Lecture 3

Xuanxi Zhang

Topics

- Conditional Independence ↔ Graphical Representation
- Existence and Unicity
- Independence Maps and D-separation

Recall from Lecture 1: Brute-force approach to estimate a joint probability distribution does not scale — need to introduce structure!

Recall from Lecture 2: First attempt at structure: focus on first two moments and disregard the rest → Gaussian Estimation. Efficient in high dimensions, but limited power to model many important phenomena.

Alternative route to structure: Statistical Independence.

1 Independence

Let $(X, Y) \sim P$ with marginals P_X and P_Y.

- X and Y are independent if and only if:

$$P = P_X \otimes P_Y$$

- When P has density $p(x, y)$:

$$p(x, y) = p_X(x)p_Y(y)$$

Key Property The marginals in the right-hand side (RHS) are of lower dimension.

Example: Hypercube Consider again the hypercube $\mathbb{H}^d = \{\pm 1\}^d$.

Suppose that $\{X_1, \ldots, X_{d/2}\}$ is independent from $\{X_{d/2+1}, \ldots, X_d\}$. Then,

$$P_\theta(X_1, \ldots, X_d) = P_{\theta_1}(X_1, \ldots, X_{d/2}) \cdot P_{\theta_2}(X_{d/2+1}, \ldots, X_d)$$

where $\theta = (\theta_1, \theta_2)$. Given n i.i.d. observations $x^1, \ldots, x^n$, the MLE for θ is:

$$\frac{1}{n}\sum_{i=1}^{n} \log P_\theta(x^i) = \frac{1}{n}\sum_i \log P_{\theta_1}(x^i_{1:d/2}) + \frac{1}{n}\sum_i \log P_{\theta_2}(x^i_{d/2+1:d})$$

$$= \mathcal{L}(\theta_1) + \mathcal{L}(\theta_2)$$

This corresponds to **multivariate models in d/2 dimensions**.

Gain in Sample Complexity

- Sample complexity reduces from $O(2^d)$ to $O(2^{d/2})$.
- This breaks the curse of dimensionality by hierarchically splitting variables into independent components.

Conditional Independence

Independence is a (very) strong assumption!
Can a weaker assumption lead to similar gains?
Recall the Bayes factorization of a joint distribution:

$$p \in \mathcal{P}(\mathcal{X}^d), \quad p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{1:i-1})$$

The right-hand side (RHS) can be represented using a **directed graph**:

$$G = (V, E)$$

where the **vertex set** is $V = \{X_i\}_{i=1}^d$.
The probability factorization can be rewritten as:

$$p(x_i \mid \{x_j\}_{j \in A(i)}) \Rightarrow A(i) = \{j \in V \mid j \to i \in E\}$$

- The **parents** of each node encode the conditioning set.
- Notice that the **in-degree** D_i of each node:

  $$D_i = |A(i)|$$

  determines the dimension of the associated conditional probability model.
- Thus, when $\max_i D(i) = O(d)$, we are **cursed by dimension**.
- In this example, we have, in fact, the **densest possible** graph.

Directed Acyclic Graphs (DAGs)

A **directed, acyclic graph (DAG)** is similar to the complete graph but under Directed Acyclic (DA) constraints.

Why Acyclic?

Question: Why must the graph be acyclic?
Answer:

- A Bayes decomposition starts with an **ordering of variables**, (which can be arbitrary).
- This defines a **topological ordering** in G, such that:

  $$(i, j) \in E \Rightarrow j > i.$$

- It follows that G **cannot have cycles**. Suppose a cycle exists:

  $$i_1 \to i_2 \to i_3 \to \cdots \to i_k \to i_1$$

  but this contradicts the topological ordering, where:

  $$i_1 < i_2 < i_3 < \cdots < i_k < i_1,$$

  which is impossible.

Alleviating the Curse of Dimensionality

To alleviate the **Curse of Dimensionality (CoD)**, we need to **sparsify** the graph by removing edges so that the **degree** is reduced.

