I}(P)$ via $I(G)$?

Furthermore, can we always find a graph G such that $\mathcal{I}(P) = \mathcal{I}(G)$? No, as demonstrated by this counter-example:

$$P(x,y,z) = \begin{cases} 1/12 & x \oplus y \oplus z = 0 \\ 1/6 & x \oplus y \oplus z = 1 \end{cases}$$

- It is simple to show that $X \perp\!\!\!\perp Y$, and from symmetry also that $Y \perp\!\!\!\perp Z$ and $Z \perp\!\!\!\perp X$
- On the other hand, $X \not\perp\!\!\!\perp Y \mid Z$

Conclusion: $\mathcal{I}(G) \neq \mathcal{I}(P) = \{X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z, Y \perp\!\!\!\perp Z\}$ for all G

Conclusion

- Bayesian Networks are an intuitive language (yet "imperfect") to encode conditional independence and factorization of distributions
- Useful for
  - More efficient learning (estimating less parameters)
  - We'll see in the future: enables generic graph-based inference algorithms
  - More...
- Recitation: examples of HMMs, Next lectures: undirected models, latent variables, variational inference ...