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 Probability Basic

1.1 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**.

1.2 Conditional Independence

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

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

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:

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

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

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.

2 Bayes Factorization and Bayesian Networks

For a joint distribution $p \in \mathcal{P}(\mathcal{X}^d)$, we have:

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

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

$$G = (V, E)$$

where the **vertex set** is $V = \{X_i\}_{i=1}^d$. And edges are induced by the **conditioning** in the factorization:

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

properties of the graph: 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.

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

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

where the conditioning set is smaller, with:

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

2.1 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 $(i \to j) \in E \iff i \in A(j)$.

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.

2.2 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:

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

Spam is detected given X if:

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

is greater than:

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

3 Directed Acyclic Graphs (DAGs)

We can see that the graph build from bayes factorization is a DAG.
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.

4 Mapping Between Bayesian Probability and DAGs

Definition 1. 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 2 (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 3 (Independence Map)**.** A DAG G is an **Independence Map** for P if:

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

is satisfied under P. This means that $I_l(G)$ 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 **F & K Theorems 3.1 and 3.2** (extending our previous argument).

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

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

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

5 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.

Definition 4 (D-Separation)**.** Let X, Y, and Z be disjoint sets of nodes in a DAG. We say X and Y are d-separated by Z if every path from any node in X to any node in Y is blocked by Z.

A path is blocked by Z if at least one triple along the path:

1. Chain or Fork: $A \to B \to C$ or $A \leftarrow B \to C$. The path is blocked if the middle node B is **in** Z.
2. Collider: $A \to B \leftarrow C$, The path is blocked if the middle node B is **not in** Z and none of its descendants are in Z.

If a path is not blocked, we sometimes say it is "active." If exist a path between X and Y is active, then X and Y are not d-separated.

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.

Prop: 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.)

Note that this does **NOT** imply a tight equivalence between **DAGs** and **probability models**. Under the concept of **equivalence in independence relation**, we have:

1. 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:
$$I(G_1) = I(G_2).$$
. For example, consider:
$$X \to Y \to Z \quad \text{and} \quad X \leftarrow Y \to Z.$$

2. Non-Existence of DAG Representations: For some distributions P, ther does not exist a DAG G such that $I(G) = I(P)$. 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.