Edge Removal and Conditional Independence

Removing edges from this graph amounts to stating:

$$p(X_i \mid X_{1:i-1}) = p(X_i \mid X_{A(i)})$$

where the conditioning set is smaller, with:

$$|A(i)| < i - 1.$$

First Non-Trivial Setting

$$p(X_3 \mid X_1, X_2) = p(X_3 \mid X_1) \quad \text{(or equivalently, } p(X_3 \mid X_1, X_2)\text{)}$$

This means that X_2 and X_3 are **conditionally independent** given X_1.

Fact: Conditional Independence Property

Assume $p(X_2 \mid X_1) > 0$ for all X_1, X_2. Then:

$$p(X_3 \mid X_1, X_2) = p(X_3 \mid X_1) \iff p(X_2, X_3 \mid X_1) = p(X_2 \mid X_1) \cdot p(X_3 \mid X_1)$$

Proof:

$$\Rightarrow : \; p(X_3, X_2 \mid X_1) = p(X_2 \mid X_1) \cdot p(X_3 \mid X_1, X_2) = p(X_2 \mid X_1) \cdot p(X_3 \mid X_1)$$

$$\Leftarrow : \; p(X_3 \mid X_1, X_2) = \frac{p(X_2, X_3 \mid X_1)}{p(X_2 \mid X_1)} = p(X_3 \mid X_1).$$

Important Fact

Unconditional independence does not imply conditional independence:

$$X \perp Y \not\Rightarrow X \perp Y \mid Z$$

$$X \perp Y \not\Leftarrow X \perp Y \mid Z$$

Counter-Examples

Example 1:

- X: Alice is working late at night.
- Y: Bob is working late at night.
- Z: There is a conference deadline tomorrow!

Example 2:

- X: Parent 1 is left-handed.
- Y: Parent 2 is left-handed.
- Z: The child is left-handed.

Example: Naïve Bayes Classifier

Consider discrete covariates $X = (X_1, \ldots, X_d)$ and a label Y.
For example:

- $Y = \mathbf{1}(\text{email is spam})$
- $X_i = \mathbf{1}(\text{word}_i \in \text{email})$

Naïve Bayes Model:

$$p(X_1, \ldots, X_d, Y) = p(Y) \cdot \prod_{i=1}^{d} p(X_i \mid Y)$$

Naïve Bayes and Spam Detection

Spam is detected given X if:

$$P(Y=1)\prod_{i=1}^{d} P(X_i \mid Y=1)$$

is greater than:

$$P(Y=0)\prod_{i=1}^{d} P(X_i \mid Y=0).$$

Bayesian Networks

- First, pick an **ordering** of variables and consider the model:

  $$P(X_1,\ldots,X_d)=\prod_{i=1}^{d} P(X_i \mid X_{A(i)})$$

  with $A(i)\subseteq\{1,\ldots,i-1\}$.
- We saw how to represent this model using a **Directed Acyclic Graph (DAG)**:

  $$G=(V,E),$$

  where $(i\to j)\in E \iff i\in A(j)$.
- **Mapping Between Joint Probability Distributions and DAGs.**

Open Questions

- What are the best **algorithms** for inference and querying the probability model?
- How accurate is the **translation** between joint probability distributions and DAGs?

To be discussed later.

Example

Consider the following variables:

- G: Gender
- A: Age
- MH: Mother's Height
- FH: Father's Height
- H: Child's Height
- W: Child's Weight

Graphical Representation

The Bayesian network structure implies the following joint probability decomposition:

$$P(A,G,H,W,MH,FH)=P(G)\cdot P(A)\cdot P(MH)\cdot P(FH)\cdot P(H \mid G,MH,FH,A)\cdot P(W \mid G,A)$$

Topological Ordering

The topological ordering of the nodes in the network helps in identifying conditional independences.

Reading Off Conditional Independences

From the structure of the graph, we can infer the following conditional independences:

$$X_1 \perp X_2, \quad X_1 \perp X_3, \quad X_2 \perp X_3, \quad X_3 \perp X_4$$

$$X_5 \perp X_6 \mid X_4$$

Key Property

A variable is **independent** of its **non-descendants**, given its **parents**.

Takeaway So Far

We have two different notions of **separation**:

- A **probabilistic** one: $X \perp Y \mid Z$ (between random variables).
- A **topological** one between nodes in a graph.

Question: Given a probabilistic model, can we always build a DAG G such that these two notions agree?

Independence Sets and Maps

Definition: Independence Set

Let P be a probability distribution on $\mathcal{X}$ (input space). We define $I(P)$ as the set of all **conditional independence** relations satisfied by P, i.e.,

$$X \perp Y \mid Z, \quad Z \subseteq \mathcal{X}, \quad X, Y \in \mathcal{X}.$$

Ideally, we would like the **graphical representation** to capture $I(P)$.

Definition: Factorization

Let G be a Directed Acyclic Graph (DAG) over variables $X_1, \ldots, X_d$. We say that P **factorizes** over G if:

$$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i \mid X_{A(i)})$$

where $A(i)$ represents the **parents** of node i in G.

Definition: Independence Map

A DAG G is an **Independence Map** for P if:

$$I_P(G) = \left\{ X_i \perp \text{Non-Descendants}(X_i) \mid X_{A(i)} \right\}, \quad i = 1, \ldots, d$$

is satisfied under P. This means that $I_P(G)$ contains the **local independencies**.

Key Fact

$$P \text{ factorizes according to } G \iff G \text{ is an independence map for } P.$$

Proof

This follows from **Frydenberg & Lauritzen Theorems 3.1 and 3.2** (extending our previous argument).

Interpretation

In other words, knowing that P **factorizes** according to G provides an explicit list of **local conditional independencies**, meaning:

$$I_P(G) \subseteq I(P).$$

Open Question

Can we also obtain non-local conditional independencies from G?

D-Separation

Goal: Given a DAG G and variables $X, Y \in \mathcal{X}$, determine whether:

$$X \perp Y \mid \mathcal{Z}, \quad \mathcal{Z} \subseteq \mathcal{X}$$

Main Idea: Determine whether there exists an **active path** between X and Y when $\mathcal{Z}$ is observed.

Local Rules

1. **Cascade:**

   $$X_i \to Z \to X_j$$

   Path is blocked if $Z \in \mathcal{Z}$, i.e., Z is observed.

2. **Common Parent (Deadline Example):**

   $$X_i \leftarrow Z \to X_j$$

   Path is blocked if $Z \in \mathcal{Z}$.

3. **V-Structure (Leftie Example):**

   $$X_i \to Z \leftarrow X_j$$

   Path is blocked if $Z \notin \mathcal{Z}$.

Definition

If no active path exists between X and Y given $\mathcal{Z}$, we say that X and Y are **d-separated**.

D-Separation and Conditional Independence

Given a set $\mathcal{Z}$, we define:

$$I(G) = \{(X \perp Y \mid \mathcal{Z}) \mid X, Y \text{ are d-separated with respect to } \mathcal{Z}\}$$

Claim

If a distribution P factorizes over G, then:

$$I(G) \subseteq I(P).$$

In other words, any **conditional independence** that we extract from **d-separation** is satisfied by P.

Proof

(To be covered in the next lecture, using undirected graphs.)

Open Question

What about the case when:

$$I(G) = I(P)?$$

Answer

For any distribution P that factorizes over G, except for a set of measure zero, we have:

$$I(G) = I(P).$$

(This accounts for corner cases from degenerate distributions.)

Important Note

This does **NOT** imply a tight equivalence between **DAGs** and **probability models**.

Non-Uniqueness of DAG Representations

We can have two **non-isomorphic** DAGs G_1 and G_2 (i.e., $G_1 \neq G_2$), such that:

$$I(G_1) = I(G_2).$$

This means that they have the same **independence structure**, leading to a **lack of uniqueness**.

Lack of Existence of a DAG Representation

More importantly, some distributions P **cannot** be represented by any DAG, meaning:

$$I(G) \neq I(P) \quad \forall G.$$

For example, consider:

$$X, Y \sim \text{Bernoulli}(1/2) \text{ independently}, \quad Z = X \oplus Y.$$

The joint distribution (X, Y, Z) is **exchangeable** and **not adapted** to a directed graphical structure.

Summary

DAGs and Bayesian Networks provide a **powerful language** to express **conditional independencies**.

However, this **language is imperfect** due to:

- **Lack of existence**: Some distributions do not have a DAG representation.
- **Lack of uniqueness**: Different DAGs can encode the same independence structure.